The Value of Data: Analyst Vs. Machine*

Stefano Bonini[†‡]
sbonini@stevens.edu

Thomas Shohfi[§]
tshohfi@wayne.edu

Majeed Simaan[‡]
msimaan@stevens.edu

Guofu Zhou[¶]
zhou@wustl.edu

September 12, 2024

*We thank conference participants at the 2024 CICF, the 2022 FMA Annual Meeting, the 2022 European FMA Meeting, the 2022 Applied FMA Meeting, the 2020 FMA Annual Meeting, the 2019 Eastern Finance Association Annual Meeting, and R/Finance 2019. We are also grateful to participants at Shanghai Jiao Tong University, Zhejiang University, Tsinghua University, Bentley University, Rensselaer Polytechnic Institute, Venice University, Wayne State University, and the Finance/Financial Engineering Brownbag Series at Stevens Institute of Technology for helpful comments and suggestions. We thank Hoang Nguyen (discussant), John Cotter (discussant), Banks Osborne (discussant), Nan Li (discussant), Laura Veldkamp, Ying Wu, Mustafa Gültekin, and Khaldoun Khashanah for comments and feedback. We thank the Donald Shohfi Financial Research Fund for computing support, and Yuvraj Chopra and Avnish Grover for excellent research assistance. All errors are our own.

[†]Corresponding author. Tel: (201) 215-3528, Email: sbonini@stevens.edu.
[‡]School of Business, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, USA.
[§]Mike Ilitch School of Business, Wayne State University, 2771 Woodward Ave, Detroit, MI 48201, USA.
[¶]Olin Business School, Washington University in St. Louis, 1 Snow Way Dr, St. Louis, MO 63130, USA.

The Value of Data: Analyst Vs. Machine

Abstract

Our paper studies the role of analysts' research outputs in predicting the cross-section of stock returns when a large number of firm characteristics provide competing information. Utilizing a 25-year dataset with 2,515 stocks and 1,309,335 analyst reports, we find that applying machine learning (ML) algorithms to 94 factors from Gu et al. (2020) yields 10-20% annual alphas. Analyst-generated information adds value only when the representative ML agent fails to process these baseline characteristics efficiently. With comprehensive factors and complex ML, analyst incremental value is minimal. Post-2003 Global Research Settlement analysis confirms this, showing a near-zero value of analysts' contributions.

Keywords: Information Processing, Machine Learning, Portfolio Construction, Sell-Side Analysts, Textual Analysis
JEL Codes: C19, C45, C55, C63, G10, G20, G24, M41

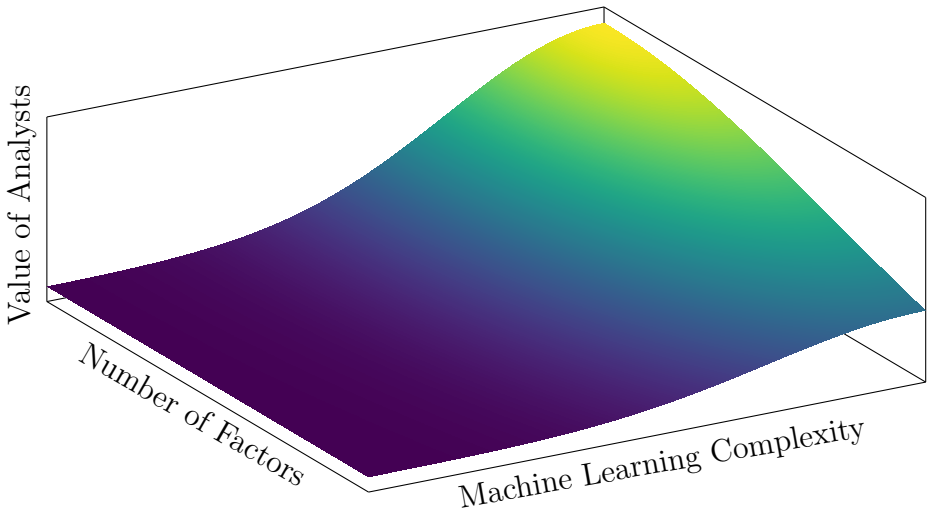
1 Introduction

Portfolio formation relies on accurately estimating asset inputs to determine optimal allocations. In a mean-variance paradigm, these inputs are the mean vector and covariance matrix of asset returns. However, it is well-known that such a model is sensitive to estimation risk and can result in extreme or unrealistic positions (Jagannathan and Ma, 2003). Alternatively, asset managers may utilize what is called a style-based investment approach, where assets are grouped into different categories based on certain characteristics. To generate expected returns beyond a given benchmark, investors can tilt their portfolios towards an ex-ante-chosen style to load on the associated risk premium. However, empirical evidence shows that risk premia are time-varying and are difficult to measure, making it one of the major challenges in asset pricing over the past decades (Nagel and Xu, 2023). For instance, many factors have been proposed in the literature to have investment value, generating the so-called "zoo of factors" (see, e.g., Chen and Zimmermann (2021)), which can be broadly classified into three categories: market-driven (returns, momentum, liquidity, beta, etc.), company fundamentals (valuation ratios, earnings/sales-to-price, etc.), and analyst-based (recommendation, price targets, earnings forecast, and qualitative content).

In isolation, many of the proposed factors generate alphas in-sample; yet, identifying ex-ante which factors have out-of-sample investment value is challenging (Harvey and Liu, 2021). Differently, recent contributions have shown that aggregating market and firm-related fundamental factors through advanced machine learning (henceforth ML) techniques yield superior returns and persistent alphas (see, e.g., Gu et al. (2020)). These studies, however, do not incorporate the contribution of analyst-produced information that a large body of literature argues has significant investment value (e.g., Bradshaw et al. (2012)) due to equity analysts' access to superior information and more sophisticated processing skills. In light of the previous results on the value of big data processing via ML (e.g., Chen et al. (2022)) and given that analysts process public (rather than private) data, whether analysts' information has any incremental value in forming portfolios remains an open question.

We address the above question by utilizing a large-scale sample of 2,515 stocks over 25 years and recommendations and textual content of over 1,309,335 million sell-side analyst reports. Based on this rich data set, we propose an integrated framework to isolate the contribution of commonly available factors, data analysis techniques, and human information processing for portfolio formation. We obtain several important results. First, we estimate a baseline benchmark deploying seven different algorithms - linear regression with elastic net, two random forests, three artificial neural net, and support vector regression with radial

Figure 1: **Value of Analysts as a Function of Complexity**



Notes: This figure summarizes the main findings of the manuscript by visualizing the marginal contribution of the analysts' data as a function of complexity. Complexity is captured using two dimensions. The first is the number of firm-specific characteristics used to train the machine learning algorithm. In the largest case, we consider 94 predictors based on Gu et al. (2020). We consider five predictors based on Ball and Ghysels (2018) (unreported) in the smallest case. The second complexity dimension qualitatively ranks machine learning (ML) algorithms based on nonlinearity. For instance, a linear model would be considered the lowest in complexity, whereas random forest or neural nets would be the highest. Such complexity ranking is qualitative; however, the decomposition approach by Molnar et al. (2020) provides further guidance on quantifying ML complexity.

kernel - to aggregate the 94 factors studied by Gu et al. (2020).[1] Conditional on the ML algorithm, we generate annual risk-adjusted returns (alphas) using a six-factor model based on Fama and French (2015) and Carhart (1997). Overall, we find robust evidence of economic and statistically significant alphas from the baseline models, ranging between 10% and 20%.[2]

Second, given the baseline benchmark, we estimate the value of the analyst's content by adding recommendation and textual content, both separately and jointly, to the complete set of 94 predictors. We find that analysts' information has a positive value when the factor space is small, and the algorithms are relatively simpler. However, when portfolio estimates include all available factors and highly complex ML algorithms, the incremental value of analysts disappears. Figure 1 provides an intuitive plot of the estimated incremental value of analyst-generated information, conditional on the ML technology complexity and the magnitude of the factors space.

[1]The main manuscript reports the results for three algorithms: the linear model, one random forest, and one artificial neural net. Due to space constraints, the results for the other algorithms are summarized in Section 6.1 as a robustness check, whereas the corresponding tables are reported in Section IA.4.2 of the Internet Appendix.

[2]The top three performers are two random forests and one linear model. Two of the artificial neural net algorithms result in an alpha of 15%.

These striking results suggest that publicly available information, if processed through a sufficiently sophisticated data analysis technology, contains all the investment value available to the market. As a consequence, analysts' contributions are fully absorbed by the technology. Differently, if data are processed with naïve techniques, analysts can provide a valuable integration to market and firm-based information. Indeed, we can think of ML algorithms as different representative agents with various capacities to process information. The main takeaway from our findings is that when the agent fails to process public information on their own, the analyst's content has some merit. However, when the agent is sophisticated enough to process a large scale of public predictors, the analyst's content emerges less effective.

The above intuition is further supported by our results on sub-samples before and after the Global Research Settlement (GRS) Act. In fact, prior to 2003 (when sophisticated techniques were largely unavailable), the contribution of analysts was meaningful and significant. Differently, post-GS, their incremental value has diminished to the point of fully flattening. Our results provide two important implications. On the one hand, they point at a potential widening of the gap in access to abnormal returns: only investors who can handle relatively complex techniques can fully exploit the investment value of available information and generate substantial abnormal returns. On the other hand, analysts' input serves as a useful benchmark for price efficiency, particularly for those who cannot employ complex ML methods.

A battery of robustness checks supports the findings and conclusions of our paper. In particular, we utilize recent tools in interpretable machine learning (also known as explainable artificial intelligence) - see Molnar (2020). These tools allow us to conduct sensitivity analysis given the black-box nature of ML complex models. For instance, in the case of a random forest agent, the influence of analyst-related content diminishes gradually over time, displaying cyclicality primarily driven by the recommendation part. On the other hand, the multilayer neural network algorithm reveals that the contribution of analyst content does not diminish over time, suggesting that such technology places a comparatively greater emphasis on such information throughout the sample period. Despite this observation, the baseline model of this algorithm, which relies on market and fundamental data, does not emerge as the optimal performer. Hence, such complexity implies that a more pronounced reliance on analyst-related content might enhance predictive ability, however, only when the baseline model is relatively weak.

Next, we put our experiment to one final test using mean-variance analysis that utilizes the return and volatility forecast made by the seven algorithms and the different data sources studied in our paper. Such analysis underscores the robustness of our main results: while analysts' data may offer some merit in constructing optimal portfolios, the enduring strength

of the market and fundamental data provide investors with a more reliable benchmark, mitigating the risk associated with model selection and potential suboptimal risk-adjusted returns.

Our paper makes two important contributions to the literature. First, we address an important question related to "valuing financial data" (Farboodi et al., 2024). In particular, our model-agnostic ML approach provides a framework to value analyst data from the investors' point of view. On the one hand, the literature finds that analyst content, especially recommendations, is informative about future returns (Barber et al., 2010), while controlling for a large set of firm characteristics. On the other hand, analysts could also attract noise traders, undermining their informativeness (Easley et al., 1998). In this context, we identify an ex-ante optimal and - most importantly - applicable approach that blends different sources of information available to the investment public. Nonetheless, the value of the analysts' information is complex and may require nonlinear technology to utilize its informativeness. While linear technology fails to utilize this information, it still results in a strong benchmark utilizing the baseline data. On the other hand, the marginal value of analysts is more evident when complexity is high, which comes at the cost of a weaker baseline model.

Second, our paper contributes to the literature on portfolio selection under model/estimation risk (Kan et al. (2022)). The identification of nonlinearity, especially for mean-variance investors, could present a valuable layer of insight into portfolio optimization strategies. However, the crucial point emerges that the effectiveness of this nonlinearity is most evident when the feature space excludes market and fundamental data or when such a benchmark is relatively weak, which is evident in only one out of seven models in our investigation. Our analysis, therefore, stresses the opportunity cost facing economic agents in utilizing information and model selection, where they could be worse off by trying to ex-ante select the optimal data and model.

Overall, our ML approach provides insight into how optimal, dynamic information processing across numerous heterogeneous sources has the potential to enhance portfolio construction. The main results of our research stress the lack of relative and absolute value of both recommendations and textual content to ML-aided portfolio formation.

The rest of the paper is organized as follows. Section 2 provides a summary of related literature. In Section 3, we describe the data and provide an overview and summary. Section 4 presents single-sorting strategies and a double-sorting benchmark strategy. In Section 5, we introduce our ML methodology, illustrate its performance, and evaluate the strategy's sensitivity to the different sources of information. We provide additional analysis and results in Section 6, whereas Section 7 concludes. The Internet Appendix (IA) contains additional robustness checks and supplementary results.

2 Related Literature

The literature on the information content of analysts' data and stock returns is extensive and can be split into two sub-streams: quantitative-based (recommendation) and qualitative-based (textual content). Earlier studies by Dimson and Marsh (1984); Barber and Loeffler (1993); Stickel (1995); Womack (1996); Michaely and Womack (1999) evaluate the recommendation-based data in terms of stock return prediction power. Such predictability typically comes from shifts in analysts' opinions about the covered firm. Later work by Barber et al. (2001) provides further support in terms of forecasting power by analysts but undermines whether such predictability can be utilized after taking into consideration transaction costs. While these papers typically support the notion that, on average, stock prices tend to move in the direction predicted by analysts, there remains a question about their predictability in the cross-section of stock returns (Sorescu and Subrahmanyam, 2006). In this regard, Sorescu and Subrahmanyam (2006) show that the analyst's experience and the degree to which the revision is made matters for future stock returns. Their analysis is motivated by behavioral finance in terms of how investors form beliefs for financial decision-making.

Although analysts play a crucial role as information intermediaries in financial markets, evidence suggests biases in their recommendations (see, e.g., Michaely and Womack (1999)), raising concerns about the reliability of information provided by analysts through recommendation forecasts. However, Bradshaw (2011) suggest that their contribution should be assessed in a qualitative manner. Consistent with this conjecture, Huang et al. (2014) utilize a Naïve Bayes classifier to study a large corpus of analysts' reports and find that their content in terms of sentiment has a significant effect around report release dates. Additionally, the authors consider an in-sample predictive regression and show that the value of analyst report text is statistically significant for future earnings growth in subsequent years. Another paper by Caylor et al. (2017) proposes trading a strategy using their analyst-text-based signal and shows that it generates economically significant returns (about 14% excess return over the CRSP benchmark). Further, Chhaochharia et al. (2023) study analysts' sentiment and show that a long-short portfolio utilizing industry-level analyst sentiment corresponds to an annualized alpha of over 7%. Our results do not contradict these findings, nor do theirs contradict ours. We document some value of analysts' text and quantitative data; however, this is primarily when the ML-based representative agent does not process market and fundamental data efficiently. Nevertheless, our investigation differs mainly in focusing on the long-term economic and statistical significance of the analyst data rather than the short-term news-diffusing/reinforcing mechanism.

Our paper also relates to Cao et al. (2024) study comparing forecasts generated by different ML algorithms and analysts, both in isolation and jointly. The authors conclude that a better performance can be attained by combining *man and machine*. Similarly, Chen et al. (2022) provide evidence using machine learning in predicting the future direction of one-year ahead earnings changes, outperforming conventional models as well as analysts. Different from them, we do not focus on analysts' statistical performance per se but rather on the economic value of analysts' data and their marginal contribution from a portfolio selection point of view. Additionally, our paper relates to Gu et al. (2020) in that we build on their data to construct our baseline analysis on which we deploy our empirical strategy using different ML models and feature spaces to parse the value of analysts' data. Our study is also related to Bali et al. (2024), who utilize different ML algorithms to represent economic agents with heterogeneous beliefs. The authors find that their ML agents are considered less biased than analysts and show that greater disagreement among agents translates to economic and statistically significant future returns.[3]

For the *value of data*, our portfolio selection investigation is inspired by Farboodi et al. (2024), allowing us to put the analyst data to an important test and provide implications about the important question of "how much should investors be willing to pay for a stream of analyst data?." Consistent with Farboodi et al. (2024), we find that the answer depends on the sophistication of the ML-representative agent, leading to differing valuations for the same data. Related to this stream of literature is the recent work by Cakici and Zaremba (2024), who investigate the value of market versus accounting data in stock return prediction and find that the former plays a more critical role. Additionally, Ke (2023) raises a question about the use of data mining in terms of price efficiency, i.e., the *value of data mining*. The author argues that such practice could result in greater complexity, undermining the efficiency channel. On the other hand, Kelly et al. (2022) emphasize the appeal of complexity in terms of return predictability, providing empirical evidence supporting the effectiveness of complex models.

While previous studies predominantly underscore the predictive performance of analysts' data in terms of future financial performance and stock reaction, our focus shifts to the economic value of this information, particularly within the context of portfolio selection. To the best of our knowledge, our paper is the first to conduct an experiment where the analysts' data is put against state-of-the-art ML algorithms and big financial data (Gu et al., 2020). Additionally, unlike analyses that primarily focus on analysts' statistical performance, we underscore the importance of understanding the economic implications and marginal

[3]A simple premise behind such finding is that greater disagreement among market participants makes stock prices upward biased since pessimistic agents face short-sale constraints and limits to arbitrage (Miller, 1977).

contributions of analysts' data from a portfolio selection standpoint. Finally, our paper also leverages state-of-the-art explainable artificial intelligence tools that help us provide deeper implications and interpretability of the determinants of the value of data.

3 Methodology and Data Sample

In this section, we describe our sample construction and the different data sources used in our analysis. We set our discussion by first describing the analyst data before delving into the data merging process with other conventional databases.

3.1 Analyst Data

For the analyst data, we consider two data sources corresponding to both quantitative and qualitative characteristics. Analysts provide information in two ways: first, they provide summary measures such as recommendations and target prices; second, they elaborate on these summary measures in lengthy report commentaries, the content of which may convey significant additional information captured by the language, tone, and/or use of tables and figures (Huang et al., 2014). To analyze both components of analyst-provided information, we collect sell-side analyst reports obtained from Investext and pair them with stock recommendations collected by I/B/E/S. In the discussion below, we provide further information on the data gathering and processing.

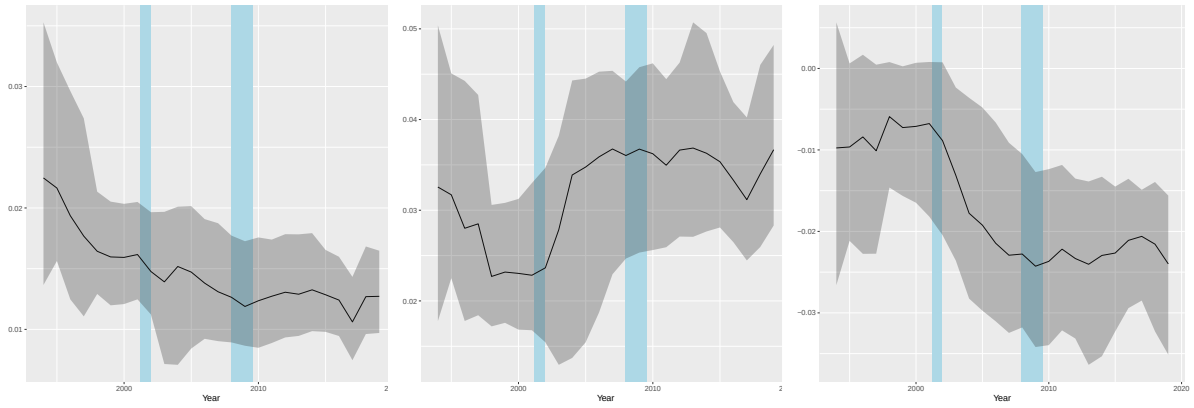
3.1.1 Analyst Reports

We download reports in portable document format (PDF) and convert each report to text using the application pdftotext.[4] If a PDF document does not contain internally searchable text (i.e., the PDF is image-based), we use Tesseract open source optical character recognition (OCR) software to extract text from each analyst report.[5] Investext also provides summary information for each analyst report file, including report title/subtitle, date of release, brokerage name, analyst name, and language. We limit this initial sample only to English language reports, which includes 1,568,002 analyst reports with unique identification numbers covering 2,941 randomly selected U.S. exchange-listed firms from 1984 through 2018.[6]

[4]pdftotext is part of the open source project Xpdf available at http://www.xpdfreader.com/.

[5]Tesseract is available at https://github.com/tesseract-ocr/. Older analyst reports in Investext (i.e., prior to 2000) are more frequently image-based and require OCR for textual analysis. We note that the

Figure 2: **Analyst Report Sentiment**



(a) Positive (b) Negative (c) Z Score

Notes: This figure summarizes the sentiment of sell-side analyst reports in our sample from 1993 through 2018. In all cases, the blue bars denote the recession periods identified with respect to the dates provided by the National Bureau of Economic Research. The solid black lines represent the cross-sectional median over each single year in the data. The gray shaded areas denote the cross-sectional bottom and top 10% percentiles. Panels (a) and (b) demonstrate positive and negative tone, respectively. Panel (c) corresponds to the Z score from Equation (3.1), denoting the net tone, i.e., the number of positive terms minus negative ones adjusted by the total number of words in the document. The positive and negative terms are based on the dictionaries by Loughran and McDonald (2011).

3.1.2 Sentiment and Textual Features

For each analyst report, we gauge sentiment using the standard dictionaries by Loughran and McDonald (2011). In particular, we compute three metrics following their proposed dictionaries: positive, negative, and net tone. The last one is computed as the difference between the number of positive and negative terms adjusted by the total number of words in the document. We note that for each underlying firm, multiple analyst reports are released over any given time period. For instance, for firm i at the end of month t, there are $N_{i,t}$ analyst reports covering the company in that specific month. We map each report covering company i at the end of month t by analyst A into a sentiment score as

$$Z_{i,t,A} = \frac{N^{+}_{i,t,A} - N^{-}_{i,t,A}}{N_{i,t,A}} \tag{3.1}$$

resulting dataset of PDF (converted text) reports is more than 300 (30) gigabytes in size.

[6]As an initial perspective on the analyst data, we provide basic summary statistics in Table IA.3 of the Internet Appendix. For each year, we report the total number of reports, analysts, brokerage firms, and firms covered in the sample. In addition, we compute both the reports-to-firms ratio and the reports-per-analyst ratio. Overall, the table indicates that the volume of available coverage increases significantly over time. This evidence can be attributed to two concurrent factors: the limited availability of very old reports and the endogenously lower research intensity at the beginning of our sample. Additionally, Section IA.2.2 of the Internet Appendix provides a detailed time series analysis of the content of reports.

where $N^{+}_{i,t,A}$ ($N^{-}_{i,t,A}$) denotes the number of positive (negative) words in a single report covering company i at the end of month t by analyst A. Moreover, $N_{i,t,A}$ denotes the number of words in that specific report.

Figure 2 illustrates the evolution of tone over the sample period, along with positive and negative sentiment. We observe that positive sentiment declines steadily until the end of the financial crisis and stabilizes afterward. The negative tone declines in the early part of the sample, rebounding following the tech bubble recession and stabilizing through the following years. For instance, there is an increase in positive tone around 2000 with a simultaneous decrease in negative sentiment. Examining the Z sentiment score in Panel (c), we note a net tone decline over time driven by years between the two recessions from 1993 to 2018.

3.1.3 Additional Dictionaries

Several studies highlight that the tone of analysts' reports might be affected by their goal of forecasting future performance. In particular, Brochet et al. (2015) focus on temporally oriented words, both short and long-term oriented, and develop a dictionary that captures different time horizons. We refer to these dictionaries as "Short-Term" and "Long-Term", capturing the relative number of related terms in the analyst report. Additionally, we consider a list of financially-oriented terms based on conference calls from Matsumoto et al. (2011). The argument is that managers in conference calls are less constrained in providing information relative to financial statements or other press releases. In this regard, we conjecture that these terms also have incremental information content when it comes to analyst reports. We leverage the Matsumoto et al. (2011) dictionary and name it "Financial".

Further, Brown et al. (2015) find that sell-side analysts support their recommendations using detailed valuation models that are, however, highly sensitive to the inputs, assumptions, and methodologies employed. In this respect, they suggest using dictionaries incorporating valuation-related terms (e.g., discounted cash, dividend discount, etc.). Additionally, Painter (2018) and Dessaint et al. (2024) show that when traditional valuation metrics are not applicable or sufficient, analysts incorporate information beyond traditional financial market and accounting metrics that appear to have a significant impact on markets relative to peers. Their results suggest, for example, that investors with relevant satellite imagery data (e.g., parking lot traffic) possess an advantage over other market participants. We follow these contributions and compile two additional dictionaries aimed at capturing the usage of valuation-related words and alternative, unconventional information sources. We name these dictionaries as "Valuation" and "Alternative". In Table IA.1 of the Internet Appendix, we detail the list of terms used in each dictionary and their time series properties.[7]

[7]In particular, we illustrate these dictionaries and other characteristics in Figures IA.1 and IA.2 from

Finally, given the implications of information uncertainty on stock returns (see, e.g., Zhang (2006); Baltussen et al. (2018)), we consider uncertainty related-terms in line with the uncertainty dictionary proposed Loughran and McDonald (2011). In total, we have six text-related features in addition to the sentiment-related dictionaries.

3.1.4 Firm-Month Panel

The release of analyst reports takes place frequently. As a result, the timing of these reports is not constant over time. Additionally, the analyst reports data correspond to a three-level panel. For this reason, at the end of each month, we aggregate the textual content on the firm-month level by taking the average. In this regard, we map the analyst-firm-month panel into a firm-month one. This procedure is similar to analysts' consensus data covering different firms. Moreover, in the event that there were no releases in a single month, we refer to the most recent data point and create a complete month-firm panel.

3.1.5 Recommendations

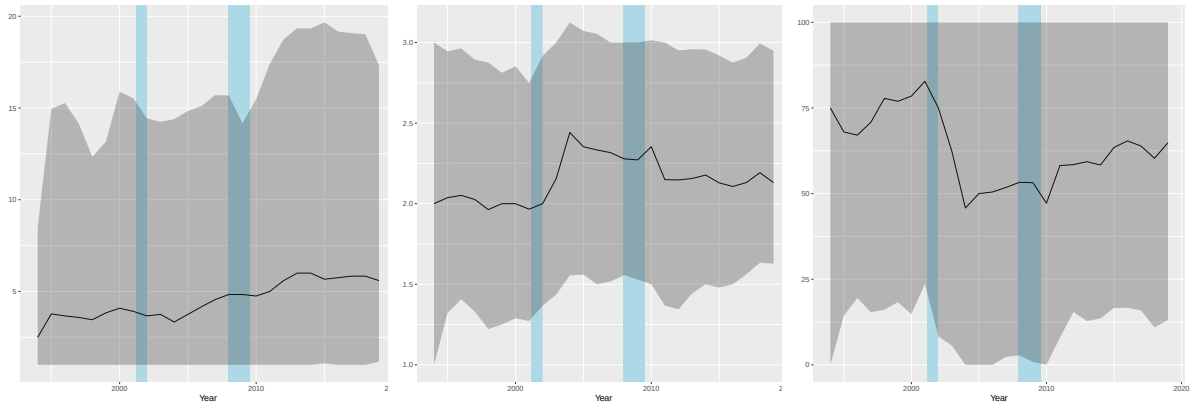
We refer to the I/B/E/S database for analyst consensus recommendations. The recommendation data dates back to Nov 1993. In Figure 3, Panels (a) through (c) plot the report recommendation statistics over our sample period. The intensity of analyst coverage measured as the number of recommendations per firm also increases over time. The median recommendation is 2, corresponding to BUY, until the bust of the dot-com boom, after which the median view becomes more conservative. The over-optimism of analysts during the dot-com bubble is further supported by a sharp drop in the percentage of buy recommendations - falling roughly from 75% at the peak of the bubble to around 50% afterward. We also note that recommendations became more negative during the global financial crisis, but they quickly reverted to more bullish views in the subsequent period.

3.2 Common Data Sources

Our objective is to link the analyst textual reports and recommendations with several common data sources and establish a final panel data. In the following steps, we match analyst report data with a rich set of stock characteristics commonly used in the cross-section of stock returns (Gu et al., 2020).

the Internet Appendix.

Figure 3: **Time Series of Analyst Recommendations**



(a) Recommendations per Firm (b) Mean Recommendation (c) Percentage of Buy Recommendation

Notes: This figure summarizes sell-side analyst recommendations included in our sample from 1993 through 2018. In all cases, the blue bars denote the recession periods identified with respect to the dates provided by the National Bureau of Economic Research. The solid black lines represent the cross-sectional median over each single year in the data. The gray shaded areas denote the cross-sectional bottom and top 10% percentiles. In panels a-c, we present summaries of the recommendations as obtained from I/B/E/S. Panel (a) reports the number of recommendations per firm. Panel (b) reports the evolution of the mean (consensus) recommendation. Panel (c) reports the percentage of buy recommendations out of total recommendations.

3.2.1 Market and Fundamentals

For each analyst report, we extract the firm's NCUSIP identification to link the analyst content with other non-analyst data. We refer to the CRSP database for stock returns and market data. Specifically, we keep data points that correspond to a unique NCUSIP-PERMCO identification. As common in the literature, we focus on common shares alone (i.e., share code (SHRCD) equals 10 or 11) and remove any missing values for prices and returns. We consider two main features from the CRSP database. Those correspond to stock returns and realized volatility. For the latter, we refer to the CRSP daily database and compute the average squared return for each firm on a quarterly basis.[8] As a proxy for size, we consider market capitalization, computed as share price times the total number of shares outstanding.

In addition to the above three market variables, firm fundamentals convey additional information on firm financials at any given time to allow investors a reliable assessment of future stock prices. Such information should be valuable in forming investment portfolios as an alternative or complementing current or past market data. The literature covers an extensive list of potential financial ratios (accounting fundamentals) as potential firm characteristics that matter in the cross-section of expected returns (see, e.g., Chen and

[8]We scale the proxy of realized volatility by $\sqrt{21 \times 3}$ to reflect a quarterly measure.

Zimmermann (2021); Gu et al. (2020)). In this regard, we supplement our data sample using a firm-month panel of 94 characteristics according to Gu et al. (2020). These characteristics correspond to market and fundamental-based variables identified in the literature to provide cross-section predictability.[9]

Earlier research by Ball and Ghysels (2018) runs a horse race between analysts' earnings forecasts and forecasts automatically produced by processing a small set of company fundamentals through a mixed data sample regression method (MIDAS). In unreported ancillary results, we ran our analysis using their much smaller set of fundamentals, which includes changes in inventory, accounts receivable, capital expenditures, selling, general, and administrative expenses (SG&A), and changes in leverage. Since our investigation focuses on predicting the cross-section of stock returns rather than analysts' forecasts, these ratios are less relevant than the more extensive set by Gu et al. (2020). Additionally, some of these variables are less applicable to a subset of the firms in our sample, which includes financials. These caveats notwithstanding, using this smaller set of factors, we identify some value of analysts' data, both text and recommendation, when we deploy complex ML technologies. This result will support our main conclusions, which we cover in greater detail in Section 5.

3.2.2 Final Panel

Since the data sources correspond to either monthly or quarterly frequencies, we aggregate the data at the quarterly level and merge them to create our final firm-quarter panel. Merging all altogether, we obtain a panel covering 89,454 firm-quarter observations ranging between 1994-Q1 and 2018-Q4. This corresponds to 2,333 unique stocks. In Table IA.4 of the Internet Appendix, we provide summary statistics for the final panel. We note that the final panel serves as the primary dataset for our empirical analysis. Specifically, this is most relevant for the ML implementation (Section 5 of this manuscript). Additionally, linked with the monthly CRSP database, the final panel allows us to form portfolios with monthly returns that can be rebalanced on a quarterly/yearly basis. We devote the following section to this discussion.

4 Portfolio Sorts

In this section, we perform portfolio sorting to evaluate future returns with respect to each variable from the final panel. Such investigation serves as a baseline analysis, where we

[9]The data from Gu et al. (2020) is based on open-source code utilizing common data sources such CRSP and COMPUSTAT and is publicly available On Dacheng Xiu's webpage (link to data). For further details about the 94 characteristics, see Table A.6 from Gu et al. (2020).

study the value of each individual stock characteristic in forming portfolios. We begin by evaluating portfolios based on market and fundamentals. Then, we form portfolios using both single and double sorts to investigate the value of analysts' content. In all cases, portfolios are sorted into ten deciles at the end of the year, and their returns are computed using equal weighting (see, e.g., Bali et al. (2017)). Regarding risk-adjusted returns, we price portfolios with respect to the five risk factors by Fama and French (2015) and the momentum factor by Carhart (1997). Additionally, we compute Newey and West (1986) standard errors with six lags to control for autocorrelation and heteroskedasticity.

4.1 Single Sorting: Market and Fundamentals Data

We begin with a simple benchmark strategy that sorts stocks based on previous characteristics covering both market and fundamental indicators.[10] These characteristics contain a wide range of metrics, including but not limited to realized volatility, market beta, momentum, company size, liquidity, CAPEX, ROE, ROA, and various others. As the following analysis serves as a baseline, we report the results in Table IA.5 from the Internet Appendix. The table reports the risk-adjusted returns of the ten portfolios along with a long-short portfolio that longs the high decile and shorts the low one.

Table IA.5 contains the summary of 97 characteristics: the first three (returns, realized volatility, and market capitalization) are added as an additional control for analysis, and the other 94 represent the complete set used by Gu et al. (2020). For brevity, we summarize results only for a subset of relevant characteristics. First, the realized volatility (RV) result is consistent with the risk-based explanation, where stocks with high realized volatility exhibit a higher risk-adjusted return of 16% annually. This difference is statistically significant at the 5% level. Second, we note that small-cap stocks outperform large-cap stocks significantly when we consider either a standard measure market capitalization (MKTCAP) given by shares outstanding times the current price or the one from Gu et al. (2020) (mve1), resulting in an annual risk-adjusted return of 24% per annum. While we control for size risk factor by including the five risk factors from Fama and French (2015), we note that the factor has a minor effect in explaining the returns of the lower decile. This implies that the returns could be potentially attributed to other hidden factors. We return to the issue later in the paper.

Third, among the other characteristics from Gu et al. (2020), we find the following to be statistically significant at the 1% level: dollar trading volume (dolvol with -17%) by Chordia et al. (2001), 12-month momentum (mom12 with -19%) by Jegadeesh (1990), industry-

[10] In total, we have 97 characteristics, where 94 come from Gu et al. (2020) and 3 others that we include as additional control.

adjusted change in asset turnover (chatoia with 8%) by Soliman (2008), depreciation and PP&E (depr with 14%) by Holthausen and Larcker (1992), change in current ratio (pchcurrat with -8%), R&D to market capitalization (rdmve with 14%) by Guo et al. (2006), secured debt (secured with 8%) by Valta (2016), and stock liquidity (ill with 16%) by Amihud (2002). These findings provide mixed support to prior extant evidence. For instance, the result from the illiquidity characteristic is consistent with the original paper results and intuition, where one expects a liquidity premium for trading less liquid stocks. On the other hand, the momentum, for instance, demonstrates that stocks exhibit reversal rather than momentum in our sample, i.e., past 12-month winners (losers) underperform (outperform).

Our baseline findings illustrate that risk premia are time-varying and are difficult to measure (Nagel and Xu, 2023). Many factors have been proposed in the literature to have investment value, generating the so-called "zoo of factors" (see, e.g., Chen and Zimmermann (2021)). In isolation, many of the proposed factors generate alphas in-sample; yet, identifying which factors have out-of-sample investment value is challenging (Harvey and Liu, 2021). Indeed, out of 94 characteristics studied in our baseline analysis, only 8 stand robust in our experiment. This demonstrates the difficulty in forming portfolios ex-ante based on a given characteristic, where choosing a single characteristic from a zoo of factors implies an overfitting problem. In this regard, Gu et al. (2020) propose a data-driven ML approach to combine the 94 factors to form portfolios. We pursue this experiment later in Section 5 while putting the analysts' data to the main test in order to evaluate its marginal value after processing a large set of observable market and fundamental characteristics.

4.2 Single Sorting: Analyst Recommendations

Next, we follow a similar analysis to evaluate whether there is any cross-sectional predictability using analysts' recommendation data. In particular, at the end of each year, we sort stocks into ten portfolios based on each variable from the recommendation set. Similar to the results of the portfolio sorts based on market and fundamental characteristics, Table 1 reports the risk-adjusted returns of each decile as well as the difference between the top and bottom legs. Additionally, we consider three panels. The first panel corresponds to the full sample period and, hence, is consistent with the previous portfolio formation. In the second and third panels, we consider a subset of the full period to distinguish the era following the GRS Act in 2003.[11] In this regard, Panel (b) reports the portfolio returns pre-2003, whereas Panel (c) denotes the post-regulation (i.e., Regulation FD) period.

Starting with mean recommendation (MEANREC), we note that the variable ranges be-

[11]For further information on the effects of the GRS, see Kadan et al. (2008).

tween 1 and 5, where 1 (respectively 5) value denotes strong buy (respectively strong sell). Hence, we expect the top decile to be associated with stocks with weak buy or sell recommendations as we form them into ten groups. Interestingly, stocks in the top decile outperform those in the bottom. The difference between the top and bottom decile results in a risk-adjusted return of 3.3% per annum; however, such difference is statistically insignificant. At the same time, we note that the portfolio risk-adjusted returns exhibit a U-shaped function with a greater skew on the top leg. Notably, a similar pattern emerges when we consider the percentage of analysts with buy recommendations (BUYPCT) for portfolio sorts, whereas the long-short portfolio corresponds to an annual risk-adjusted return of -10%, which is significant at the 1% level.

We observe that stocks with less favorable recommendations exhibit higher risk-adjusted returns for either mean recommendation (MEANREC) or percentage of analysts with buy recommendations (BUYPCT). This evidence supports recent research by Kong et al. (2021) showing that institutional investors treat analyst downgrades as significant and negative signals. In particular, downgrades from "strong buy" to "buy" and from "buy" to "sell" are associated with a significant net-selling reaction. From a risk-based point of view, such evidence indicates that investors require higher compensation for holding stocks with less favorable analyst recommendations.

The second row of Table 1 shows that stocks with greater dispersion among analysts' recommendations (STDEV) are associated with lower risk-adjusted returns. In particular, longing stocks in the top decile while shorting those in the lower one results in a -10.7% annual mean return that is significant at the 1% level. One potential explanation is that stocks with greater dispersion are associated with higher uncertainty. Such uncertainty could also be attributed to smaller firm size and, hence, STDEV captures a size effect. Indeed, this is consistent with the evidence for the number of analysts (NUMREC). Another explanation is that such uncertainty is priced. According to the limited market participation argument (see, e.g., Baltussen et al. (2018)), stocks with higher uncertainty are traded by more informed investors who require a lower uncertainty premium.[12]

Finally, comparing the pre- and post-regulation periods, we note that risk-adjusted returns of the long-short portfolios exhibit stronger economic and statistical significance in the pre-2003 period. For instance, when sorting stocks based on the mean recommendation, the long-short portfolio corresponds to 11% risk-adjusted return per annum before 2003. On the other hand, post-2003, the risk-adjusted return is about 1%. One potential explanation is

[12]More recently, Lundholm and Rogo (2020) compare trading rules based on analysts' forecast dispersion and show that higher dispersion is associated with lower risk-adjusted returns - proxied by Sharpe ratios. The authors attribute this evidence to a behavioral bias, where analysts "curry favor with management or to differentiate themselves from other analysts."

the greater sell-side analyst conflict of interest that preceded the regulation era. In this regard, analyst recommendations may have increased demand for strong buy recommendation stocks, bidding their prices up relative to weak recommendation stocks, the result of which exacerbated the difference between the top and bottom deciles.[13]

4.3 Single Sorting: Analyst Textual Content

As in the previous analysis, we sort stocks into ten deciles based on each text-related feature at the end of each year and assess their portfolio returns over the following twelve months. We report the corresponding results in Table 2 with respect to each of the seven textual features. We summarize the results as follows. First, we observe that a self-financing portfolio that longs stocks with the highest net tone while short those with the lowest net tone results in a 3.4% annual risk-adjusted return; however, this result is statistically insignificant. Nonetheless, post-2003 period (i.e., post-GRS), the risk-adjusted return of this long-short portfolio increases to 5% and becomes statistically significant at the 5% level. On the other hand, in the pre-GRS period, we observe that the alpha is roughly zero. One possible conjecture is that in the aftermath of the settlement, the tone of the analysts is more objective and, hence, more aligned with actual stock performance.

Second, uncertainty emerges as the only dictionary that provides strong statistical evidence among the other five textual features. In particular, a higher text-related uncertainty is associated with lower future risk-adjusted returns, such that longing the bottom decile whole shorting the top one results in an annual risk-adjusted return of 7.8%. This evidence aligns with previous results obtained by sorting stocks based on the dispersion in recommendations (STDEV). Additionally, the economic significance appears consistent both before and after the GRS period, although the statistical significance is only evident in the post-GRS period.[14]

4.4 Double Sorting: Analyst Recommendations and Net Tone

The results from single sorting illustrate mixed evidence in terms of investment value stemming from the analyst data. In the following analysis, we consider the interaction be-

[13]Figure IA.3 in the Internet Appendix visualizes the risk-adjusted returns of two long-short portfolios: one based on analyst mean recommendation (MEANREC), and the other is based on recommendation dispersion (STDEV). While the figure captures a consistent trend over time for each long-short portfolio, the trends' slopes are of higher magnitude during the pre-2003 period, an observation that is compatible with the results from Table 1.

[14]Figure IA.4 in the Internet Appendix shows that while the net tone's alpha has been steadily increasing since the post-2003 period (i.e., post-GRS), the long-short portfolio based on the uncertainty measure displays a negative trend over the sample period, with a slight upward kink during the Great Recession.

tween analysts' recommendations and net tone. We pursue this investigation since it is well-known that recommendations serve as a coarse proxy of an analyst's view on a company and, consequently, may be accompanied by divergent sentiment captured by tone. For instance, a BUY recommendation might be issued with a lukewarm report (Hirshleifer et al., 2024), whereas an otherwise similar BUY might be supported by an enthusiastic view of the company. Additionally, Huang et al. (2014) document that only 34% of BUY recommendations in their sample are issued with overall positive report sentiment, and a non-trivial 12% are characterized by a negative view. In this respect, it is reasonable to expect that selecting stocks through a joint weighting of recommendations and contemporaneous report tone might yield superior results.

We design a double-sorting scheme that jointly considers these two factors, aligning with the self-financing strategy used in the previously conducted single-sorting portfolio analysis. Specifically, we implement a dependent double sorting based on recommendations first and net tone second. Initially, firms are grouped into terciles based on their mean recommendation (MEANREC). Within each recommendation group, firms are then ranked according to their net tone, resulting in nine recommendation-tone groups. We calculate the equal-weighted return for each cluster over the subsequent twelve months, producing nine time series of portfolio returns. The risk-adjusted mean returns for each portfolio are presented in Table 3. Additionally, we report the return differences between strong and weak recommendation groups within each tone category, as well as the return differences between low and high sentiment within each recommendation group. All returns are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with six lags, consistent with the single-sorting analysis.

Among all recommendation terciles, we observe that higher net tone is associated with higher risk-adjusted returns. However, this significance is most pronounced within the weak recommendation group and is primarily driven by the post-2003 period. For instance, a long-short portfolio that buys stocks in the weak-high group while shorting the weak-low portfolio yields a 6.1% risk-adjusted return per annum in the post-2003 period. This result is significant at the 5% level. Moreover, we observe that firms ranked on the high net tone tercile across the strong to weak recommendation groups produce a 6.2% risk-adjusted return per annum, which is significant at the 1% level. There is a monotonic increase in the risk-adjusted return of the self-financing portfolio (W)-(S) as we move from low to high net tone. In other words, the 6.2% alpha becomes weaker within the lower net tone terciles. At the same time, this alpha seems to be driven mainly by the pre-2003 period, with weaker evidence post-2003.

4.5 Challenges in Portfolio Sorting

Before we proceed to the next section, which presents the primary analysis and findings of our research, we address a key caveat associated with the portfolio sorts conducted thus far. Specifically, the use of ex-post realizations to form portfolios and identify strategies that maximize returns is prone to look-ahead bias, thereby offering limited out-of-sample guidance. Additionally, some of these strategies lack economic intuition and are not considered investable. To illustrate this point, we identify *corner portfolios*, due to their position at the corner of Table 3, obtained by an arbitrage strategy that longs both the weak-high stocks and shorting the strong-low. Such a strategy results in an 8.3% annual risk-adjusted return over the whole sample period, which is significant at the 1% level, as Table IA.7 of the Internet Appendix demonstrates.

This example highlights the difficulty of forming a robust ex-ante strategy given the richness and complexity of our dataset, which includes numerous features. While we have explored various portfolio sorts that provide mixed statistical and economic evidence, these approaches are inherently limited to univariate and bivariate sorting due to the challenges associated with high-dimensional sorting (Cong et al., 2022). The combination of mean recommendation and net tone is just one possible pair among many features, but determining the "optimal pair" ex-ante is not feasible and often results in data mining with limited future guidance. We address this issue in the following section, which constitutes the primary contribution of our paper.

5 Machine Learning-Based Information Aggregation

In this section, we move from the descriptive analysis of Section 4 to predictive analyses. We refer to Section 4 as descriptive since it remains unclear how to *a priori* select a single characteristic or a combination of which to form portfolios accordingly. The previous analysis demonstrates that, when considered individually, each source of information offers some degree of investment value. While it is reasonable to assume that investors are aware of this and likely form their strategies by combining multiple information sources, the exact process of aggregation remains unknown - and, more importantly, it varies widely across investors.[15] A potential solution to this challenge is to adopt an artificial intelligence approach. For example, an ML algorithm could identify an optimal *pro-tempore* portfolio, continuously

[15]For instance, superior strategies may be obtained by counterintuitive aggregations, as the corner portfolios exercise from Table IA.7 illustrates.

adapting it to the arrival of new and heterogeneous information.[16]

5.1 Machine Learning Design

Having recognized the limitations of traditional portfolio sorting methods and the potential of ML to overcome these challenges, we now shift our focus to the design and implementation of our proposed ML framework. We develop a predictive model tailored to high-dimensional data, through which we propose an ML decision rule to form equal-weighted portfolios and evaluate their performance out-of-sample. This predictive approach allows us to measure and test the significance of the incremental value of heterogeneous data types. Consistent with the previous analysis, we cluster the full set of variables into three main categories: market and fundamentals, analyst recommendations, and analyst textual content. The core idea behind supervised ML is to map these indicators into a representative index, a process that requires supervision in terms of target (response) variables. To determine the optimal mapping, we consider two targets: the realized stock return over the next year and the realized volatility over the next year. We are mainly interested in forecasting stock returns and forming portfolios accordingly, in line with Gu et al. (2020). However, we also consider realized volatility to conduct additional robustness checks where we value the data using mean-variance portfolio rules that rely on both return and volatility forecasts.

5.2 Optimal Mapping

The first major issue concerning our research questions is how to optimally aggregate information from the investment point of view. A simple decision rule groups firms based on a single or double sort as in Section 4. For instance, consider the case of realized volatility (RV). If the investor assumes that the current RV is associated with the next year's RV, then sorting portfolios based on RV provides an appropriate simple decision rule. A more sophisticated decision rule, however, could involve deploying a predictive model for RV using a variety of indicators. This model would generate ex-ante forecasts, allowing investors to sort stocks into different portfolios based on these predictions.[17] In this regard, the predictive model approach underscores the value of employing ML in our investigation to assess the incremental value of the diverse data sources analyzed in this paper.

The following discussion further details the mapping process and the different algorithms

[16] In recent work, Cong et al. (2022) utilize a new class of tree-based models (P-Trees) to address the problem of high-dimensional sorting. Their analysis, however, does not investigate the value of data (specifically analysts), which is the objective of our research.

[17] For instance, a similar analysis is demonstrated in Table 7 from Gu et al. (2020), where stocks are sorted into deciles using an ML forecast that integrates all 94 features.

studied in our paper. We introduce some notation, where we let $\hat{f}_y^M$ denote a predictive function $M \in \{1, 2, 3, 4, 5, 6, 7\}$ at the end of year y. The predictive model is a fitted function that maps a set of predictors $X_y(M)$, available at the end of year y, into next year's return or RV denoted, respectively, by $\hat{\mu}_{y+1}$ and $\hat{\sigma}_{y+1}$. The $\hat{\cdot}$ denotes that the value is an estimate (fitted value) rather than the true one. On the other hand, μ_{y+1} and σ_{y+1} correspond to the true/realized target values. The learning process, hence, identifies a mapping function (technology) $\hat{f}_y^M$ based on the predictors and the target variables.

In terms of technology, we deploy several nonlinear algorithms such as the support vector regression (SVR) model with radial kernel (Drucker et al., 1997), random forest model (RF) (Breiman, 2001), and three different neural nets (NNET) models (Ripley, 2007). Additionally, we consider one linear model with an elastic net (LMNET) penalty (Zou and Hastie, 2005). The linear model serves as a simple benchmark to evaluate value ML complexity in processing the different data sources. Each algorithm requires tuning to choose an optimal mapping function, which we determine using cross-validation. In all specifications, the model results in an out-of-sample score for each stock available at the end of December of year y. We dedicate the discussion below to describing the training process and the different algorithms employed in our analysis.

5.2.1 Training and Cross-Validation

Our training methodology comprises four layers of computational complexity. Initially, training is conducted using an annual rolling sample. Next, cross-validation (CV) is performed for each rolling training sample. This process is then replicated across seven feature spaces. Finally, multiple ML algorithms are applied to execute the entire procedure. Each layer is discussed in detail below.

In the first two layers, we utilize a rolling window approach. Starting with a four-year observation sample (training sample), our first fitted model is generated by the end of December 1997, marking the initial cross-sectional forecast. Subsequently, we shift the window forward by one year (December 1998) and repeat the procedure. This process continues until 2017, producing a sequence of out-of-sample cross-section forecasts for the years $y = 1998, ..., 2018$. Hyperparameter tuning plays a crucial role in this process. To mitigate overfitting concerns, we implement a 10-fold CV design with three repetitions. In particular, the CV procedure randomly groups the training sample into 10 clusters using a fixed seed. Beginning with the first cluster, we train the model on the other nine groups, testing predictive power on the first cluster using a hyperparameter grid.[18] This process is

[18]In all cases, the CV is attained by minimizing the mean-squared error (MSE) between the true values and the fitted ones.

repeated for each subsequent cluster until the last one is tested. Due to the random nature of cluster selection, we repeat the entire CV procedure three times to determine the final optimal hyperparameters.[19]

Our analysis consistently applies the above process across all feature spaces and ML algorithms. For a given ML algorithm, we consider seven feature spaces denoted by $M = 1, 2, 3, 4, 5, 6, 7$, where each M corresponds to a specific feature set:

1. When $M = 1$, the feature space $X_y(1)$ consists of the market and fundamental characteristics.
2. When $M = 2$, the feature space $X_y(2)$ consists of $X_y(1)$ plus analyst recommendation data.
3. When $M = 3$, the feature space $X_y(3)$ consists of $X_y(1)$ plus analyst report textual content.
4. When $M = 4$, the feature space $X_y(4)$ combines all the above variables, encompassing market, fundamental, recommendation, and textual data.
5. When $M = 5$, the feature space $X_y(5)$ consists solely of analyst textual data.
6. When $M = 6$, the feature space $X_y(6)$ consists solely of analyst recommendation data.
7. When $M = 7$, the feature space $X_y(7)$ combines analyst textual and recommendation data.

Note that $M = 1$ stands for the baseline feature space that does not rely on any analyst data, while the other specifications either focus on analyst-specific data or combine it with the baseline features. For instance, specifications $M = 5, 6, 7$ allow us to evaluate the standalone value of analyst data, whereas $M = 2, 3, 4$ assess the marginal contribution of this data to the comprehensive market and fundamental feature space.

5.2.2 ML Algorithms

RF: The first algorithm is Breiman's (2001) RF model that creates a collection of decision trees, each trained on a slightly different subset of the data. This ensemble approach enhances the model's accuracy and robustness. During training, the algorithm introduces randomness by considering only a subset of features at each decision point in each tree. This helps prevent overfitting and ensures the model can generalize well to new data. Its performance and popularity are well-documented in the literature (see, e.g., Gu et al. (2020); Cao et al.

[19]For all ML algorithms and specifications, the implementation, training, tuning, and testing are conducted using the caret R library by Kuhn et al. (2008).

(2024)).[20] While the algorithm relies on three hyperparameters, it works well with default parameters. The first hyperparameter captures the number of trees sampled. For each tree, the algorithm considers a sub-sample of size N_S and a sub-feature space of size m, corresponding to the second hyperparameter. We tune both hyperparameters using the CV procedure described in Section 5.2.1.[21]

ANN: The second algorithm is the Multilayer Perceptron (MLP) that represents an artificial neural network (ANN), originally introduced by Rumelhart et al. (1985). The MLP denotes an ANN that consists of multiple layers of nodes (neurons), each connected to the nodes in neighboring layers. This architecture allows the model to capture complex relationships within the data. During training, the MLP adjusts the weights on these connections to learn patterns and representations that optimize predictive performance. The hyperparameters of the MLP, which we fine-tune for optimal results using the CV procedure, include the number of hidden layers, the number of neurons in each layer, and parameters that control the learning rate and regularization. By leveraging the flexibility of the MLP, we aim to capture complex patterns in the analysts' and market/fundamental data, providing a sophisticated and adaptive model for our research investigation. As additional robustness, we consider two supplemental ANN-based algorithms. One relies on a single layer denoted by NNET, whereas the other runs principal component analysis (PCA) on the data before feeding into the NNET algorithm.

LMNET: The third algorithm we consider is a linear model denoted by LMNET (Zou and Hastie, 2005). This model is simpler and less complex than the algorithms mentioned above, as it produces a linear mapping of the original data into the fitted values. Nonetheless, the model is also subject to the choice of two hyperparameters determined using CV. The main idea behind the model is to find an optimal mapping of the current data into future response values based on linear mapping while addressing two major issues. First, what features should be included in the model (first norm)? Second, how much should we shrink the coefficients of these features toward zero (second norm)? The combination of these two elements constitutes elastic net regularization, hence the name of the model. While the LMNET model is still linear and is associated with lower complexity relative to the above nonlinear models, its statistical inference is less transparent than an ordinary least

[20] The algorithm has also gained huge popularity in terms of predictive power. For instance, RF is one of the most used algorithms in Kaggle competitions (see, e.g., https://www.kaggle.com/code/faressayah/decision-trees-random-forest-for-beginners).

[21] As additional robustness, we consider an RF algorithm with predetermined hyperparameters. For the first, we fix the number of trees to be 1000. For the other, the value of m is typically considered less than the square root of the total number of features, which we set $m = 10$ across all specifications given that the non-analyst feature space covers roughly 100 firm characteristics. This rule results in a more efficient predictive function with lower variance (James et al., 2013).

squares (OLS) model. We follow the same CV procedure from Section 5.2.1 and reproduce the same empirical design using this model. We conjecture that due to the complexity and heterogeneity of our data sources, the LMNET model fails to capture complex dynamics compared to a nonlinear technology such as RF or ANN. This hypothesis aligns with recent work that employs ML models for forming analysts' forecasts, such as price targets (Cao et al., 2024).[22]

5.3 Hypothesis

For a given ML algorithm, there exists an optimal mapping/forecast for the model feature space M. Each optimal mapping corresponds to a portfolio decision rule that enables us to ex-ante sort stocks into ten deciles. This sorting scheme corresponds to the same univariate sorts from Section 4, allowing us to evaluate the difference between the high and low legs based on the ML predictions, corresponding to the economic performance $\pi(M)$ of feature space M. Formally, our main hypotheses can be stated as follows:

$$\begin{cases} H_0 & \pi(M) = \pi(1) \\ H_1 & \pi(M) \neq \pi(1) \end{cases}, \forall M = 2, ..., 7. \tag{5.1}$$

The null hypothesis indicates that specification $M \in 2, ..., 7$ offers no incremental value relative to the baseline one ($M = 1$), which relies solely on market and fundamental data, i.e., non-analyst data. The $M = 2$ and $M = 3$ specifications test the incremental value of the analyst recommendations and textual report content, respectively. The $M = 4$ specification tests the incremental value when combining both analyst data types with the non-analyst data. Finally, specifications $M = 5, 6, 7$ focus on the marginal value of analyst data alone, without the baseline market and fundamental data. For instance, the double-sorting portfolios indicate that a combination of mean recommendation and net tone yields a risk-adjusted return of 8% per annum (see, e.g., Table IA.7). This states that there are potential synergies among the features that help decision makers to sort stocks with complex characteristics and, hence, form portfolios accordingly. We put this hypothesis to the test given the diverse set of ML algorithms examined in our analysis.

[22] In particular, Cao et al. (2024) demonstrate that complex ML technologies such as RF capture important nonlinear and intertwined features that baseline models such as OLS or LMNET fail to detect.

5.4 Main Results and Findings

In the following, we present the main results based on the three ML algorithms described in Section 5.2.2. Let $\mathcal{A}$ denote the set of all ML algorithms considered in our experiment. For a given algorithm $a \in \mathcal{A}$ and feature space M, we group stocks at the end of each year into ten deciles based on the corresponding mapping function. Consistent with univariate analysis from Section 4, we report the results using portfolio formation with risk-adjusted returns. Additionally, we evaluate a specification that examines the long-short portfolio across different features M for the same algorithm model.

5.4.1 RF Results

We begin our discussion of the RF results summarized in Table 4. The rows correspond to models $M = 1, 2, ..., 7$, whereas (H)-(L) denotes the difference in portfolio risk-adjusted returns between the high and low deciles. Finally, column $\pi(M) - \pi(1)$ computes our main statistic of interest that evaluates any enhancement of ML technology using the different feature spaces of $M = 2, ..., 7$ in line with our hypothesis from Section 5.3. Starting with row $M = 1$, we observe a monotonic increase in both risk-adjusted returns as we move from the bottom decile to the top one. Based on market and fundamentals features (model $M = 1$), the difference between the top and bottom deciles corresponds to a risk-adjusted return of 18.6% per annum, which is significant at the 1% level. This result establishes a relatively high benchmark as we proceed to evaluate the contribution of the analysts' data.

When we incorporate analyst recommendation data on top of the market and fundamental data (row $M = 2$), the ML technology provides better discriminative power compared to the baseline one, as highlighted by a larger risk-adjusted return on the top decile but not so much at the bottom level. In particular, the difference between the top and low deciles results in an annual risk-adjusted return of 21.2% per annum, which is significant at the 1% level. This result indicates that including analyst recommendations provides some marginal improvement. To price the incremental value of the analysts' recommendations, we compute the statistic $\pi(2) - \pi(1)$, which denotes a difference in differences. While this difference corresponds to an improvement of 2.7% annual risk-adjusted return, it is statistically insignificant.

In the following stage, we evaluate the incremental value of analyst report textual data on top of market and fundamental data. Overall, we find no evidence of the marginal value by incorporating the textual data on top of the baseline characteristics. On the contrary, we observe that the (H)-(L) results in a lower risk-adjusted return than the baseline model. This is evident from the $\pi(3) - \pi(1) = -1\%$ column, which is statistically insignificant.

This result indicates that the analyst report textual data does not provide a considerable value compared to analyst recommendations as in model $M = 2$. Additionally, when both analyst recommendations and reports are included in the ML mapping, we do not observe any meaningful positive synergy, resulting in no significant incremental improvement as the difference $\pi(4) - \pi(1)$ is statistically indistinguishable from zero. The mapping technology based on the feature space of $M = 2$ results in the highest risk-adjusted return for the long-short portfolio, even though the 2.7% improvement from the recommendation component is statistically insignificant.

Finally, we test the analyst data independently to assess whether it offers any marginal value without relying on the rich feature space of the market and fundamental characteristics. This is reflected in rows $M = 5$, $M = 6$, and $M = 7$. In all cases, we observe that the ML mapping process does not yield any incremental contribution, i.e., there is no monotonic increase in the risk-adjusted returns as we move from the short leg to the top one. Indeed, the risk-adjusted return of the (H)-(L) portfolio among these three specifications is statistically indistinguishable from zero. Additionally, the $\pi(M) - \pi(1)$ column indicates both statistically and economically significant reduction in value compared to the baseline model, which excludes analyst data. In summary, the findings of the RF suggest that the inclusion of analyst data does not provide a meaningful improvement. The marginal or negative impact of analyst data on predictive power suggests that, in this framework, market and fundamental features play a more crucial role in generating positive returns.

5.4.2 MLP Results

Next, we discuss the findings based on the MLP model, one of the three ANN models considered in our research. As with the previous model, we present the results of the MLP model in Table 5. Notably, the baseline model of $M = 1$ results in a lower risk-adjusted return of the (H)-(L) portfolio (11%) compared to the RF algorithm. Additionally, the economic significance of the long-short portfolio is weaker, even though the result is significant at the 10% level. In line with the insight from the previous model, we discern that integrating the analyst recommendation data into the baseline model ($M = 2$) within the MLP framework results in a marginal improvement of 2.6%; however, this improvement is statistically insignificant. A similar observation follows when adding the textual data on top of market and fundamental data ($M = 3$), with a 2.1% statistically insignificant improvement.

Combining both analyst recommendations and reports within the MLP framework ($M = 4$) shows a synergistic gain, as $\pi(4) - \pi(1)$ indicates some economic significance of 4% annually, though it remains statistically insignificant. When extending the analysis to the cases where the MLP model relies solely on analyst data ($M = 5, M = 6, M = 7$), the MLP

algorithm exhibits some tangible contribution when training it based on analysts' recommendation data alone, resulting in an annual risk-adjusted return of 8.5% for the (H)-(L) portfolio. However, this return is still lower than that of the baseline model ($M = 1$). Overall, the $\pi(M) - \pi(1)$ column for $M = 5, 6, 7$ does not indicate any statistical and economic significance, a result that is consistent with the RF model. Overall, the $\pi(M) - \pi(1)$ column for $M = 5, 6, 7$ does not reveal any statistical or economic significance, consistent with the results from the RF model. The MLP model demonstrates some merit when including analysts' data alongside market and fundamental data, but it is important to note that the baseline model's performance is relatively weak. This outcome suggests that the artificial representative agent may fail to efficiently process market and fundamental data, which leads to a marginal benefit from adding analysts' information. Indeed, when we consider the robustness checks in Section 6.1 using two additional ANN algorithms, we find that when the agent processes the market and fundamental data effectively - yielding strong results for the (H)-(L) portfolio - there is no statistical or economic evidence supporting the inclusion of the analyst data.

5.4.3 LMNET Results

Finally, we repeat the previous tests using a linear mapping function, as summarized in Table 6. Despite the linear nature of the LMNET algorithm, the results are largely consistent with those of the RF model. Across the first four specifications ($M = 1, 2, 3, 4$), we observe that the model barely integrates any information from the analysts, resulting in roughly the same risk-adjusted returns across all deciles - and the same goes for the long-short portfolios. Nonetheless, the model can process the market and fundamental information efficiently, resulting in a 17.9% annual risk-adjusted return for the (H)-(L) portfolio. Given the popularity of the characteristics considered in the feature space of $M = 1$, this evidence suggests that the LMNET is capable of identifying time-varying risk-premium based on these observable features, resulting in a relatively high benchmark.[23]

Prior evidence from Gu et al. (2020) demonstrates the virtue of complexity (Kelly et al., 2022). Our results based on the above three algorithms offer mixed evidence. On the one hand, the RF model delivers the highest performance, as measured by the risk-adjusted return of the long-short portfolio when $M = 1$. On the other hand, the linear model of LMNET ranks second. One potential explanation is that our analysis employs a rolling window by retraining the ML algorithms at the end of each year and rebalancing the portfolios annually, whereas Gu et al. (2020) rebalance the portfolios monthly even though the models

[23]In fact, among the seven ML algorithms studied in our paper, LMNET ranks third after two RF algorithms in terms of the (H)-(L) risk-adjusted returns using the baseline feature space of $M = 1$.

are retrained annually. Our approach implies some trade-offs. On the positive side, it allows us to determine a data-driven mapping technology that is not too sensitive to the hyperparameters tuning. In this regard, our analysis can be considered as a brute force where the same systematic training is applied to each model. However, on the downside, training each model differently, as in Gu et al. (2020), provides greater flexibility, potentially leveraging model complexity and nonlinearity more effectively.

6 Further Analysis

In the following section, we discuss additional analyses in support and extension of our main findings. These results cover several angles. First, we perform robustness checks using additional ML algorithms. Second, given the black-box nature of the nonlinear ML models used in our analysis, we consider recent advancements in interpretable ML to gain deeper statistical insights into the important features (if any) that drive the marginal contribution of the analysts' data. Finally, we put the analysts' data to a final test, where we leverage the ML prediction results to form mean-variance portfolio rules that depend on both return and volatility forecasts.

6.1 Additional Algorithms

We consider four additional ML algorithms as a robustness check. These four can be viewed as a simplification of the RF and ANN models. The first ML algorithm is support vector regression (SVR) (Drucker et al., 1997), which identifies certain features in the data that can distinguish the response variable's attributes within a certain confidence level. The confidence level is also known as the margin, which provides a certain confidence that the model is not overfit and is robust against outliers. Additionally, the algorithm deploys a radial basis function kernel that maps the original data into a new space, where the attributes of the response variable can be better identified. In this regard, one can think of the SVR model with a radial kernel as a plain vanilla artificial neural network in which the original data are mapped into a new set of hidden features (neurons) that are subsequently mapped into the output variable (e.g., future realized volatility).

The second specification is an RF algorithm with predetermined hyperparameters, where we set the number of trees to 1,000 and the number of sampled features to 10. The latter value is consistent with the rule of thumb used in the machine learning literature, which corresponds to the floor of the square root of the total number of features (James et al.,

2013).[24] We denote this model by RF2. This is consistent with the RF model from Table 4, which relies on cross-validation for hyperparameters tuning. For the other two specifications, we explore two ANN algorithms: one is a single-layer ANN, referred to as NNET, and the other is similar but incorporates PCA before feeding the data into the model. We refer to the latter algorithm as NNETpca.

For brevity, the additional results are presented in Section IA.4.2 of the Internet Appendix. Specifically, the results of the SVR, RF2, NNET, and NNETpca algorithms are reported in Tables IA.8, IA.9, IA.10, and IA.11, respectively. We summarize the main findings as follows. First, we find no statistical evidence across the three models in support of the incorporation of analysts' data, i.e., $\pi(M) - \pi(1) > 0$ for $M = 2, 3, 4$. In fact, for the NNETpca model, we find that the analysts' recommendation worsens the performance of the (H)-(L) portfolio. Second, we conclude that the RF algorithm with the predetermined hyperparameters yields the highest performance when evaluating the long-short portfolio of the baseline feature space, resulting in a 21.9% annual risk-adjusted return.

Overall, these additional results, combined with the main findings from Section 5.4, carry important implications. The lack of a tangible positive contribution from analyst data, even within more complex ML models, suggests that traditional analyst content, whether recommendation or textual-based, may not consistently provide a robust advantage in the long term. This challenges existing evidence that expert opinions in the form of analyst information significantly enhance predictive models. This is highly relevant given the widespread availability of rich public data, as utilized in studies like ours and that of Gu et al. (2020). Furthermore, the superior performance of the simplified RF model (i.e., RF2) with predetermined hyperparameters highlights the importance of model simplicity and feature space relevance. It suggests that, in certain contexts, a more straightforward approach might outperform more complex models. The performance of the LMNET model reinforces this notion, as it outperforms all ANN algorithms.

6.2 Sensitivity Analysis

One major challenge in the application of ML models is their interpretability. For instance, ANN-based models, such as the MLP, often function as black-box technologies. These models transform the original data into a new feature space through multiple layers of nonlinear mappings, ultimately relying on features that are not directly observable or interpretable. While the results of the MLP (see Table 5) show some improvement by incorporating sell-side analyst data, the key drivers behind the model remain unknown. Specifically, it is unclear

[24] Note that we have more or less than 100 features across the $M = 1, 2, 3, 4$ specifications.

how different sources of information contribute to the model's performance over time. Understanding this is crucial, as it provides deeper insights into how information flows to the market under varying exogenous conditions.

In the following investigation, we deploy an ML-specific sensitivity analysis that helps us to determine the key features (variables) that influence the mapping function. This approach provides a managerial overview of ML algorithms and their inner workings. To serve this objective, several techniques have been proposed in the literature. One common approach builds on identifying the important variables using data shuffling, a technique known as "Permutation Feature Importance" (see, e.g., Fisher et al. (2019)). In this method, a feature is considered "important" if shuffling its values reduces out-of-sample performance. On the other hand, a feature is "unimportant" if shuffling its values results in no effect on the predictive function. Thus, if a feature's contribution is marginal, the model likely disregards it in the mapping process, whereas a meaningful contribution indicates that the feature plays a significant role.

We leverage the implementation by Biecek (2018) to investigate variables' importance using a couple of models and their impact on the out-of-sample fit, measured using R^2. We focus on two models of interest: RF and MLP. Among the three models summarized in Section 5.4, RF yields the highest performance in terms of the baseline model, whereas MLP demonstrates the largest marginal improvement by incorporating both analyst recommendation and textual content. In both cases, we focus on feature space $M = 4$ that includes all data sources. For each year, the sensitivity analysis computes an importance score for each variable defined as follows:

$$S_{y,v} = \frac{R^2 - R^2(v)}{R^2}, \tag{6.1}$$

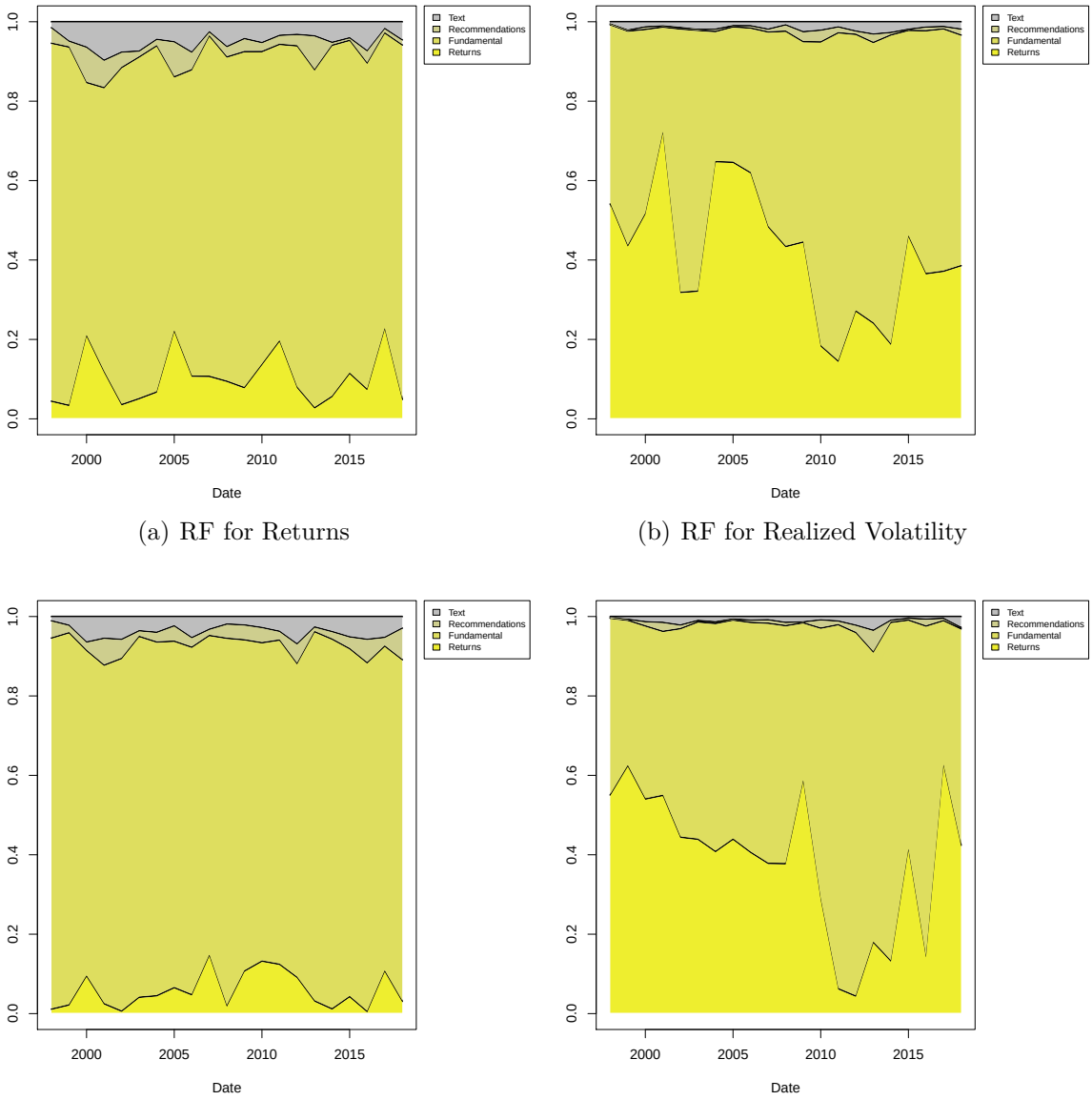
with $S_{y,v}$ denoting the sensitivity of the out-of-sample R^2 to permutation to variable v at year y. Here, R^2 proxies the out-of-sample prediction power of the full model, whereas $R^2(v)$ denotes the out-of-sample R^2 when variable v is shuffled with noise.[25]

In Figure 4, we illustrate the relative importance of each variable using the sensitivity score from Equation (6.1). In particular, we aggregate the weighted score at four different group levels: our three market features, the 94 features from Gu et al. (2020), our analysts' recommendation features, and our analysts' related textual features. The results are illustrated based on feature space $M = 4$, which covers all four categories for both RF and MLP algorithms.[26] Additionally, the analysis covers both return and realized volatility predictions.

[25]In Table IA.12 of the Internet Appendix, we report summary statistics of $S_{y,v}$ for the top 20 features, which are sorted based on the average sensitivity score when the response variable is the future stock return.

[26]The results are consistent with those reported in Table IA.12 from the Interest Appendix.

Figure 4: **Statistical Performance: Features Importance over Time**



(a) RF for Returns

(b) RF for Realized Volatility

(c) MLP for Returns

(d) MLP for Realized Volatility

Notes: This figure demonstrates the importance of the features within each data category. The results follow from the machine learning process that maps previous data (features) into future returns (Panel (a)) and future realized volatility (Panel (b)). The data correspond to four categories: market data (bottom yellow), fundamentals, sell-side analyst recommendations, and analyst report textual content (top gray). The sensitivity scores are standardized to one and aggregated at the data category level to highlight the relative importance of each category. The sensitivity levels rely on interpretable machine learning techniques discussed in Section 6.2.

We obtain several noteworthy results. First, Panel (a) of Figure 4 reports the results for return predictions using the RF algorithm. Overall, most of the model's applicability stems from the market and fundamental features. At the top of the figure, we note two gray shaded

areas corresponding to the analysts' related content, whose contribution diminishes over time, showing some cyclicality primarily driven by the recommendation component. Second, we observe that the majority of realized volatility explainability for the RF algorithm reported in Panel (b) originates from the small feature space of market data that contains prior returns, volatility, and market capitalization. Additionally, we note that the analyst-related content plays a much smaller role in explaining volatility predictions than returns.

Finally, the sensitivity analysis for the MLP (Panels (c) and (d)) shows similar insights as the RF; however, there is one main difference when it comes to return predictions: the analysts' content contribution over time does not diminish as it does with the RF algorithm. This suggests that an MLP agent relies more consistently on such content throughout the sample period. Nevertheless, it is important to note that the baseline MLP model does not emerge as the top performer. For this reason, incorporating more analyst-related content could offer benefits for portfolio selection. In contrast, for the RF model, the baseline feature space serves as a strong benchmark, making the impact of analyst contributions less significant.

6.3 Mean-Variance Analysis

We subject our ML predictions to one final test. By design, the ML experiment produces predictions for both returns and volatility. The portfolio sorting results discussed in Section 5.4 focused exclusively on return predictions, i.e., ranking stocks based on future return forecasts. However, these portfolio sorts did not utilize the volatility-based predictions. To address this, we now leverage both return and volatility predictions using several mean-variance portfolio rules.

6.3.1 Portfolio Rules

For brevity, the detailed experiment description is provided in Section IA.5 of the Internet Appendix. However, it is worthwhile to briefly outline the analysis to follow. In particular, we utilize both return and volatility forecasts to form five mean/variance portfolio rules commonly studied in the literature (see, e.g., Kan et al. (2022)). These rules rely on mean returns as well as volatilities. We consider a mean-variance portfolio rule without short-sales (Jagannathan and Ma, 2003). Additionally, we leverage the global minimum variance portfolio with and without short-sale constraints. It is important to note that these two rules rely solely on volatility forecasts. The other two rules are related to volatility and risk-to-reward timing (Kirby and Ostdiek, 2012). The former is based exclusively on volatility predictions, whereas the latter incorporates both mean and volatility forecasts. The two rules

from Kirby and Ostdiek (2012) can be viewed as ad-hoc modifications of the conventional mean-variance portfolios.

6.3.2 Experiment

We conduct our experiment on a subset of the stocks studied in the baseline analysis. In particular, at the end of each year, we choose the largest 100 stocks in terms of market capitalization (see, e.g., DeMiguel et al. (2009); Kan et al. (2022)). For each stock, we utilize the ML forecasts to form the aforementioned five portfolio rules and evaluate their monthly returns over the next year. At the end of the following year, we repeat the same procedure: choose the largest market capitalization stocks, utilize their stock and volatility forecasts, and evaluate their performance over the next 12 months. We repeat this procedure until the end of the sample period. Additionally, the experiment is replicated for the smallest 100 market capitalization stocks to control for size.

Given the out-of-sample portfolio returns, we evaluate the marginal contribution of each feature space M conditional on algorithm a. Specifically, we compute the certainty-equivalent return (CER) for different portfolio rules, ML algorithms, and feature spaces. The CER of an investor with a risk aversion level of γ based on algorithm a and feature space M is given

$$CER(a, M) = \mu_p(a, M) - \frac{\gamma}{2}\sigma_p^2(a, M) \tag{6.2}$$

where $\mu_p(a, M)$ and $\sigma_p^2(a, M)$ denote the out-of-sample portfolio mean return and volatility based on algorithm a and feature space M. Similar to our main analysis from Section 5.4, we compute $CER(a, M) - CER(a, 1)$, where $CER(a, 1)$ is the CER return of the baseline model that does not include any analyst data based on algorithm a. We note that this incremental value is similar to the metric used by Farboodi et al. (2024) in valuing financial data. To compute $CER(a, M)$, we set $\gamma = 3$ (and $\gamma = 5$ as additional robustness).

6.3.3 Results

For a given γ and portfolio rule, the experiment produces 98 $CER(a, M)$ values, covering seven ML algorithms, seven feature spaces, and two stock groups (small and large). These values are detailed in Table 7 in the Internet Appendix for all five portfolio rules. To interpret these results and assess the content of analyst data in this mean-variance portfolio exercise, we employ an OLS regression, where the CER is regressed on three dummy variables, as shown in Table 7. The first variable (NonLinear) equals one if the machine learning algorithm is nonlinear and zero otherwise, with LMNET being the only linear model in the experiment. The second variable (Analyst) equals one if the feature space includes analyst data, i.e.,

$M > 1$. Finally, we include a size dummy that indicates whether the experiment involves the largest market capitalization stocks. Table 7 is organized by portfolio rule type, with CER values aggregated by taking the average across the three rules that rely on mean-variance forecasts and the two rules that rely solely on variance forecasts.

We summarize the findings of Table 7 as follows. First, we find weak evidence supporting the inclusion of the analyst data, with a positive coefficient that is significant at the 10% level. The effect is more pronounced when considering portfolio rules that depend on mean and variance forecasts. This finding aligns with the sensitivity analysis from Section 6.2, which suggests that analyst data plays a minor role in predicting future realized volatility. Second, nonlinearity reveals some value, an outcome that is consistent with our complexity conjecture. However, this result is primarily evident among more risk-averse investors focused on volatility-managed portfolios. Third, it is notable that this weak evidence primarily arises when the feature space excludes market and fundamental data while relying on analyst data only, i.e., the case of $M = 5, 6, 7$. This is evident when comparing Panel (a) with Panel (b) from Table 7, where the coefficient on the analyst dummy variable becomes statistically insignificant when the feature space excludes analyst-only data, i.e., omitting the experiments with $M = 5, 6, 7$.

A closer look at the actual data points from Table IA.13 in the Internet Appendix reveals considerable heterogeneity across the results, making it difficult, if not impossible, to ex-ante choose a model that consistently delivers a higher CER. Overall, the MV analysis of this section reinforces our main results and the additional evidence conducted throughout this section. The key takeaway is that while analyst data may have some value in constructing out-of-sample optimal portfolios, investors would likely achieve more reliable outcomes by focusing on market and fundamental data, which consistently provide a robust benchmark. This approach mitigates model risk and avoids the potential pitfalls of selecting a "best" feature ex-ante that could lead to significantly lower risk-adjusted returns.

7 Conclusion

While asset managers arguably use all sources of information, the actual aggregation process is highly idiosyncratic and essentially unknown. In light of the sheer magnitude of the global asset management industry, a better understanding of the structure and dynamics of the contribution of different sources of information to the formation of efficient portfolios is of a great importance for both practice and academic research. Our contribution to this discussion is multi-faceted. On a very large-scale sample of over 1,309,335 sell-side analyst reports, covering almost 2,515 stocks over 25 years, we implement a machine learning

approach and show how economic agents with varying levels of sophistication and beliefs can leverage diverse types of information to dynamically form forward-looking portfolios.

Our results indicate that portfolios formed using time-varying loadings of information inputs generate strongly significant abnormal returns over the testing period. However, we find that the relative contribution of inputs exhibits substantial time-series variation. Parsing such evidence, we highlight that analyst-related information has indeed marginal economic value in terms of the textual content of analyst reports - particularly as a complement to more simple analyst outputs (e.g., recommendations, number of estimates, etc.) - but only when the agent is unable to efficiently process common market and fundamental data. Our approach is model-agnostic, but extracts the unobservable, heterogeneous information processing functions within the portfolio design process of the investment industry, an industry influenced by multiple time-varying factors. Consequently, the performance of traditional buy-side portfolio managers might be more appropriately measured against machine learning algorithms that emulate a sophisticated agent with access to the same set of heterogeneous inputs rather than against typical passive or peer-based benchmarks. In this view, the value of analysts information is limited. However, it is unknown whether the information has any value for forming corporate bond portfolios, an interesting question we leave for future research.

References

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

Bali, T. G., Brown, S. J., Murray, S., Tang, Y., 2017. A lottery-demand-based explanation of the beta anomaly. Journal of Financial and Quantitative Analysis 52, 2369–2397.

Bali, T. G., Kelly, B. T., Mörke, M., Rahman, J., 2024. Machine forecast disagreement. Tech. rep., National Bureau of Economic Research.

Ball, R. T., Ghysels, E., 2018. Automated earnings forecasts: Beat analysts or combine and conquer? Management Science 64, 4936–4952.

Baltussen, G., Van Bekkum, S., Van Der Grient, B., 2018. Unknown unknowns: uncertainty about risk and stock returns. Journal of Financial and Quantitative Analysis 53, 1615–1651.

Barber, B., Lehavy, R., McNichols, M., Trueman, B., 2001. Can investors profit from the prophets? security analyst recommendations and stock returns. The Journal of Finance 56, 531–563.

Barber, B. M., Lehavy, R., Trueman, B., 2010. Ratings changes, ratings levels, and the predictive value of analysts' recommendations. Financial Management 39, 533–553.

Barber, B. M., Loeffler, D., 1993. The "dartboard" column: Second-hand information and price pressure. Journal of Financial and Quantitative Analysis 28, 273–284.

Biecek, P., 2018. Dalex: explainers for complex predictive models in r. The Journal of Machine Learning Research 19, 3245–3249.

Bradshaw, M. T., 2011. Analysts' forecasts: what do we know after decades of work? Available at SSRN 1880339 .

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

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Brochet, F., Loumioti, M., Serafeim, G., 2015. Speaking of the short-term: Disclosure horizon and managerial myopia. Review of Accounting Studies 20, 1122–1163.

Brown, L. D., Call, A. C., Clement, M. B., Sharp, N. Y., 2015. Inside the "black box" of sell-side financial analysts. Journal of Accounting Research 53, 1–47.

Cakici, N., Zaremba, A., 2024. Accounting vs market information: What matters more for stock return predictability? Available at SSRN 4637008 .

Cao, S., Jiang, W., Wang, J., Yang, B., 2024. From man vs. machine to man+ machine: The art and ai of stock analyses. Journal of Financial Economics 160, 103910.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Caylor, M., Cecchini, M., Winchel, J., 2017. Analysts' qualitative statements and the profitability of favorable investment recommendations. Accounting, Organizations and Society 57, 33–51.

Chen, A. Y., Zimmermann, T., 2021. Open source cross-sectional asset pricing. Critical Finance Review, Forthcoming .

Chen, X., Cho, Y. H., Dou, Y., Lev, B., 2022. Predicting future earnings changes using machine learning and detailed financial data. Journal of Accounting Research 60, 467–515.

Chhaochharia, V., Kumar, A., Rantala, V., Zhang, A., 2023. Artificially intelligent analyst sentiment and aggregate market behavior. Available at SSRN 4249442 .

Chordia, T., Subrahmanyam, A., Anshuman, V. R., 2001. Trading activity and expected stock returns. Journal of Financial Economics 59, 3–32.

Cong, L. W., Feng, G., He, J., He, X., 2022. Growing the efficient frontier on panel trees. NBER Working Paper .

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The Review of Financial studies 22, 1915–1953.

Dessaint, O., Foucault, T., Frésard, L., 2024. Does alternative data improve financial forecasting? the horizon effect. The Journal of Finance 79, 2237–2287.

Dimson, E., Marsh, P., 1984. An analysis of brokers' and analysts' unpublished forecasts of uk stock returns. The Journal of Finance 39, 1257–1292.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al., 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9, 155–161.

Easley, D., O'Hara, M., Paperman, J., 1998. Financial analysts and information-based trade. Journal of Financial Markets 1, 175–201.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Farboodi, M., Singal, D., Veldkamp, L., Venkateswaran, V., 2024. Valuing financial data. The Review of Financial Studies .

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20, 1–81.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Guo, R.-J., Lev, B., Shi, C., 2006. Explaining the short-and long-term ipo anomalies in the us by r&d. Journal of Business Finance & Accounting 33, 550–579.

Harvey, C. R., Liu, Y., 2021. Lucky factors. Journal of Financial Economics 141, 413–435.

Hirshleifer, D., Shi, Y., Wu, W., 2024. Do sell-side analysts say "buy" while whispering "sell"? Review of Finance 28, 1275—1310.

Holthausen, R. W., Larcker, D. F., 1992. The prediction of stock returns using financial statement information. Journal of Accounting and Economics 15, 373–411.

Huang, A. H., Zang, A. Y., Zheng, R., 2014. Evidence on the information content of text in analyst reports. The Accounting Review 89, 2151–2180.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. The Journal of Finance 58, 1651–1683.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning, vol. 112. Springer.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. The Journal of Finance 45, 881–898.

Kadan, O., Madureira, L., Wang, R., Zach, T., 2008. Conflicts of interest and stock recommendations: The effects of the global settlement and related regulations. The Review of Financial Studies 22, 4189–4217.

Kan, R., Wang, X., Zhou, G., 2022. Optimal portfolio choice with estimation risk: No risk-free asset case. Management Science 68, 2047–2068.

Ke, S., 2023. The double-edged sword of data mining: Implications on asset pricing and information efficiency. Available at SSRN .

Kelly, B. T., Malamud, S., Zhou, K., 2022. The virtue of complexity everywhere. Available at SSRN .

Kirby, C., Ostdiek, B., 2012. It's all in the timing: simple active portfolio strategies that outperform naive diversification. Journal of Financial and Quantitative Analysis 47, 437–467.

Kong, D., Lin, C., Liu, S., Tan, W., 2021. Whose money is smart? individual and institutional investors' trades based on analyst recommendations. Journal of Empirical Finance 62, 234–251.

Kuhn, M., et al., 2008. Building predictive models in r using the caret package. Journal of Statistical Software 28, 1–26.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Lundholm, R., Rogo, R., 2020. Do excessively volatile forecasts impact investors? Review of Accounting Studies 25, 636–671.

Matsumoto, D., Pronk, M., Roelofsen, E., 2011. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. The Accounting Review 86, 1383–1414.

Michaely, R., Womack, K. L., 1999. Conflict of interest and the credibility of underwriter analyst recommendations. The Review of Financial Studies 12, 653–686.

Miller, E. M., 1977. Risk, uncertainty, and divergence of opinion. The Journal of Finance 32, 1151–1168.

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, Springer, pp. 193–204.

Nagel, S., Xu, Z., 2023. Dynamics of subjective risk premia. Journal of Financial Economics 150, 103713.

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

Painter, M., 2018. Unlevelling the playing field: The investment value and capital market consequences of alternative data. Available at SSRN 3222741 .

Ripley, B. D., 2007. Pattern recognition and neural networks. Cambridge university press.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al., 1985. Learning internal representations by error propagation.

Soliman, M. T., 2008. The use of dupont analysis by market participants. The Accounting Review 83, 823–853.

Sorescu, S., Subrahmanyam, A., 2006. The cross section of analyst recommendations. Journal of Financial and Quantitative Analysis 41, 139–168.

Stickel, S. E., 1995. The anatomy of the performance of buy and sell recommendations. Financial Analysts Journal 51, 25–39.

Valta, P., 2016. Strategic default, debt structure, and stock returns. Journal of Financial and Quantitative Analysis 51, 197–229.

Womack, K. L., 1996. Do brokerage analysts' recommendations have investment value? The Journal of Finance 51, 137–167.

Zhang, X. F., 2006. Information uncertainty and stock returns. The Journal of Finance 61, 105–137.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Tables

Table 1: **Single Portfolio Sorting Based on Analyst Recommendations**
At the end of each calendar year, firms are grouped into ten deciles based on sell-side analyst recommendation data. The table presents the risk-adjusted returns of equal-weighted portfolios. The (H)-(L) column denotes the return of a self-financing portfolio that goes long on the top decile and shorts the bottom one. In all cases, portfolios are rebalanced annually. Risk-adjusted returns are computed using the five risk factors by Fama and French (2015) and the momentum risk factor by Carhart (1997). The full period corresponds to portfolio time series returns from January 1998 to December 2018. The sub-samples, pre-2003 and post-2003, account for the enactment of the Global Research Settlement. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags. For a full definition of the variables, see Table IA.2 in the Internet Appendix.

	(L)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)
						Full Period					
MEANREC	0.068**	−0.019	0.007	−0.002	−0.0004	0.049**	0.023	0.015	0.078***	0.101***	0.033
	(0.028)	(0.019)	(0.021)	(0.021)	(0.018)	(0.024)	(0.016)	(0.020)	(0.024)	(0.029)	(0.022)
STDEV	0.109***	0.098***	0.064*	0.022	0.004	0.022	0.002	−0.004	0.001	0.002	−0.107***
	(0.031)	(0.034)	(0.034)	(0.023)	(0.014)	(0.018)	(0.016)	(0.015)	(0.015)	(0.018)	(0.027)
NUMREC	0.138***	0.103**	0.048*	−0.009	0.029	0.002	−0.020	−0.001	0.006	0.029*	−0.109***
	(0.036)	(0.043)	(0.026)	(0.022)	(0.019)	(0.017)	(0.016)	(0.010)	(0.015)	(0.016)	(0.033)
BUYPCT	0.151***	0.057***	0.012	0.001	0.006	0.001	−0.001	0.024	0.023	0.047	−0.104***
	(0.030)	(0.021)	(0.021)	(0.017)	(0.014)	(0.016)	(0.016)	(0.023)	(0.027)	(0.034)	(0.028)
						Pre-2003					
MEANREC	0.229***	0.043	−0.004	0.030	0.093**	0.179**	0.135***	0.147***	0.204***	0.337***	0.108***
	(0.065)	(0.045)	(0.053)	(0.029)	(0.045)	(0.075)	(0.041)	(0.033)	(0.038)	(0.069)	(0.032)
STDEV	0.246***	0.282***	0.265***	0.173***	0.060*	0.107***	0.106***	−0.0005	0.028	0.124***	−0.122**
	(0.065)	(0.061)	(0.093)	(0.059)	(0.033)	(0.029)	(0.036)	(0.037)	(0.052)	(0.037)	(0.059)
NUMREC	0.320***	0.297***	0.224***	0.021	0.139***	0.096**	0.063	0.057**	0.066	0.112***	−0.208**
	(0.070)	(0.108)	(0.063)	(0.041)	(0.045)	(0.039)	(0.040)	(0.027)	(0.042)	(0.034)	(0.079)
BUYPCT	0.372***	0.165***	0.176***	0.113***	0.025	0.017	0.031	0.153***	0.121***	0.208**	−0.164**
	(0.078)	(0.035)	(0.033)	(0.028)	(0.048)	(0.034)	(0.041)	(0.053)	(0.039)	(0.085)	(0.076)
						Post-2003					
MEANREC	0.026	−0.023	0.016	−0.007	−0.021*	0.003	0.001	−0.022	0.053**	0.040**	0.014
	(0.026)	(0.016)	(0.021)	(0.023)	(0.013)	(0.010)	(0.011)	(0.015)	(0.023)	(0.019)	(0.025)
STDEV	0.075**	0.051	0.007	−0.013	−0.004	−0.006	−0.018*	−0.0001	0.002	−0.023*	−0.097***
	(0.030)	(0.033)	(0.025)	(0.016)	(0.013)	(0.014)	(0.011)	(0.011)	(0.011)	(0.012)	(0.029)
NUMREC	0.092**	0.049	0.0005	−0.010	0.003	−0.011	−0.038***	−0.011	−0.007	0.005	−0.086**
	(0.035)	(0.037)	(0.021)	(0.022)	(0.014)	(0.012)	(0.011)	(0.009)	(0.011)	(0.013)	(0.035)
BUYPCT	0.093***	0.036	−0.029**	−0.023	0.006	0.004	−0.011	−0.008	0.004	0.002	−0.091***
	(0.020)	(0.023)	(0.015)	(0.015)	(0.011)	(0.015)	(0.014)	(0.018)	(0.025)	(0.028)	(0.023)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: **Single Portfolio Sorting Based on Analyst Report Sentiment**

At the end of each calendar year, firms are grouped into ten deciles based on sell-side analyst textual data. The table presents the risk-adjusted returns of equal-weighted portfolios. The (H)-(L) column denotes the return of a self-financing portfolio that goes long on the top decile and shorts the bottom one. In all cases, portfolios are rebalanced annually. Risk-adjusted returns are computed using the five risk factors by Fama and French (2015) and the momentum risk factor by Carhart (1997). The full period corresponds to portfolio time series returns from January 1998 to December 2018. The sub-samples, pre-2003 and post-2003, account for the enactment of the Global Research Settlement. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags. For a full definition of the variables, see Table IA.2 in the Internet Appendix.

	(L)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)
						Full Period					
Net Tone	0.048**	−0.009	0.015	−0.004	0.037	−0.008	0.041**	0.070***	0.043**	0.082***	0.034
	(0.024)	(0.021)	(0.020)	(0.018)	(0.023)	(0.017)	(0.019)	(0.025)	(0.020)	(0.025)	(0.023)
Uncertainty	0.106***	0.079**	0.048*	0.015	0.016	0.010	0.00004	0.011	0.004	0.027	−0.078***
	(0.024)	(0.034)	(0.026)	(0.020)	(0.016)	(0.016)	(0.018)	(0.022)	(0.021)	(0.023)	(0.023)
Valuation	0.061**	0.027	0.026	0.055**	0.007	0.043**	0.028	0.005	0.028*	0.032	−0.030
	(0.028)	(0.023)	(0.024)	(0.023)	(0.024)	(0.022)	(0.017)	(0.014)	(0.016)	(0.020)	(0.031)
Alternative	0.043***	0.030*	−0.017	−0.001	0.052***	0.030	0.023	0.045**	0.026	0.080***	0.037
	(0.016)	(0.017)	(0.022)	(0.018)	(0.020)	(0.021)	(0.018)	(0.021)	(0.030)	(0.025)	(0.025)
Long Term	0.018	0.025	0.004	0.045**	0.036*	0.035	0.052*	0.024	0.028*	0.047***	0.029
	(0.031)	(0.023)	(0.017)	(0.021)	(0.022)	(0.021)	(0.028)	(0.018)	(0.015)	(0.018)	(0.034)
Short Term	0.055**	0.019	0.030*	0.010	0.006	0.029	0.025	0.042	0.042	0.058*	0.002
	(0.026)	(0.026)	(0.018)	(0.015)	(0.017)	(0.020)	(0.018)	(0.030)	(0.026)	(0.031)	(0.031)
Financial	0.046	0.018	0.016	0.004	0.010	0.021	0.032	0.017	0.054**	0.097***	0.051
	(0.034)	(0.028)	(0.018)	(0.014)	(0.021)	(0.019)	(0.024)	(0.024)	(0.023)	(0.023)	(0.034)
						Pre-2003					
Net Tone	0.214***	0.086*	0.080	0.099***	0.173***	0.007	0.170***	0.185**	0.162***	0.201***	−0.013
	(0.060)	(0.044)	(0.057)	(0.031)	(0.061)	(0.031)	(0.037)	(0.075)	(0.034)	(0.054)	(0.061)
Uncertainty	0.247***	0.283***	0.164***	0.101**	0.054*	0.092***	0.048	0.136***	0.077**	0.177***	−0.070
	(0.058)	(0.097)	(0.052)	(0.044)	(0.031)	(0.021)	(0.043)	(0.043)	(0.034)	(0.039)	(0.055)
Valuation	0.213***	0.120**	0.088*	0.181***	0.166***	0.209***	0.091**	0.084***	0.079*	0.129***	−0.085
	(0.058)	(0.048)	(0.048)	(0.045)	(0.057)	(0.051)	(0.044)	(0.030)	(0.041)	(0.035)	(0.070)
Alternative	0.060*	0.121***	0.116***	0.108***	0.202***	0.166***	0.083**	0.190***	0.099	0.217***	0.157***
	(0.031)	(0.037)	(0.040)	(0.030)	(0.046)	(0.043)	(0.035)	(0.055)	(0.069)	(0.035)	(0.039)
Long Term	0.117**	0.132**	0.062	0.205***	0.157***	0.114*	0.261***	0.126***	0.082**	0.110***	−0.008
	(0.058)	(0.051)	(0.039)	(0.053)	(0.041)	(0.063)	(0.074)	(0.044)	(0.036)	(0.033)	(0.082)
Short Term	0.153***	0.042	0.159***	0.057*	0.096**	0.155***	0.060**	0.216**	0.217***	0.220**	0.067
	(0.042)	(0.052)	(0.027)	(0.033)	(0.044)	(0.051)	(0.027)	(0.090)	(0.052)	(0.098)	(0.083)
Financial	0.185***	0.062**	0.101***	0.035	0.156***	0.106***	0.123	0.173***	0.171**	0.265***	0.080
	(0.058)	(0.028)	(0.029)	(0.043)	(0.048)	(0.035)	(0.081)	(0.051)	(0.067)	(0.055)	(0.078)
						Post-2003					
Net Tone	0.004	−0.032	0.0003	−0.031**	0.005	−0.002	0.008	0.036*	0.021	0.054**	0.050**
	(0.018)	(0.020)	(0.017)	(0.015)	(0.014)	(0.013)	(0.015)	(0.018)	(0.021)	(0.025)	(0.023)
Uncertainty	0.068***	0.018	0.026	−0.002	0.017	−0.007	−0.008	−0.021	−0.015	−0.011	−0.080***
	(0.020)	(0.020)	(0.025)	(0.016)	(0.015)	(0.014)	(0.016)	(0.019)	(0.022)	(0.020)	(0.023)

Valuation	0.026	0.008	0.013	0.015	−0.034*	0.001	0.015	−0.016	0.022	0.017	−0.009
	(0.028)	(0.022)	(0.021)	(0.019)	(0.018)	(0.014)	(0.014)	(0.015)	(0.016)	(0.018)	(0.032)
Alternative	0.050***	0.006	−0.050***	−0.023	0.017	−0.009	0.015	0.008	0.003	0.049*	−0.001
	(0.018)	(0.015)	(0.018)	(0.014)	(0.014)	(0.016)	(0.018)	(0.013)	(0.027)	(0.027)	(0.027)
Long Term	0.005	0.004	−0.004	0.005	0.004	0.015	−0.005	−0.010	0.015	0.038**	0.032
	(0.034)	(0.021)	(0.017)	(0.015)	(0.018)	(0.017)	(0.016)	(0.010)	(0.015)	(0.018)	(0.036)
Short Term	0.038	0.019	0.004	0.00004	−0.014	−0.002	0.024	−0.011	−0.001	0.009	−0.029
	(0.027)	(0.026)	(0.015)	(0.016)	(0.012)	(0.016)	(0.019)	(0.018)	(0.021)	(0.018)	(0.026)
Financial	0.017	0.024	0.002	0.001	−0.026**	0.004	0.004	−0.023	0.012	0.052***	0.036
	(0.036)	(0.032)	(0.019)	(0.011)	(0.012)	(0.016)	(0.014)	(0.017)	(0.016)	(0.016)	(0.037)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: **Double Portfolio Sorting Based on Analyst Recommendations and Net Tone**
At the end of each calendar year, stocks are grouped into three mean sell-side analyst recommendation (MEANREC) terciles. Within each recommendation group, stocks are further grouped into three terciles based on the net tone of the analyst report (Net Tone). The letters W and S denote weak (e.g., sell) and strong (e.g., buy) recommendation-related groups, respectively, while L and H denote the low and high net tone groups, respectively. The table presents the raw and risk-adjusted returns of equal-weighted double-sorted portfolios. The (H)-(L) column shows the return of a self-financing portfolio that goes long on the group with the highest net tone and shorts the group with the lowest net tone within the same recommendation bracket. Similarly, the (W)-(S) row shows the return of a self-financing portfolio that longs the group associated with the weakest recommendation while shorting the one with the strongest recommendation within the same net tone bracket. In all cases, portfolios are rebalanced annually. Risk-adjusted returns are computed using five risk factors by Fama and French (2015) and the momentum factor by Carhart (1997). The full period covers portfolio time series returns from Jan 1998 to Dec 2018. The sub-samples, pre-2003 and post-2003, control for enactment of the Global Research Settlement. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with six lags.

	Raw Returns					*Risk-Adjusted Returns*			
	(L)	(2)	(H)	(H)-(L)		(L)	(2)	(H)	(H)-(L)
					Panel (a) Full Period				
(S)	0.087	0.093	0.133**	0.046**	(S)	0.070	0.066	0.091	0.022
	(0.058)	(0.057)	(0.053)	(0.020)		(0.058)	(0.057)	(0.057)	(0.021)
(2)	0.103*	0.094*	0.136**	0.033	(2)	0.071	0.066	0.105*	0.034
	(0.056)	(0.053)	(0.055)	(0.022)		(0.055)	(0.050)	(0.054)	(0.022)
(W)	0.134**	0.138**	0.185***	0.051*	(W)	0.109*	0.106*	0.153**	0.044*
	(0.062)	(0.057)	(0.058)	(0.028)		(0.058)	(0.057)	(0.060)	(0.025)
(W)-(S)	0.047*	0.044**	0.052**		(W)-(S)	0.039	0.040*	0.062***	
	(0.027)	(0.022)	(0.020)			(0.025)	(0.024)	(0.022)	
					Panel (b) Pre-2003				
(S)	0.156	0.057	0.125	−0.031	(S)	0.054	−0.037	−0.043	−0.097
	(0.115)	(0.085)	(0.090)	(0.059)		(0.081)	(0.089)	(0.103)	(0.068)
(2)	0.157*	0.078	0.186	0.029	(2)	0.051	−0.010	0.049	−0.002
	(0.091)	(0.098)	(0.123)	(0.088)		(0.072)	(0.102)	(0.119)	(0.084)
(W)	0.212**	0.249***	0.266**	0.054	(W)	0.152**	0.173**	0.115	−0.036
	(0.080)	(0.078)	(0.111)	(0.076)		(0.071)	(0.073)	(0.106)	(0.069)
(W)-(S)	0.056	0.191***	0.141**		(W)-(S)	0.098*	0.211***	0.159***	
	(0.095)	(0.047)	(0.058)			(0.055)	(0.032)	(0.054)	
					Panel (c) Post-2003				
(S)	0.069	0.102	0.135**	0.065***	(S)	0.066	0.087	0.115*	0.049**
	(0.066)	(0.067)	(0.063)	(0.019)		(0.064)	(0.066)	(0.061)	(0.019)
(2)	0.090	0.098	0.124**	0.034**	(2)	0.073	0.082	0.111*	0.038**
	(0.066)	(0.062)	(0.061)	(0.017)		(0.063)	(0.058)	(0.057)	(0.016)
(W)	0.114	0.110	0.165**	0.050*	(W)	0.095	0.088	0.156**	0.061**
	(0.075)	(0.068)	(0.066)	(0.029)		(0.070)	(0.066)	(0.066)	(0.026)
(W)-(S)	0.045*	0.008	0.030		(W)-(S)	0.029	0.001	0.041**	
	(0.023)	(0.019)	(0.018)			(0.027)	(0.021)	(0.019)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: **Machine Learning Predictions and Future Returns Using Random Forest (RF)**

At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast using a nonlinear random forest algorithm (Breiman, 2001). For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to form the ML forecast and subsequent portfolio formation. Baseline data capturing market and fundamentals are represented by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ uses all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are included for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The $\pi(M) - \pi(1)$ column evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	−0.037	0.002	0.006	0.006	0.017	0.018	0.027	0.047**	0.082**	0.148***	0.186***	
	(0.046)	(0.029)	(0.021)	(0.020)	(0.015)	(0.012)	(0.016)	(0.022)	(0.039)	(0.039)	(0.052)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.038	−0.003	−0.005	0.007	−0.002	0.016	0.039**	0.045**	0.084***	0.175***	0.212***	0.027
	(0.044)	(0.028)	(0.020)	(0.020)	(0.019)	(0.015)	(0.018)	(0.023)	(0.031)	(0.048)	(0.054)	(0.018)
						Market and Fundamentals plus Analyst Text						
3	−0.022	0.003	−0.0001	0.048**	0.018	0.008	0.005	0.057**	0.044	0.154***	0.176***	−0.010
	(0.043)	(0.033)	(0.025)	(0.021)	(0.018)	(0.014)	(0.014)	(0.025)	(0.031)	(0.045)	(0.060)	(0.023)
						Altogether						
4	−0.047	−0.003	0.026	0.001	0.011	0.018	0.043***	0.038*	0.082**	0.146***	0.193***	0.007
	(0.044)	(0.028)	(0.021)	(0.022)	(0.016)	(0.015)	(0.015)	(0.021)	(0.040)	(0.041)	(0.053)	(0.015)
						Text Alone						
5	0.047	−0.009	0.061***	0.025*	0.065***	0.016	0.0002	0.024	0.032*	0.050**	0.004	−0.182***
	(0.036)	(0.023)	(0.023)	(0.015)	(0.020)	(0.019)	(0.016)	(0.022)	(0.019)	(0.019)	(0.033)	(0.049)
						Recommendation Data Alone						
6	0.013	−0.001	0.009	0.005	0.046**	0.066***	0.055**	0.067***	0.031	0.024	0.011	−0.174***
	(0.024)	(0.021)	(0.016)	(0.015)	(0.021)	(0.021)	(0.022)	(0.022)	(0.024)	(0.021)	(0.018)	(0.053)
						Text and Recommendations Alone						
7	0.031	0.034	0.024	0.005	0.040*	0.027**	0.046**	0.046**	0.017	0.042*	0.011	−0.175***
	(0.035)	(0.021)	(0.018)	(0.018)	(0.023)	(0.014)	(0.020)	(0.023)	(0.022)	(0.022)	(0.032)	(0.056)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: **Machine Learning Predictions and Future Returns Using Multilayer Perceptron (MLP)**
At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is nonlinear and is based on the Multilayer Perceptron (MLP), which represents an artificial neural network (ANN) mapping technology (Rumelhart et al., 1985). For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	−0.013	0.024	0.009	0.029	0.031**	0.006	0.043*	0.043**	0.056*	0.097**	0.110*	
	(0.037)	(0.032)	(0.022)	(0.022)	(0.015)	(0.016)	(0.023)	(0.021)	(0.029)	(0.045)	(0.063)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.015	−0.014	0.001	0.036*	0.030*	0.039**	0.026	0.027	0.070***	0.121***	0.136**	0.026
	(0.040)	(0.028)	(0.024)	(0.022)	(0.017)	(0.016)	(0.017)	(0.018)	(0.025)	(0.040)	(0.058)	(0.031)
						Market and Fundamentals plus Analyst Text						
3	−0.004	0.005	−0.020	0.010	0.035*	0.032*	0.029*	0.050*	0.063**	0.127***	0.130*	0.021
	(0.048)	(0.025)	(0.019)	(0.016)	(0.019)	(0.018)	(0.016)	(0.026)	(0.026)	(0.041)	(0.068)	(0.025)
						Altogether						
4	−0.017	−0.035	0.024	0.016	0.022*	0.019	0.027*	0.084***	0.042**	0.133***	0.150***	0.040
	(0.037)	(0.026)	(0.021)	(0.017)	(0.013)	(0.014)	(0.015)	(0.031)	(0.021)	(0.038)	(0.044)	(0.039)
						Text Alone						
5	0.056*	0.059**	0.035*	0.050**	0.011	0.012	0.021	0.006	0.016	0.049**	−0.006	−0.116
	(0.030)	(0.026)	(0.020)	(0.021)	(0.021)	(0.016)	(0.025)	(0.015)	(0.019)	(0.023)	(0.035)	(0.072)
						Recommendation Data Alone						
6	0.023	−0.008	0.016	0.035	−0.001	0.030	0.025	0.031	0.060**	0.109***	0.085***	−0.024
	(0.029)	(0.023)	(0.017)	(0.021)	(0.019)	(0.019)	(0.022)	(0.025)	(0.023)	(0.023)	(0.032)	(0.063)
						Text and Recommendations Alone						
7	0.042	0.038*	0.015	0.008	0.026	0.024	0.049**	0.002	0.055***	0.054**	0.012	−0.097
	(0.036)	(0.023)	(0.025)	(0.017)	(0.020)	(0.016)	(0.023)	(0.017)	(0.020)	(0.023)	(0.037)	(0.063)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: **Machine Learning Predictions and Future Returns Using Linear Model with Elastic Net (LMNET)**
At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is linear and utilizes a linear model with elastic net (LMNET) regularization (Zou and Hastie, 2005). For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
							Market and Fundamentals					
1	−0.019	−0.006	−0.031	0.006	0.022	0.045**	0.035*	0.034	0.086***	0.160***	0.179**	
	(0.048)	(0.034)	(0.027)	(0.022)	(0.017)	(0.019)	(0.020)	(0.027)	(0.029)	(0.050)	(0.078)	
							Market and Fundamentals plus Analysts Recommendation					
2	−0.012	−0.028	−0.012	0.016	−0.009	0.042**	0.028	0.048**	0.097***	0.160***	0.172**	−0.007
	(0.050)	(0.033)	(0.029)	(0.017)	(0.017)	(0.019)	(0.019)	(0.024)	(0.029)	(0.051)	(0.081)	(0.011)
							Market and Fundamentals plus Analyst Text					
3	−0.014	−0.007	−0.003	−0.009	0.013	0.058***	0.024	0.021	0.096***	0.151***	0.164**	−0.015
	(0.048)	(0.034)	(0.026)	(0.017)	(0.017)	(0.016)	(0.021)	(0.025)	(0.028)	(0.046)	(0.075)	(0.009)
							Altogether					
4	−0.010	−0.019	−0.003	−0.002	0.011	0.028**	0.028	0.041*	0.092***	0.164***	0.174**	−0.005
	(0.050)	(0.033)	(0.027)	(0.015)	(0.019)	(0.014)	(0.021)	(0.024)	(0.028)	(0.051)	(0.082)	(0.013)
							Text Alone					
5	0.050*	0.048**	0.060**	0.033	0.024	0.026	0.019	0.008	−0.001	0.046**	−0.004	−0.182**
	(0.028)	(0.021)	(0.023)	(0.026)	(0.016)	(0.023)	(0.023)	(0.017)	(0.017)	(0.023)	(0.032)	(0.081)
							Recommendation Data Alone					
6	0.030	0.002	0.001	0.030	0.030	0.033	0.008	0.0003	0.058**	0.126***	0.096***	−0.083
	(0.032)	(0.023)	(0.017)	(0.026)	(0.019)	(0.021)	(0.018)	(0.024)	(0.025)	(0.023)	(0.029)	(0.073)
							Text and Recommendations Alone					
7	0.041	0.014	0.019	0.013	0.005	0.047**	0.028	0.027	0.046**	0.077***	0.036	−0.143*
	(0.035)	(0.024)	(0.020)	(0.022)	(0.018)	(0.020)	(0.020)	(0.019)	(0.019)	(0.024)	(0.033)	(0.080)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: **Results Summary of Mean-Variance Rules**
The table below summarizes the results for the mean-variance analysis presented in Table IA.13 from the Internet Appendix using an ordinary least squares OLS regression. For each portfolio rule, the regression captures out-of-sample performance results based on a backtesting procedure that utilizes seven machine learning algorithms, seven feature spaces, and two stock universes, resulting in 98 out-of-sample outcomes. In total, there are five portfolio rules. The first is MV_NS, which denotes a mean-variance rule without short sales. The second and third, GMV and GMV_NS, denote the global minimum variance portfolio with and without short sales, respectively. The other two rules, KO_VT and KO_RT, correspond to the volatility and risk-to-reward timing rules by Kirby and Ostdiek (2012), respectively. See Section 6.3 along with Section IA.5 from the Internet Appendix for more details. At the end of each year, the largest (or smallest, depending on the experiment) 100 stocks by market capitalization are selected. The mean return and volatility are estimated using a machine learning (ML) technology that maps data model M into either next year's return or realized volatility. This procedure is repeated annually, resulting in 20×12 monthly returns for each data-algorithm-rule combination. The main dependent variable is the certainty equivalent return (CER), which depends on a risk aversion parameter γ. Columns (1) and (2) (respectively Columns (3) and (4)) present the results for $\gamma = 3$ (respectively $\gamma = 5$). The OLS regression distinguishes between rules that rely on both mean and volatility inputs (Columns (1) and (3)) and those that rely on volatility alone (Columns (2) and (4)). Size is a dummy variable indicating whether the experiment uses the 100 largest market cap stocks versus the smallest ones. NonLinear is a binary variable that equals 1 if the machine learning algorithm is nonlinear and 0 otherwise, with LMNET being the only linear model in the experiment. Analyst is a dummy variable indicating whether the feature space includes analyst data.

	Dependent variable: CER			
	$\gamma = 3$		$\gamma = 5$	
	Three Rules Mean and Volatility	Two Rules Volatility Only	Three Rules Mean and Volatility	Two Rules Volatility Only
	(1)	(2)	(3)	(4)
	Full Feature Space $M = 1, 2, 3, 4, 5, 6, 7$			
Constant	0.715	2.478***	−3.169**	−1.126
	(1.191)	(0.559)	(1.495)	(0.685)
Size	0.729	1.615***	0.017	1.437***
	(0.637)	(0.299)	(0.799)	(0.366)
NonLinear	−0.628	0.289	−1.009	0.954*
	(0.910)	(0.427)	(1.142)	(0.523)
Analyst	1.582*	0.846*	1.833	0.942*
	(0.910)	(0.427)	(1.142)	(0.523)
Observations	98	98	98	98
Adjusted R^2	0.018	0.240	0.004	0.164
	Sub Feature Space $M = 1, 2, 3, 4$			
Constant	0.150	2.625***	−3.475**	−0.615
	(1.322)	(0.340)	(1.576)	(0.480)
Size	1.443*	1.715***	0.780	1.506***
	(0.797)	(0.205)	(0.950)	(0.289)
NonLinear	−0.386	0.060	−1.096	0.318
	(1.139)	(0.293)	(1.358)	(0.413)
Analyst	0.465	0.140	0.514	0.181
	(0.921)	(0.237)	(1.097)	(0.334)
Observations	56	56	56	56
R^2	0.066	0.575	0.029	0.350
Adjusted R^2	0.012	0.550	−0.027	0.313
Note:				*p<0.1; **p<0.05; ***p<0.01

Internet Appendix

The Value of Data: Analyst Vs. Machine

The following Internet Appendix (IA) provides supplementary details, including feature construction, dictionary and variable definitions, as well as other descriptive results such as summary statistics and data visualizations, offering readers a deeper understanding of the rich data used in the main manuscript. Additionally, the IA presents further analyses, including single and double portfolio sorts (tables and figures), descriptive statistics of the final features used in the machine learning training process, and supplementary materials about the additional machine learning algorithms and their findings. The IA also includes further results and detailed discussions on the sensitivity analysis and the mean-variance experiment conducted in Section 6 of the manuscript.

IA.1 Dictionary and Variable Definitions

Table IA.1: **Valuation and Alternative Dictionary Terms**

Dictionary	Terms
Valuation	cf, discounted, cash, discount, cash, valuati, sum, of, the, part, sum, of, part, dividend discount, comparables, comparable, firm, comparable, compan, liquidation, val, replacement cost, peer, group, ratio, multiple, price-to-, price, to, present, value
Alternative	channel, check, data, satellite, imag, geolocation, sensors, social, media, web search, parking, lot, traffic, sentiment, textual, logistics, biometric

Table IA.2: **Variable Definitions** This table summarizes the main variables from the final panel data. The definition column indicates the mnemonic variable names corresponding to the database definitions for each variable. For other categories, the definition column describes the related variable. For definitions of the 94 variables capturing the broader feature space for market and fundamental data, see Table A.6 from Gu et al. (2020). Table IA.4 provides summary statistics for all variables/features considered in our investigation.

Variable	Definition
	Analyst Textual Content
Net Tone	Number of positive terms minus negative terms adjusted by total words in the document (Loughran and McDonald, 2011)
Uncertainty	Number of uncertainty-related terms adjusted by total words in the document (Loughran and McDonald, 2011)
Valuation	Number of valuation-related terms adjusted by total words in the document. Terms are motivated by Brown et al. (2015). See Table A.1 of the appendix for the full list
Alternative	Number of alternative data-related terms adjusted by total words in the document. Terms are motivated by Painter (2018) and Dessaint et al. (2024). See Table A.1 of the appendix for the full list
Long-Term	Number of long-term related terms adjusted by total words in the document (Brochet et al., 2015)
Short-Term	Number of short-term related terms adjusted by total words in the document (Brochet et al., 2015)
Financial	Number of financial-related terms adjusted by total words in the document (Matsumoto et al., 2011)
	Analyst Recommendation
MEANREC	Mean analyst recommendation, where 1 (5) is a strong buy (sell)
STDEV	Standard deviation of analysts' recommendations
NUMREC	Number of recommendations
BUYPCT	The percentage of buy recommendations
Ret	Stock return
RV	Realized volatility
MKTCAP	Market capitalization

IA.2 Summary Statistics

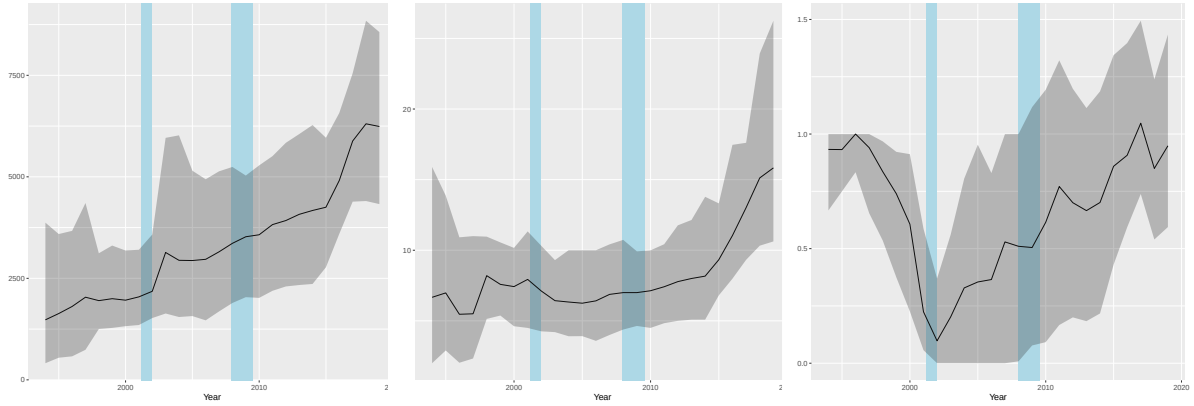
IA.2.1 Total Reports Over Time

Table IA.3: **Number of Reports and Coverage Over Time**
This table reports summary statistics for the number of sell-side analyst reports. Reports/Firm is the annual average number of reports issued on any unique company. It indicates how many reports are produced per stock in our sample on average, i.e., the ratio between the second and fifth columns. Similarly, the Reports/Analyst column computes the ratio between the second and third columns.

Year	Reports	Analysts	Brokerage Firms	Firms	Reports/Firm	Reports/Analyst
1984	1104	618	33	210	5.26	1.79
1985	922	349	33	208	4.43	2.64
1986	1498	438	39	250	5.99	3.42
1987	1871	528	39	266	7.03	3.54
1988	2306	625	51	287	8.03	3.69
1989	3058	735	58	326	9.38	4.16
1990	2611	773	75	313	8.34	3.38
1991	3034	784	98	352	8.62	3.87
1992	3269	802	112	421	7.76	4.08
1993	3578	1007	110	459	7.80	3.55
1994	5087	1278	126	550	9.25	3.98
1995	7032	1490	128	617	11.40	4.72
1996	8034	1753	149	721	11.14	4.58
1997	19551	2185	182	1317	14.85	8.95
1998	25426	2588	207	1363	18.65	9.82
1999	32456	2988	218	1424	22.79	10.86
2000	40815	3087	234	1526	26.75	13.22
2001	42254	3050	224	1612	26.21	13.85
2002	42265	3935	208	1641	25.76	10.74
2003	44376	3358	209	1100	40.34	13.22
2004	58030	3986	240	1189	48.81	14.56
2005	63463	3872	225	1407	45.11	16.39
2006	51390	3556	226	1404	36.60	14.45
2007	55836	4743	242	1547	36.09	11.77
2008	53456	3410	256	1519	35.19	15.68
2009	49669	3275	319	1461	34.00	15.17
2010	58010	3928	371	1529	37.94	14.77
2011	76058	3960	387	1704	44.63	19.21
2012	86087	3230	362	1770	48.64	26.65
2013	93141	3101	360	1826	51.01	30.04
2014	104804	3484	364	2270	46.17	30.08
2015	129246	3860	413	2381	54.28	33.48
2016	140270	3635	364	2354	59.59	38.59
2017	116383	2497	346	2311	50.36	46.61
2018	94542	2282	320	2107	44.87	41.43
Average	43455	2434	209	1192	27.23	14.20

IA.2.2 Other Analyst Report Features

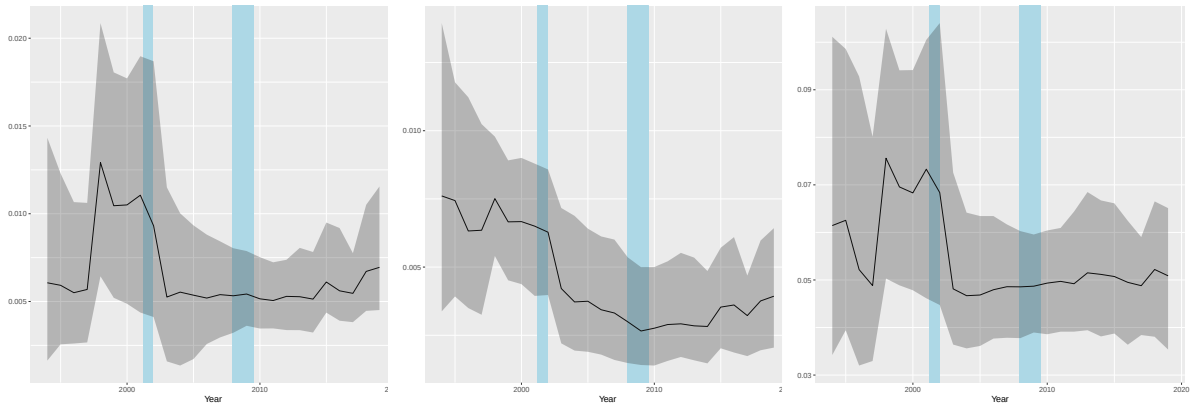
Figure IA.1: **Report Characteristics**
This figure summarizes the basic characteristics of sell-side analyst reports in our sample from 1993 through 2018. In all cases, the blue bars denote the recession periods identified with respect to the dates provided by the National Bureau of Economic Research. The solid black lines represent the cross-sectional median over each year in the data. The gray shaded areas denote the cross-sectional bottom and top 10% percentiles. Panels (a) and (b) summarize the number of words and pages in each report, respectively. Panel (c) illustrates the number of images per page over time.



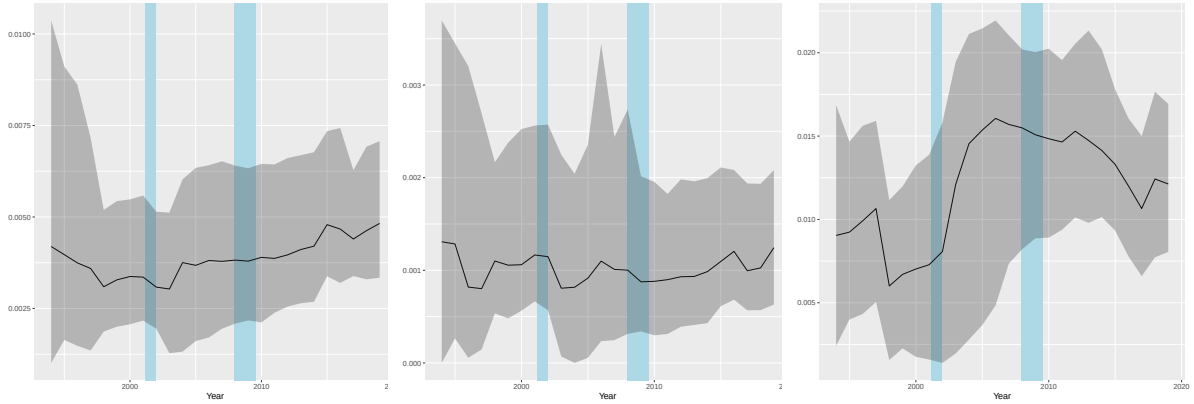
(a) Number of Words (b) Number of Pages (c) Number of Images per Page

IA.2.3 Alternative Dictionaries

Figure IA.2: **Additional Dictionaries Over Time**
This figure summarizes sell-side analyst reports in our sample from 1993 through 2018 based on the six dictionaries discussed in Section 3.1.3. In all cases, the blue bars denote recession periods identified by the National Bureau of Economic Research. The solid black lines represent the cross-sectional median for each year in the data, while the gray shaded areas denote the cross-sectional bottom and top 10th percentiles. Panels (a) and (b) plot the short- and long-term related dictionaries according to Brochet et al. (2015). Panel (c) corresponds to the financial dictionary by Matsumoto et al. (2011). Panels (d) and (e) are based on Brown et al. (2015) and Painter (2018), covering valuation and alternative data-related terms (see Table IA.1). Finally, Panel (f) covers uncertainty-related terms defined by Loughran and McDonald (2011).



(a) Short-Term (b) Long-Term (c) Financial



(d) Valuation (e) Alternative (f) Uncertainty

IA.2.4 Final Panel

Table IA.4: **Summary Statistics for Final Panel**
This table reports the summary statistics for the final firm-quarter panel. Merged altogether, the panel covers 89,454 firm-quarter observations. It ranges between 1994-Q1 and 2018-Q4, covering in total 2,333 unique stocks identified with respect to CUSIP.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Panel (a) Analyst Textual Content							
Net Tone	89,454	−0.020	0.012	−0.215	−0.026	−0.013	0.053
Uncertainty	89,454	0.013	0.006	0.000	0.009	0.016	0.111
Valuation	89,454	0.004	0.002	0.000	0.003	0.005	0.038
Alternative	89,454	0.001	0.002	0.000	0.001	0.001	0.054
Long Term	89,454	0.004	0.003	0.000	0.002	0.005	0.067
Short Term	89,454	0.007	0.004	0.000	0.004	0.008	0.090
Financial	89,454	0.054	0.019	0.000	0.044	0.059	0.367
Panel (b) Analyst Recommendation							
MEANREC	89,454	2.228	0.607	1.000	1.853	2.667	5.000
STDEV	89,454	0.617	0.398	0.000	0.380	0.890	2.830
NUMREC	89,454	7.541	6.999	1	2	10.3	56
BUYPCT	89,454	57.030	33.575	0.000	32.792	88.240	100.000
Panel (c) Fundamentals							
mvel1	89,454	0.179	0.545	−0.999	−0.249	0.643	1.000
beta	89,454	0.155	0.523	−1.000	−0.261	0.611	1.000
betasq	89,454	0.154	0.523	−1.000	−0.264	0.610	1.000
chmom	89,454	−0.004	0.596	−1.000	−0.539	0.528	1.000
dolvol	89,454	0.203	0.531	−1.000	−0.204	0.649	1.000
idiovol	89,454	0.090	0.519	−1.000	−0.337	0.533	1.000
indmom	89,454	0.021	0.598	−1.000	−0.509	0.552	1.000
mom1m	89,454	0.011	0.600	−1.000	−0.532	0.552	1.000
mom6m	89,454	0.006	0.600	−1.000	−0.534	0.548	1.000
mom12m	89,454	0.010	0.596	−1.000	−0.519	0.545	1.000
mom36m	89,454	0.024	0.563	−1	−0.4	0.5	1
turn	89,454	0.150	0.516	−0.999	−0.237	0.582	1.000
absacc	89,454	0.005	0.548	−1	−0.4	0.5	1
acc	89,454	−0.011	0.547	−1	−0.5	0.4	1
age	89,454	0.053	0.584	−1	−0.4	0.5	1
agr	89,454	−0.020	0.559	−1	−0.5	0.4	1
bm	89,454	−0.127	0.523	−1.000	−0.577	0.289	1.000
bmia	89,454	−0.048	0.525	−1.000	−0.481	0.386	1.000
cashdebt	89,454	0.003	0.568	−1.000	−0.463	0.478	1.000
cashpr	89,454	0.069	0.556	−1.000	−0.379	0.544	1.000
cfp	89,454	−0.034	0.554	−1.000	−0.503	0.415	1.000
cfpia	89,454	−0.017	0.551	−1.000	−0.480	0.441	1.000
chatoia	89,454	0.002	0.547	−1	−0.4	0.4	1
chcsho	89,454	0.042	0.563	−1	−0.4	0.5	1
chempia	89,454	0.028	0.555	−1	−0.4	0.5	1
chinv	89,454	0.026	0.556	−1	−0.4	0.5	1
chpmia	89,454	0.009	0.554	−1	−0.4	0.5	1
convind	89,454	−0.741	0.652	−1	−1	−1	1
currat	89,454	0.052	0.537	−1.000	−0.358	0.501	1.000
depr	89,454	0.033	0.529	−1.000	−0.378	0.467	1.000
divi	89,454	−0.881	0.395	−1	−1	−1	1
divo	89,454	−0.877	0.403	−1	−1	−1	1
dy	89,454	−0.077	0.579	−1	−0.5	0.4	1
egr	89,454	0.024	0.565	−1	−0.4	0.5	1
ep	89,454	−0.017	0.558	−1.000	−0.501	0.442	1.000
gma	89,454	0.032	0.563	−1	−0.4	0.5	1
grcapx	89,454	0.021	0.515	−1	−0.3	0.4	1
grltnoa	89,454	0.004	0.487	−1	−0.3	0.3	1
herf	89,454	0.024	0.564	−1.000	−0.461	0.504	1.000
hire	89,454	0.034	0.554	−1	−0.4	0.5	1
invest	89,454	0.026	0.534	−1	−0.4	0.4	1
lev	89,454	−0.060	0.554	−0.997	−0.536	0.397	1.000
lgr	89,454	0.012	0.552	−1	−0.4	0.5	1
mveia	89,454	0.036	0.571	−1.000	−0.443	0.541	1.000
operprof	89,454	0.015	0.555	−1	−0.4	0.5	1

orgcap	89,454	0.038	0.456	−1	−0.2	0.3	1
pchcapxia	89,454	0.009	0.537	−1	−0.4	0.4	1
pchcurrat	89,454	0.003	0.546	−1	−0.4	0.5	1
pchdepr	89,454	−0.008	0.535	−1	−0.4	0.4	1
pchgmpchsale	89,454	0.005	0.542	−1	−0.4	0.4	1
pchquick	89,454	0.002	0.546	−1	−0.4	0.4	1
pchsalepchinvt	89,454	−0.003	0.462	−1	−0.3	0.2	1
pchsalepchrect	89,454	0.004	0.536	−1	−0.4	0.4	1
pchsalepchxsga	89,454	0.013	0.488	−1	−0.3	0.3	1
pchsaleinv	89,454	−0.001	0.459	−1	−0.2	0.2	1
pctacc	89,454	0.028	0.534	−1	−0.4	0.5	1
ps	89,454	0.009	0.553	−1	−0.4	0.5	1
quick	89,454	0.033	0.550	−1.000	−0.403	0.497	1.000
rd	89,454	−0.614	0.745	−1	−1	−1	1
rdmve	89,454	−0.002	0.469	−1.000	−0.070	0.154	1.000
rdsale	89,454	0.006	0.465	−1.000	−0.051	0.119	1.000
realestate	89,454	0.004	0.415	−1.000	−0.049	0.088	1.000
roic	89,454	0.024	0.589	−1.000	−0.490	0.533	1.000
salecash	89,454	0.008	0.585	−1.000	−0.498	0.520	1.000
saleinv	89,454	−0.015	0.475	−1.000	−0.303	0.242	1.000
salerec	89,454	0.044	0.539	−1.000	−0.377	0.484	1.000
secured	89,454	−0.068	0.525	−1.000	−0.289	0.153	1.000
securedind	89,454	−0.040	0.986	−1	−1	1	1
sgr	89,454	0.020	0.543	−1	−0.4	0.5	1
sin	89,454	−0.960	0.229	−1	−1	−1	1
sp	89,454	−0.024	0.575	−1.000	−0.526	0.467	1.000
tang	89,454	0.020	0.559	−1.000	−0.439	0.498	1.000
tb	89,454	0.020	0.542	−1.000	−0.410	0.467	1.000
aeavol	89,454	0.042	0.536	−1.000	−0.397	0.491	1.000
cash	89,454	0.037	0.583	−1.000	−0.467	0.554	1.000
chtx	89,454	0.013	0.567	−1.000	−0.470	0.500	1.000
cinvest	89,454	0.003	0.549	−1.000	−0.443	0.450	1.000
ear	89,454	0.003	0.582	−1.000	−0.507	0.512	1.000
nincr	89,454	−0.209	0.755	−1	−1	0.4	1
roaq	89,454	0.002	0.587	−1.000	−0.527	0.510	1.000
roavol	89,454	0.001	0.548	−1	−0.4	0.4	1
roeq	89,454	0.009	0.586	−1.000	−0.515	0.515	1.000
rsup	89,454	0.011	0.550	−1.000	−0.440	0.474	1.000
stdacc	89,454	−0.020	0.507	−1	−0.4	0.3	1
stdcf	89,454	−0.017	0.509	−1	−0.4	0.3	1
ms	89,454	0.058	0.551	−1.000	−0.342	0.479	1.000
baspread	89,454	0.112	0.505	−0.999	−0.303	0.544	1.000
ill	89,454	−0.173	0.549	−1.000	−0.647	0.285	1.000
maxret	89,454	0.077	0.528	−1.000	−0.359	0.527	1.000
retvol	89,454	0.078	0.524	−0.999	−0.357	0.526	1.000
stddolvol	89,454	−0.160	0.540	−1.000	−0.629	0.269	1.000
stdturn	89,454	0.090	0.526	−0.999	−0.329	0.531	1.000
zerotrade	89,454	−0.171	0.500	−1.000	−0.588	0.202	1.000
pricedelay	89,454	−0.079	0.532	−1.000	−0.530	0.346	1.000
				Panel (d) Market			
Ret	89,454	0.010	0.103	−0.975	−0.034	0.049	5.216
RV	89,454	0.248	0.197	0.006	0.135	0.304	18.385
MKTCAP	89,454	5,906,106.000	28,528,313.000	769.200	154,942.900	2,200,104.000	1,073,390,587.000

IA.3 Additional Portfolio Sorts

IA.3.1 Single Portfolio Sorts

Table IA.5: **Single Portfolio Sorting based on Common Predictors**
At the end of each calendar year, firms are grouped into ten deciles based on 94 different firm characteristics from Gu et al. (2020). The table illustrates the risk-adjusted returns of equal-weighted portfolios. The (H)-(L) column denotes the return of a self-financing portfolio that goes long on the top decile and shorts the bottom one. In all cases, portfolios are rebalanced annually. Risk-adjusted returns are computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The full period covers portfolio time series returns from January 1998 to December 2018. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with six lags.

	(L)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)
Ret	0.066	0.040	0.025	0.038**	0.019	0.012	0.031**	0.020	0.013	0.051	−0.015
	(0.060)	(0.033)	(0.022)	(0.016)	(0.012)	(0.012)	(0.016)	(0.015)	(0.028)	(0.031)	(0.062)
RV	0.011	0.003	0.006	0.001	0.021**	0.017	0.031	0.009	0.059	0.173**	0.162**
	(0.013)	(0.011)	(0.012)	(0.012)	(0.010)	(0.017)	(0.022)	(0.024)	(0.039)	(0.073)	(0.075)
MKTCAP	0.242***	0.044	0.030	0.031	−0.017	0.004	−0.014	−0.008	0.018	0.006	−0.237***
	(0.075)	(0.043)	(0.030)	(0.021)	(0.017)	(0.015)	(0.012)	(0.015)	(0.014)	(0.009)	(0.074)
mvel1	0.241***	0.045	0.029	0.028	−0.017	0.001	−0.012	−0.002	0.020	0.006	−0.235***
	(0.071)	(0.044)	(0.028)	(0.022)	(0.017)	(0.014)	(0.010)	(0.013)	(0.014)	(0.009)	(0.070)
beta	0.037	0.030*	0.034**	0.011	0.028	0.011	0.035*	0.040	0.011	0.076	0.039
	(0.025)	(0.017)	(0.017)	(0.016)	(0.019)	(0.016)	(0.020)	(0.025)	(0.031)	(0.054)	(0.065)
betasq	0.039	0.030*	0.034**	0.011	0.028	0.011	0.036*	0.041	0.011	0.076	0.038
	(0.025)	(0.017)	(0.017)	(0.016)	(0.019)	(0.016)	(0.020)	(0.025)	(0.031)	(0.054)	(0.065)
chmom	−0.018	0.011	0.026	0.043**	0.020	0.047***	0.034*	0.068***	0.039	0.042	0.060**
	(0.024)	(0.024)	(0.021)	(0.017)	(0.018)	(0.017)	(0.018)	(0.021)	(0.026)	(0.030)	(0.028)
dolvol	0.195***	0.056	0.050*	0.033	0.014	−0.027	−0.025**	−0.003	0.012	0.023	−0.173***
	(0.052)	(0.041)	(0.029)	(0.024)	(0.018)	(0.017)	(0.013)	(0.013)	(0.015)	(0.014)	(0.050)
idiovol	0.016	0.009	0.010	−0.005	0.005	0.011	0.026	0.033	0.068	0.153**	0.137**
	(0.010)	(0.009)	(0.010)	(0.012)	(0.013)	(0.014)	(0.023)	(0.032)	(0.047)	(0.064)	(0.065)
indmom	0.079**	0.078**	0.053*	0.029	0.014	0.024	0.022	−0.012	0.013	0.012	−0.068
	(0.039)	(0.037)	(0.028)	(0.025)	(0.034)	(0.034)	(0.027)	(0.026)	(0.027)	(0.025)	(0.050)
mom1m	0.067	0.028	0.043	0.005	0.015	0.016	0.028**	0.007	0.031*	0.065*	−0.002
	(0.054)	(0.037)	(0.027)	(0.019)	(0.013)	(0.014)	(0.013)	(0.015)	(0.018)	(0.035)	(0.058)
mom6m	0.083*	0.075*	0.024	0.049**	0.027*	0.017	0.020	0.021	0.018	−0.019	−0.102**
	(0.049)	(0.038)	(0.019)	(0.019)	(0.016)	(0.013)	(0.014)	(0.019)	(0.019)	(0.025)	(0.051)
mom12m	0.152**	0.052	0.049*	0.041**	0.023*	0.003	0.017	0.015	0.007	−0.037	−0.189***
	(0.065)	(0.034)	(0.026)	(0.019)	(0.012)	(0.010)	(0.013)	(0.014)	(0.019)	(0.023)	(0.065)
mom36m	0.108*	0.030	0.035*	0.003	0.053**	0.024	0.017	−0.002	0.016	0.028	−0.080
	(0.055)	(0.031)	(0.020)	(0.016)	(0.026)	(0.023)	(0.010)	(0.014)	(0.018)	(0.024)	(0.054)
turn	0.081**	0.056**	0.050*	0.039	0.024*	0.046**	−0.003	0.015	0.002	0.005	−0.076
	(0.036)	(0.027)	(0.026)	(0.023)	(0.014)	(0.018)	(0.020)	(0.024)	(0.022)	(0.034)	(0.048)
absacc	0.033**	0.030**	0.041**	0.020	0.007	0.019	0.025	0.070**	0.022	0.044	0.011
	(0.016)	(0.015)	(0.020)	(0.013)	(0.017)	(0.020)	(0.018)	(0.029)	(0.025)	(0.047)	(0.048)
acc	0.064	0.041	0.043**	0.026	0.016	−0.003	0.037**	0.044***	0.024	0.018	−0.047

	(0.045)	(0.026)	(0.020)	(0.019)	(0.017)	(0.018)	(0.017)	(0.017)	(0.016)	(0.031)	(0.037)
age	0.035	0.046	0.009	0.032	0.048*	0.064***	0.042**	0.028*	0.007	−0.003	−0.038
	(0.034)	(0.030)	(0.025)	(0.027)	(0.025)	(0.023)	(0.017)	(0.016)	(0.010)	(0.011)	(0.036)
agr	0.004	0.001	0.063**	0.029*	0.014	0.026*	0.024*	−0.001	0.068**	0.089*	0.085*
	(0.031)	(0.020)	(0.025)	(0.015)	(0.015)	(0.014)	(0.014)	(0.016)	(0.029)	(0.051)	(0.044)
bm	0.079*	0.029	0.028	0.027	0.014	0.029	0.003	−0.003	0.040**	0.066*	−0.013
	(0.041)	(0.022)	(0.020)	(0.019)	(0.018)	(0.021)	(0.017)	(0.013)	(0.019)	(0.034)	(0.037)
bmia	−0.005	−0.011	0.041*	0.057*	0.047*	0.051*	0.028	0.011	0.062***	0.030	0.035
	(0.022)	(0.018)	(0.023)	(0.031)	(0.027)	(0.027)	(0.024)	(0.024)	(0.023)	(0.028)	(0.030)
cashdebt	0.110	0.048	0.004	0.020	0.032*	0.032*	0.016	0.001	0.033**	0.023	−0.087
	(0.068)	(0.044)	(0.022)	(0.013)	(0.017)	(0.017)	(0.013)	(0.011)	(0.013)	(0.019)	(0.060)
cashpr	0.020	0.036**	0.037	0.027	0.046	0.059*	0.027	0.040*	0.018	0.002	−0.019
	(0.018)	(0.016)	(0.030)	(0.022)	(0.034)	(0.031)	(0.024)	(0.021)	(0.015)	(0.019)	(0.025)
cfp	0.134**	0.006	0.025	0.020	0.037***	0.016	0.015	0.0002	0.011	0.053**	−0.081
	(0.064)	(0.043)	(0.021)	(0.014)	(0.013)	(0.014)	(0.012)	(0.013)	(0.015)	(0.024)	(0.056)
cfpia	0.077	−0.018	0.0002	0.028	0.015	0.042***	0.045***	0.049**	0.030**	0.042**	−0.035
	(0.048)	(0.027)	(0.024)	(0.019)	(0.016)	(0.016)	(0.016)	(0.021)	(0.015)	(0.020)	(0.043)
chatoia	0.002	0.031	0.035	0.018	0.012	0.037	0.006	0.070***	0.020	0.081***	0.079***
	(0.023)	(0.022)	(0.023)	(0.018)	(0.022)	(0.025)	(0.019)	(0.025)	(0.019)	(0.026)	(0.019)
chcsho	0.020	0.009	0.049**	0.022	0.027	0.064**	0.054**	0.061**	0.030	−0.022	−0.041*
	(0.018)	(0.012)	(0.021)	(0.018)	(0.023)	(0.025)	(0.023)	(0.025)	(0.028)	(0.027)	(0.024)
chempia	0.067*	0.044**	0.014	0.048**	0.022	0.023	0.029**	0.028	0.028	0.010	−0.058**
	(0.038)	(0.020)	(0.021)	(0.020)	(0.017)	(0.015)	(0.014)	(0.018)	(0.018)	(0.028)	(0.026)
chinv	0.040*	0.047**	0.036	0.061**	0.012	0.032	0.012	0.024*	0.012	0.037	−0.003
	(0.021)	(0.022)	(0.023)	(0.029)	(0.021)	(0.022)	(0.018)	(0.013)	(0.017)	(0.029)	(0.025)
chpmia	0.032	0.023	0.049**	0.055***	0.006	−0.014	−0.001	0.013	0.072**	0.082**	0.049*
	(0.038)	(0.024)	(0.024)	(0.018)	(0.020)	(0.022)	(0.016)	(0.021)	(0.029)	(0.033)	(0.026)
convind	0.048**	−0.002	0.033*	0.025	0.073***	0.035	0.028*	0.035*	0.023	0.014	−0.034
	(0.021)	(0.016)	(0.018)	(0.021)	(0.023)	(0.023)	(0.017)	(0.019)	(0.023)	(0.026)	(0.021)
currat	0.008	0.056***	0.025	−0.002	0.024	0.036**	0.036	0.022	0.038	0.071*	0.063*
	(0.017)	(0.018)	(0.022)	(0.018)	(0.017)	(0.016)	(0.023)	(0.021)	(0.026)	(0.039)	(0.037)
depr	−0.027	0.002	−0.001	0.004	0.027*	0.022	0.044**	0.050*	0.073***	0.117***	0.144***
	(0.025)	(0.018)	(0.018)	(0.011)	(0.016)	(0.022)	(0.022)	(0.026)	(0.025)	(0.041)	(0.040)
divi	0.052**	−0.001	0.017	0.029	0.057**	0.040**	0.014	0.032	0.038**	0.032	−0.020
	(0.021)	(0.014)	(0.020)	(0.020)	(0.024)	(0.019)	(0.021)	(0.020)	(0.018)	(0.029)	(0.020)
divo	0.053**	0.007	0.016	0.026	0.054**	0.049**	0.012	0.043*	0.024	0.026	−0.028
	(0.023)	(0.014)	(0.018)	(0.020)	(0.024)	(0.021)	(0.020)	(0.023)	(0.018)	(0.029)	(0.024)
dy	0.026	0.022	0.107***	0.043	0.047*	0.024	0.025	0.015	0.002	−0.002	−0.028
	(0.024)	(0.026)	(0.033)	(0.028)	(0.025)	(0.018)	(0.016)	(0.011)	(0.011)	(0.017)	(0.023)
egr	0.101*	0.043	0.020	0.032***	0.010	0.025**	0.016	0.025**	0.031*	0.011	−0.089**
	(0.057)	(0.031)	(0.019)	(0.011)	(0.015)	(0.011)	(0.016)	(0.011)	(0.017)	(0.030)	(0.044)
ep	0.154**	0.032	0.009	0.037**	0.019	0.021	0.027***	0.004	−0.003	0.014	−0.140**
	(0.073)	(0.047)	(0.025)	(0.019)	(0.012)	(0.014)	(0.009)	(0.011)	(0.013)	(0.020)	(0.070)
gma	0.044	0.011	0.024	0.001	0.015	0.035	0.039**	0.051***	0.066***	0.033	−0.011
	(0.054)	(0.022)	(0.019)	(0.027)	(0.021)	(0.022)	(0.016)	(0.015)	(0.017)	(0.024)	(0.057)
grcapx	0.045	0.013	0.022	0.020	0.045*	0.057***	0.026**	0.042***	0.030	0.017	−0.028
	(0.033)	(0.020)	(0.016)	(0.017)	(0.026)	(0.021)	(0.013)	(0.015)	(0.021)	(0.032)	(0.028)
grltnoa	0.051	0.042*	0.027	0.046**	0.029*	0.022*	0.043**	0.030**	0.006	0.015	−0.036
	(0.035)	(0.023)	(0.020)	(0.018)	(0.017)	(0.013)	(0.019)	(0.015)	(0.020)	(0.029)	(0.026)
herf	0.004	0.069*	0.001	0.074**	0.071**	0.061*	0.055**	−0.004	−0.024	0.009	0.004
	(0.019)	(0.040)	(0.023)	(0.034)	(0.033)	(0.032)	(0.024)	(0.021)	(0.026)	(0.019)	(0.023)
hire	0.060	0.027	0.021	0.049***	0.004	0.027*	0.040**	0.025	0.043	0.018	−0.042

	(0.039)	(0.021)	(0.019)	(0.017)	(0.018)	(0.016)	(0.017)	(0.016)	(0.027)	(0.030)	(0.029)
invest	0.041	0.022	0.062**	0.028*	0.026	0.013	0.056***	0.042**	0.026	−0.004	−0.044*
	(0.028)	(0.019)	(0.028)	(0.017)	(0.017)	(0.014)	(0.018)	(0.021)	(0.021)	(0.032)	(0.023)
lev	0.050	0.045	0.090***	0.039**	0.011	0.012	−0.004	0.010	0.028	0.030	−0.019
	(0.039)	(0.031)	(0.029)	(0.019)	(0.018)	(0.017)	(0.018)	(0.019)	(0.022)	(0.020)	(0.043)
lgr	0.051	0.031	0.044***	0.002	0.020	0.008	0.049**	0.025	0.039*	0.044	−0.007
	(0.037)	(0.022)	(0.016)	(0.011)	(0.020)	(0.014)	(0.019)	(0.015)	(0.022)	(0.035)	(0.025)
mveia	0.057*	0.015	0.077***	0.057**	0.043	0.016	0.009	−0.004	0.030*	0.019**	−0.038
	(0.033)	(0.032)	(0.029)	(0.029)	(0.029)	(0.019)	(0.014)	(0.014)	(0.016)	(0.010)	(0.032)
operprof	0.110	0.032	0.039**	0.050***	0.025	0.016	0.015	0.011	0.010	0.010	−0.100
	(0.068)	(0.035)	(0.018)	(0.019)	(0.024)	(0.020)	(0.011)	(0.009)	(0.010)	(0.015)	(0.065)
orgcap	−0.005	−0.048**	0.027	0.042**	0.062**	0.020	0.054**	0.042**	0.057***	0.064*	0.069*
	(0.017)	(0.022)	(0.019)	(0.021)	(0.028)	(0.026)	(0.023)	(0.019)	(0.020)	(0.034)	(0.037)
pchcapxia	0.033	0.042*	0.025	0.006	0.025	0.016	0.036***	0.047**	0.035	0.048	0.015
	(0.033)	(0.025)	(0.021)	(0.016)	(0.016)	(0.016)	(0.013)	(0.019)	(0.027)	(0.030)	(0.031)
pchcurrat	0.075	0.052**	0.027	0.048***	0.039**	0.004	0.022	0.021	0.030*	−0.006	−0.081***
	(0.045)	(0.026)	(0.020)	(0.015)	(0.019)	(0.013)	(0.016)	(0.015)	(0.017)	(0.027)	(0.030)
pchdepr	0.039	0.022	0.014	0.021	−0.003	0.032**	0.043***	0.039*	0.041*	0.067	0.028
	(0.035)	(0.014)	(0.011)	(0.013)	(0.016)	(0.015)	(0.016)	(0.022)	(0.024)	(0.041)	(0.027)
pchgmpchsale	0.063	0.023	0.051**	0.025*	0.013	−0.0001	0.027*	0.046***	0.032*	0.031	−0.032
	(0.042)	(0.023)	(0.020)	(0.015)	(0.023)	(0.017)	(0.015)	(0.015)	(0.018)	(0.032)	(0.026)
pchquick	0.072*	0.043*	0.055***	0.045**	0.036**	0.008	−0.005	0.032*	0.027	−0.0003	−0.073**
	(0.044)	(0.026)	(0.019)	(0.019)	(0.018)	(0.012)	(0.011)	(0.019)	(0.017)	(0.024)	(0.029)
pchsalepchinvt	0.054*	−0.004	0.027	−0.005	0.044	0.051*	0.005	0.030**	0.039**	0.073***	0.019
	(0.032)	(0.018)	(0.019)	(0.018)	(0.031)	(0.029)	(0.016)	(0.014)	(0.017)	(0.025)	(0.028)
pchsalepchrect	0.034	0.007	0.033*	0.023*	−0.008	0.023	0.032**	0.032*	0.054**	0.085**	0.050
	(0.037)	(0.021)	(0.017)	(0.013)	(0.023)	(0.022)	(0.012)	(0.016)	(0.021)	(0.037)	(0.032)
pchsalepchxsga	0.028	0.033	0.016	0.0004	0.039	0.037	0.011	0.032**	0.054***	0.064**	0.036
	(0.035)	(0.025)	(0.013)	(0.015)	(0.030)	(0.031)	(0.014)	(0.013)	(0.017)	(0.032)	(0.023)
pchsaleinv	0.045	0.006	0.029*	−0.012	0.051*	0.045*	0.006	0.026**	0.042***	0.073***	0.028
	(0.036)	(0.017)	(0.016)	(0.018)	(0.031)	(0.027)	(0.017)	(0.013)	(0.015)	(0.026)	(0.029)
pctacc	0.048**	0.038**	0.020	0.062***	0.004	0.027*	0.043*	0.052	0.029	−0.011	−0.059**
	(0.019)	(0.019)	(0.017)	(0.020)	(0.021)	(0.016)	(0.023)	(0.033)	(0.025)	(0.025)	(0.023)
ps	0.044	0.036	0.029	0.038	0.022	0.033**	0.010	0.036***	0.019	0.045***	0.0003
	(0.050)	(0.029)	(0.024)	(0.023)	(0.022)	(0.015)	(0.013)	(0.012)	(0.012)	(0.013)	(0.046)
quick	0.024	0.019	0.009	0.022	0.027	0.041**	0.024	0.019	0.058**	0.073*	0.050
	(0.023)	(0.019)	(0.013)	(0.020)	(0.019)	(0.021)	(0.019)	(0.020)	(0.029)	(0.041)	(0.038)
rd	0.031	0.001	−0.002	0.044**	0.040***	0.012	0.025*	0.015	0.058**	0.089**	0.058**
	(0.019)	(0.013)	(0.018)	(0.019)	(0.015)	(0.018)	(0.014)	(0.024)	(0.028)	(0.038)	(0.029)
rdmve	0.019	0.017	0.028**	−0.006	0.001	0.002	0.001	0.017	0.084**	0.155***	0.136***
	(0.024)	(0.018)	(0.013)	(0.019)	(0.012)	(0.017)	(0.012)	(0.026)	(0.037)	(0.054)	(0.050)
rdsale	0.019	−0.0003	0.012	−0.010	0.001	0.002	0.007	0.042*	0.146***	0.098	0.080
	(0.024)	(0.019)	(0.016)	(0.018)	(0.014)	(0.019)	(0.014)	(0.024)	(0.034)	(0.064)	(0.062)
realestate	0.059*	0.059**	0.033*	0.015	0.011	0.029	0.001	0.028	0.049***	0.024	−0.036
	(0.034)	(0.030)	(0.018)	(0.014)	(0.017)	(0.019)	(0.013)	(0.022)	(0.017)	(0.029)	(0.029)
roic	0.104	0.065*	−0.006	0.017	0.015	0.031**	0.008	0.025**	0.028**	0.035*	−0.069
	(0.075)	(0.037)	(0.026)	(0.015)	(0.013)	(0.015)	(0.013)	(0.011)	(0.014)	(0.019)	(0.069)
salecash	0.063	0.060**	0.058***	0.053***	0.019	0.017	0.016	0.024	0.018	−0.012	−0.075
	(0.060)	(0.028)	(0.020)	(0.018)	(0.017)	(0.016)	(0.019)	(0.017)	(0.016)	(0.017)	(0.058)
saleinv	0.034	0.030*	0.032*	0.006	0.050*	0.052*	0.021	0.011	0.029	0.049***	0.015
	(0.029)	(0.018)	(0.018)	(0.016)	(0.027)	(0.028)	(0.019)	(0.015)	(0.023)	(0.016)	(0.026)
salerec	0.022	0.069**	0.048	0.045**	0.049	−0.003	0.044**	0.003	0.007	0.030	0.008

	(0.017)	(0.031)	(0.030)	(0.020)	(0.032)	(0.016)	(0.022)	(0.016)	(0.019)	(0.025)	(0.028)
secured	0.0005	−0.004	0.018	0.002	0.030	0.052**	0.036*	0.030	0.077**	0.076**	0.076***
	(0.015)	(0.015)	(0.015)	(0.018)	(0.021)	(0.023)	(0.021)	(0.023)	(0.036)	(0.030)	(0.027)
securedind	0.032**	0.002	0.051***	0.007	0.019	0.041	0.020	0.044*	0.055**	0.044**	0.011
	(0.016)	(0.017)	(0.019)	(0.020)	(0.017)	(0.026)	(0.021)	(0.025)	(0.024)	(0.022)	(0.020)
sgr	0.043	0.024	0.016	0.007	−0.033*	0.058***	0.042***	0.027*	0.084***	0.047	0.004
	(0.044)	(0.023)	(0.015)	(0.014)	(0.018)	(0.021)	(0.013)	(0.016)	(0.027)	(0.033)	(0.027)
sin	0.047**	0.003	0.027	0.008	0.050**	0.061***	0.010	0.047**	0.033*	0.025	−0.022
	(0.023)	(0.015)	(0.021)	(0.017)	(0.022)	(0.018)	(0.020)	(0.021)	(0.019)	(0.021)	(0.020)
sp	0.058	0.048*	0.028	0.036**	0.039**	0.007	0.015	0.013	0.017	0.052*	−0.006
	(0.054)	(0.027)	(0.022)	(0.018)	(0.019)	(0.014)	(0.013)	(0.015)	(0.020)	(0.029)	(0.051)
tang	0.016	0.015	0.015	0.025	0.0002	0.011	0.020	0.033*	0.087**	0.095*	0.079
	(0.016)	(0.015)	(0.016)	(0.021)	(0.016)	(0.017)	(0.018)	(0.019)	(0.035)	(0.053)	(0.050)
tb	0.057*	0.040	0.063*	0.025	0.034**	0.010	0.015	0.018	0.018	0.036**	−0.021
	(0.030)	(0.042)	(0.036)	(0.023)	(0.016)	(0.019)	(0.021)	(0.018)	(0.016)	(0.015)	(0.027)
aeavol	0.056*	0.038	0.041*	0.025	0.008	0.018	0.021	0.028	0.040**	0.038	−0.018
	(0.032)	(0.025)	(0.024)	(0.017)	(0.016)	(0.016)	(0.016)	(0.019)	(0.018)	(0.023)	(0.023)
cash	−0.012	0.018	0.011	−0.002	0.011	0.006	0.045**	0.074**	0.083***	0.081	0.092*
	(0.020)	(0.016)	(0.015)	(0.015)	(0.014)	(0.016)	(0.020)	(0.031)	(0.029)	(0.056)	(0.054)
chtx	0.036	0.027*	0.050**	0.045	0.050	0.028	0.025	0.016*	0.019	0.018	−0.018
	(0.027)	(0.015)	(0.021)	(0.034)	(0.049)	(0.026)	(0.016)	(0.009)	(0.012)	(0.020)	(0.025)
cinvest	0.041	0.030*	0.026*	0.040***	0.028**	0.024	0.042**	0.039*	0.052**	−0.009	−0.050**
	(0.037)	(0.017)	(0.014)	(0.014)	(0.014)	(0.016)	(0.018)	(0.020)	(0.021)	(0.040)	(0.024)
ear	0.070*	0.025	0.028	0.046*	0.028*	0.016	0.012	0.025*	0.010	0.049**	−0.020
	(0.040)	(0.024)	(0.021)	(0.026)	(0.017)	(0.016)	(0.015)	(0.015)	(0.018)	(0.022)	(0.033)
nincr	0.037*	0.021	0.051***	0.007	0.047**	0.003	0.042*	0.061**	0.001	0.042**	0.004
	(0.019)	(0.020)	(0.019)	(0.019)	(0.019)	(0.020)	(0.023)	(0.028)	(0.019)	(0.017)	(0.015)
roaq	0.131*	0.035	0.021	0.028*	0.0004	0.025*	0.014	0.015	0.009	0.038**	−0.093
	(0.078)	(0.033)	(0.029)	(0.016)	(0.012)	(0.013)	(0.011)	(0.012)	(0.011)	(0.017)	(0.072)
roavol	0.013	0.015*	0.034***	0.015	−0.0005	0.062**	0.021	0.020	0.049	0.094	0.081
	(0.015)	(0.009)	(0.011)	(0.013)	(0.023)	(0.025)	(0.021)	(0.024)	(0.033)	(0.063)	(0.064)
roeq	0.141*	0.033	0.037	0.018	0.044***	0.012	0.004	−0.011	0.023**	0.013	−0.127*
	(0.072)	(0.036)	(0.030)	(0.013)	(0.015)	(0.010)	(0.008)	(0.013)	(0.011)	(0.018)	(0.065)
rsup	0.080**	0.063*	0.020	0.002	0.012	0.052***	0.035***	0.027**	0.006	0.017	−0.063**
	(0.035)	(0.038)	(0.032)	(0.020)	(0.017)	(0.018)	(0.012)	(0.012)	(0.019)	(0.027)	(0.030)
stdacc	0.024**	0.021	0.009	0.028*	0.046**	0.037*	0.016	0.034*	0.045	0.057	0.033
	(0.011)	(0.016)	(0.016)	(0.016)	(0.023)	(0.019)	(0.016)	(0.019)	(0.033)	(0.057)	(0.056)
stdcf	0.024**	0.024*	0.025	0.024	0.032	0.041**	0.025	0.021	0.035	0.066	0.042
	(0.011)	(0.014)	(0.017)	(0.017)	(0.021)	(0.019)	(0.015)	(0.018)	(0.036)	(0.064)	(0.064)
ms	0.061*	0.013	0.037	0.021	0.038	0.044**	0.012	0.012	0.031**	0.045***	−0.016
	(0.033)	(0.025)	(0.032)	(0.026)	(0.025)	(0.022)	(0.011)	(0.012)	(0.012)	(0.013)	(0.033)
baspread	0.020	−0.002	0.014	0.004	0.010	0.003	0.031	0.008	0.075*	0.163**	0.143*
	(0.013)	(0.011)	(0.010)	(0.012)	(0.011)	(0.015)	(0.021)	(0.026)	(0.040)	(0.074)	(0.075)
ill	0.023**	0.010	−0.015	−0.014	−0.006	−0.006	0.044*	0.053*	0.057	0.184***	0.162***
	(0.011)	(0.016)	(0.012)	(0.012)	(0.015)	(0.019)	(0.024)	(0.029)	(0.047)	(0.053)	(0.050)
maxret	−0.005	0.016	0.002	−0.001	0.039***	0.036**	0.038*	0.007	0.068**	0.121**	0.126**
	(0.015)	(0.012)	(0.016)	(0.012)	(0.012)	(0.018)	(0.021)	(0.031)	(0.031)	(0.057)	(0.062)
retvol	0.009	0.003	0.010	0.010	0.010	0.017	0.022	0.053	0.043	0.150**	0.141**
	(0.012)	(0.014)	(0.012)	(0.013)	(0.014)	(0.015)	(0.020)	(0.033)	(0.033)	(0.065)	(0.068)
stddolvol	0.028*	0.020	0.005	0.018	0.004	0.031	0.043	0.0005	0.033	0.137***	0.109**
	(0.015)	(0.016)	(0.016)	(0.019)	(0.016)	(0.021)	(0.030)	(0.030)	(0.038)	(0.041)	(0.044)
stdturn	0.049*	0.039**	0.037**	0.056**	0.027	0.028	0.023	0.044	0.016	−0.005	−0.054*

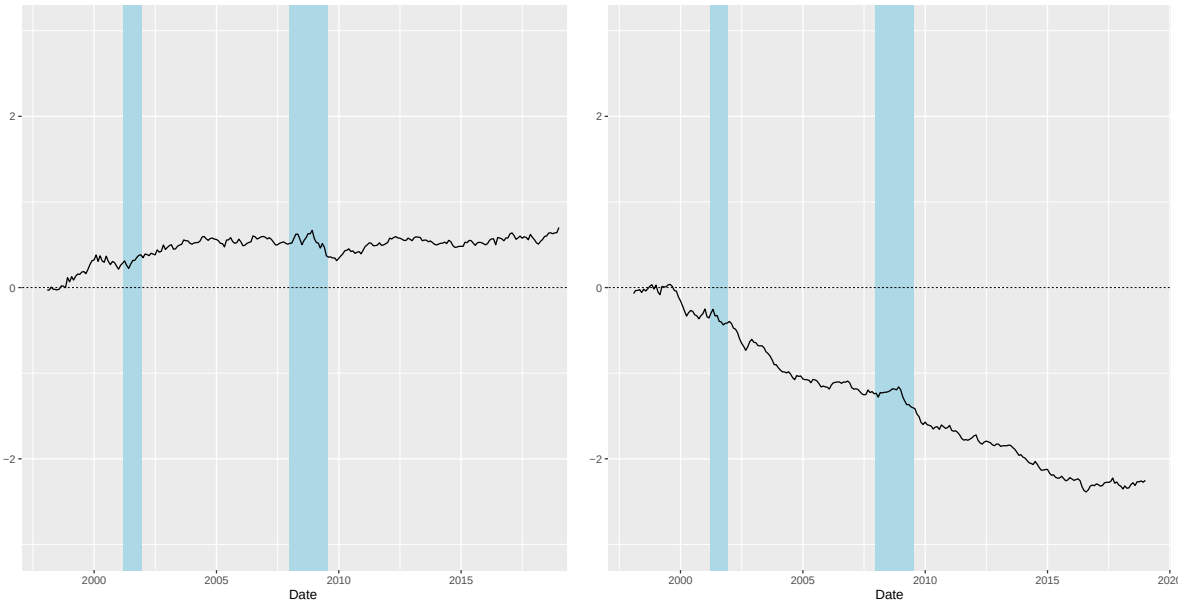
	(0.026)	(0.018)	(0.018)	(0.023)	(0.018)	(0.017)	(0.021)	(0.032)	(0.027)	(0.029)	(0.032)
zerotrade	−0.019	0.016	0.035	0.015	−0.002	0.018	0.080***	0.031	0.066**	0.074**	0.093**
	(0.032)	(0.026)	(0.023)	(0.019)	(0.017)	(0.015)	(0.023)	(0.021)	(0.028)	(0.036)	(0.044)
pricedelay	0.013	0.024	0.032**	0.025	0.027*	0.028	0.020	0.049*	0.024	0.073*	0.060
	(0.028)	(0.018)	(0.016)	(0.017)	(0.015)	(0.020)	(0.019)	(0.029)	(0.029)	(0.037)	(0.041)

Note: *p<0.1; **p<0.05; ***p<0.01

IA.3.2 Single Portfolio Sorting - Visual Illustration

In the following, we visualize the cumulative risk-adjusted returns of various long-short portfolios based on single sorting. Risk-adjusted returns are computed using the intercept from the six-factor model plus the monthly residual. Panels (a) and (b) of Figure IA.3 illustrate these cumulative returns, capturing consistent trends over time for each long-short portfolio. Notably, the slopes of these trends are steeper during the pre-2003 period, aligning with the results from Table 1. During the global financial crisis, the long-short portfolio based on STDEV shows a brief upward movement, indicating that stocks with higher uncertainty outperformed those with lower uncertainty. A similar pattern is observed during the late 2018 selloff.

Figure IA.3: **Analyst Recommendations and Future Returns**
At the end of each calendar year, Panel (a) (respectively Panel (b)) shows the case where firms are grouped into ten deciles based on the mean (respectively standard deviation) of sell-side analysts' buy/sell recommendations. For the former, a low (high) mean recommendation score denotes a strong (weak) buy recommendation, with decile 1 (10) corresponding to firms with strong (weak) recommendations. In the latter, the top decile represents stocks with the highest dispersion of recommendations, while the bottom decile represents the least dispersion. The panels illustrate the cumulative return of a self-financing portfolio that goes long on the top decile and shorts the bottom decile. Returns are equal-weighted, and portfolios are rebalanced annually. Both panels display the risk-adjusted returns (RARs) over time relative to the five risk factors from Fama and French (2015) and the momentum factor from Carhart (1997). RARs are calculated as the residuals of the six-factor regression plus the intercept.



(a) MEANREC (b) STDEV

Similar to Figure IA.3, Figure IA.4 illustrates the cumulative risk-adjusted returns of long-short portfolios sorted based on textual content, specifically net tone and uncertainty.

This experiment is consistent with the one conducted for the single sorts reported in Table 2 in the manuscript.

Figure IA.4: **Analyst Report Text and Future Returns**
At the end of each calendar year, Panel (a) (respectively Panel (b)) denotes the case in which firms are grouped into ten deciles based on the net tone (respectively uncertainty) of sell-side analyst report textual content. For net tone, a low (high) value denotes a negative (positive) tone, such that decile 1 (respectively 5) corresponds to firms with a more negative (positive) tone. For uncertainty, the top decile denotes stocks with a higher ratio of uncertainty-related terms in the analyst report, whereas the bottom decile denotes a lower ratio. The panels illustrate the cumulative return of a self-financing portfolio that longs the top and shorts the bottom decile. Returns are equal-weighted, and portfolios are re-balanced on an annual basis. Either panel displays the risk-adjusted returns (RARs) over time with respect to the five risk factors by Fama and French (2015) and the momentum factor by Carhart (1997), i.e., RARs are computed as the residuals of the six-factor regression plus the intercept.



(a) Net Tone (b) Uncertainty

IA.3.3 Double Portfolio Sorting - Additional Tables and Figures

Table IA.6 provides brief descriptive statistics to illustrate the relationship between recommendation strength and textual content. Specifically, the sample is grouped into five recommendation quintiles, and the dictionary mean for each group is presented. Firms are grouped based on their mean analyst recommendations, and the average of the extracted textual content is computed quarterly. Using the time series average of each group, we report the t-statistic testing the hypothesis that the dictionary mean for strong recommendations differs from that of weak recommendations. Several key observations emerge from Table IA.6. First, strong recommendations are associated with a higher net tone than weak recommendations, with a t-statistic of 4.1 indicating a significant difference between the two groups. Second, firms with strong recommendations exhibit higher uncertainty terms than those with weak recommendations. Third, strong recommendations are linked to a greater use of alternative-related terms than weak ones. Fourth, weaker recommendations are associated with a higher ratio of financial-related terms than stronger recommendations. Finally, all textual measures demonstrate non-linearity across recommendation deciles.

Table IA.6: **Analyst Report Textual Content by Recommendation Group**
This table reports the mean sell-side analyst report textual content for each dictionary per firm-quarter across each recommendation quintile. Each quarter, firms are grouped into five quintiles based on their analyst recommendation consensus, where S (strongest) indicates the strongest buy signal and W (weakest) indicates the weakest. The T-statistic is calculated based on the difference between the first and fifth quintiles, denoted as (S)-(W) for each text feature.

	Recommendation deciles					
Dictionary	(S)	(2)	(3)	(4)	(W)	T-stat
Net Tone	-1.604	-1.739	-1.842	-1.917	-1.837	4.104
Uncertainty	1.196	1.252	1.256	1.259	1.099	3.536
Valuation	0.420	0.422	0.416	0.422	0.418	0.316
Alternative	0.130	0.126	0.124	0.121	0.125	2.722
Long Term	0.472	0.454	0.453	0.458	0.470	0.156
Short Term	0.746	0.692	0.714	0.703	0.773	-1.107
Financial	5.575	5.415	5.612	5.614	5.877	-3.314

IA.3.4 Corner Portfolios

Next, we delve into the strategy discussed in Section 4.5, known as *corner portfolios*. This strategy corresponds to an arbitrage portfolio that goes long on both the weak-high stocks and shorts the strong-low stocks, resulting in a large risk-adjusted return. We refer to this long-short strategy as "corner portfolios" due to their position at the corner of Table 3. To quantify the returns of these corner portfolios, we present the corresponding raw and risk-adjusted returns, along with factor loadings, in Table IA.7. Notably, this self-

Table IA.7: **Corner Portfolios: Returns and Risk Factor Loadings**
This table reports the raw and risk-adjusted returns from the long-short strategy illustrated in Figure IA.5 and consistent with the results from Table 3. In particular, the corner portfolios strategy corresponds to longing the H (positive net tone) and W (weak recommendation) portfolio while shorting the L (negative net tone) and S (strong recommendation). The risk-adjusted returns are computed using the Fama-French 5 and Carhart (1997) momentum risk factors. The odd columns report the raw returns, which correspond to the intercept without any risk factors. The even columns correspond to the risk-adjusted returns. In all cases, intercepts are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with six lags.

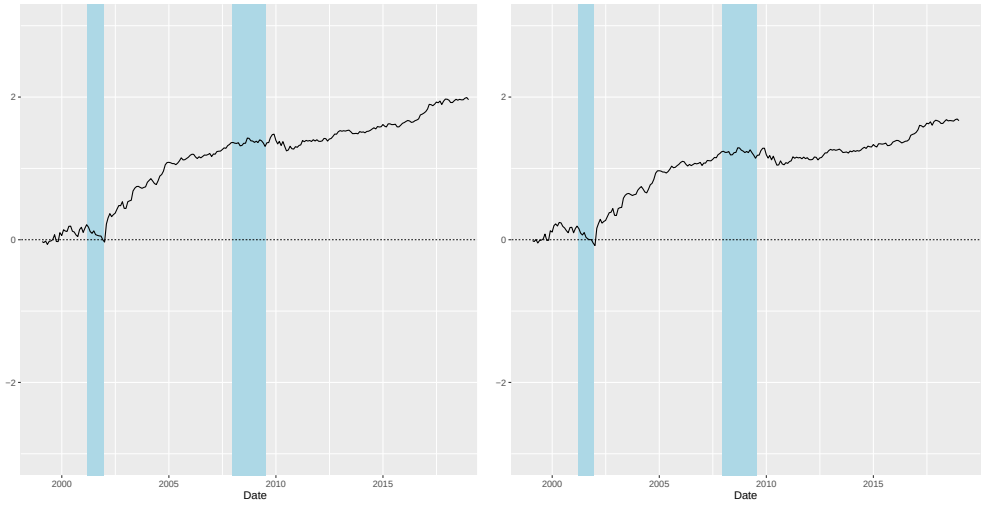
	Full Sample		Pre-2003		Post-2003	
	(1)	(2)	(3)	(4)	(5)	(6)
α	0.098***	0.083***	0.110	0.061	0.095***	0.089***
	(0.027)	(0.027)	(0.089)	(0.074)	(0.024)	(0.026)
MKT		0.068		−0.070		0.111*
		(0.072)		(0.203)		(0.061)
SMB		0.013		0.092		−0.112
		(0.084)		(0.110)		(0.110)
HML		−0.072		−0.789***		0.091
		(0.116)		(0.222)		(0.113)
RMW		0.230		0.690**		−0.054
		(0.148)		(0.277)		(0.163)
CMA		0.176		0.499		−0.007
		(0.192)		(0.309)		(0.150)
MOM		−0.053		−0.143*		0.005
		(0.070)		(0.084)		(0.055)
Observations	240	240	48	48	192	192
R^2	0.000	0.032	0.000	0.169	0.000	0.030
Note:						*p<0.1; **p<0.05; ***p<0.01

financing strategy yields an 8.3% annual risk-adjusted return over the entire sample, which is significant at the 1% level. Although the risk-adjusted return is statistically insignificant in the pre-2003 period, its magnitude is economically significant, corresponding to a 6.1% annual risk-adjusted return. In contrast, the post-2003 sample period is associated with a higher risk-adjusted return of 8.9%, which is significant at the 1% level.

To demonstrate these insights, we plot cumulative returns (both raw and risk-adjusted) of the corner portfolio strategies in Figure IA.5 of the Internet Appendix. Evidently, the strategy becomes more profitable and steady in the post-2003 period. The upward trend has been persistent since the end of the 2000 dotcom recession. At the same time, we observe that trend weakens over time, especially following the global financial crisis. Regarding risk factor loadings, the strategy is almost risk-neutral due to its arbitrage-like characteristics, with relatively low R^2. Despite that, the self-financing portfolio demonstrates statistically significant exposure to high-minus-low (HML) and robust-minus-weak (RMW) risk factors

pre-2003. However, this significance weakens after 2003. Such a decline in significance is evident across all risk factor loadings, with the only exception being the loading on the market risk factor, which is significant at the 1% level.

Figure IA.5: **Analyst Recommendation, Report Text Net Tone, and Future Returns**
At the end of each calendar year, firms are first grouped into three terciles based on mean sell-side analyst recommendations. Within each tercile, firms are further grouped into three terciles based on the net tone of analyst reports. A higher mean recommendation denotes a weaker buy signal, while a higher net tone indicates more positive sentiment. Panels (a) and (b) illustrate the cumulative raw and risk-adjusted returns, respectively, of a self-financing portfolio that goes long on weak buy and high net tone stocks while shorting strong buy and low net tone stocks. Returns are equal-weighted, and portfolios are rebalanced annually. Risk-adjusted returns are computed over time using the five risk factors from Fama and French (2015) and the momentum factor from Carhart (1997), corresponding to the residuals of the six-factor regression plus the intercept.



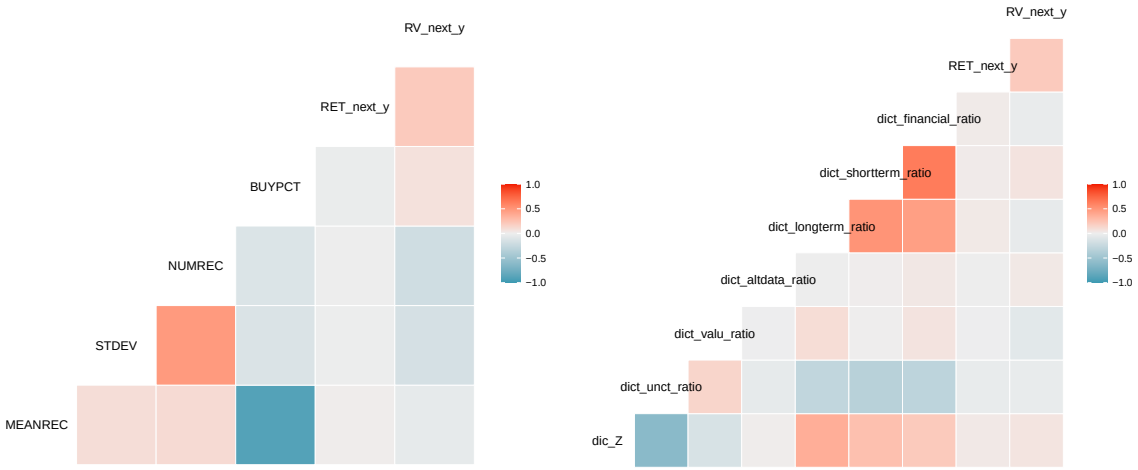
(a) Raw Returns (b) Risk-Adjusted Returns

IA.4 Machine Learning Results

IA.4.1 Correlation Matrices

Figures IA.6 and IA.7 illustrate the correlation matrices among the main predictors and target variables using a heatmap. The heatmap displays the correlation coefficients among the variables, with red indicating positive relationships, blue indicating negative relationships, and darker shades representing stronger correlations. In all cases, we observe weak correlations with next year's stock return but slightly stronger relationships with next year's realized volatility. This finding aligns with the literature, suggesting that volatility tends to be more persistent over time than mean returns. Despite the weak correlations between the predictors and target variables, our goal is to develop an optimal mapping. A complex nonlinear model can capture dynamics and relationships by design that simple measures like Pearson correlation cannot. Therefore, we expect machine learning to integrate complex information and deliver strong discriminative power for portfolio formation.

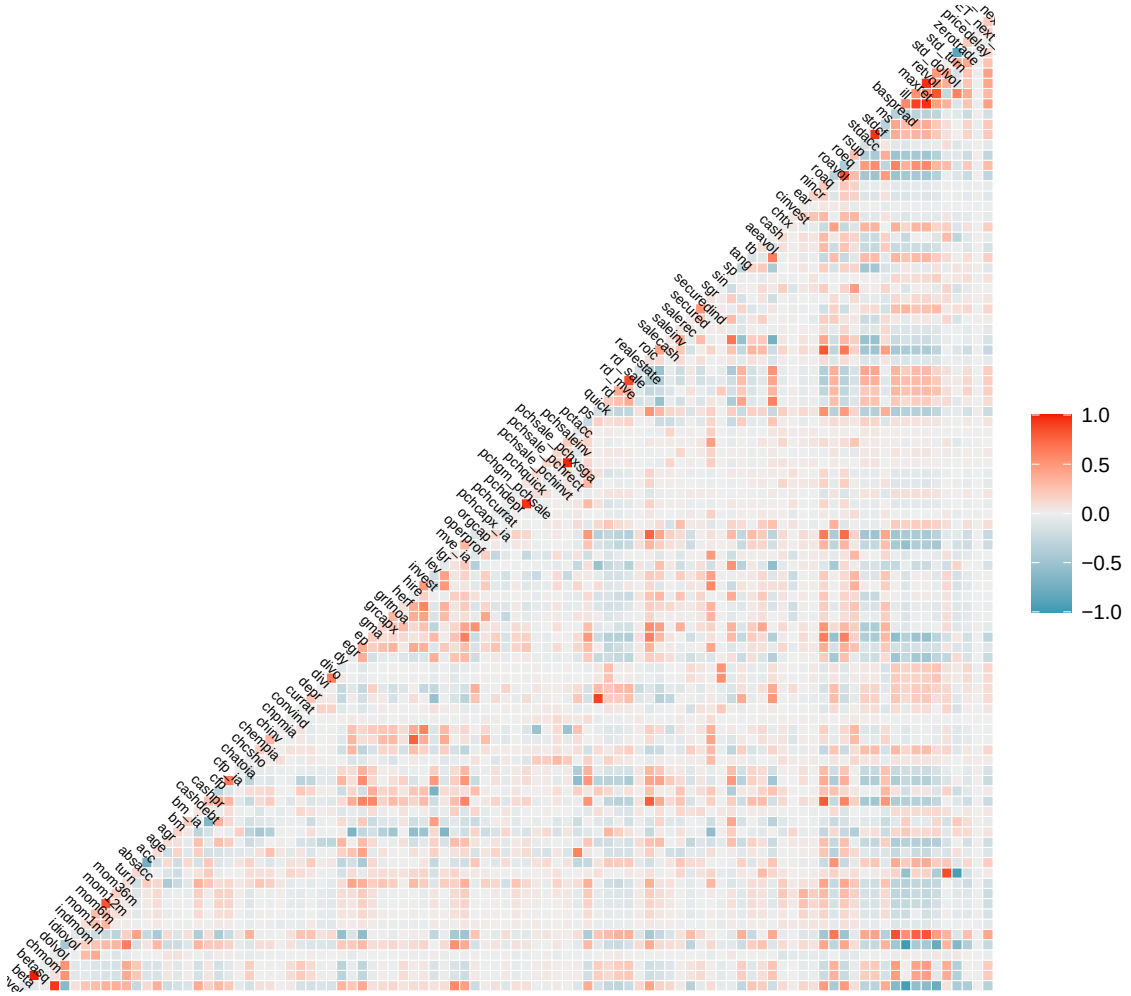
Figure IA.6: **Correlation Matrices - Analysts Data**
This figure shows the correlation between the predictors and the dependent variables across two categories. Panel (a) illustrates the correlations among the sell-side analyst recommendation variables, while Panel (b) demonstrates those for the analyst report textual content. In all cases, the dependent variables refer to either the next quarter's return or the next quarter's realized volatility, both of which are represented in the last two columns of each figure.



(a) Analyst Recommendations

(b) Analyst Report Text

Figure IA.7: **Correlation Matrices - Market and Fundamental Data**
This figure shows the correlation between the predictors and the dependent variables, where the predictor set corresponds to 94 firm characteristics from Gu et al. (2020). The dependent variables refer to either the next quarter's return or the next quarter's realized volatility, both of which are located in the last two columns of each figure, respectively.



IA.4.2 Additional Algorithms

As a robustness check, we introduce four additional ML algorithms. These algorithms represent simplified versions of the random forest (RF) and artificial neural network (ANN) models, including support vector regression (SVR). We discuss these models and their results below.

The first ML algorithm is SVR (Drucker et al., 1997), which identifies specific features in the data that can distinguish the response variable's attributes within a particular level of confidence, denoted by C. This level is also known as the margin, which provides a a degree of confidence that the model is not overfit and is robust against outliers. Additionally, the algorithm deploys a radial basis function kernel that maps the original data into a new space, where the attributes of the response variable can be better identified. This is commonly known as the "kernel trick." Additionally, one can think of the SVR model with a radial kernel as a plain vanilla ANN, in which the original data are mapped into a new set of hidden features (neurons) that are eventually mapped into the output variable (e.g., future realized volatility). The purpose of using either the returns or realized volatility as a response variable is to supervise the algorithm, such that the features are combined into a single index in a "trained" manner. The algorithm depends on two hyperparameters: the margin (confidence) and one free parameter related to the kernel. The latter is commonly referred to as intensity.

The second model is the same RF algorithm used in the manuscript, whereas the main difference is that we replicate the results from Table 4 with predefined hyperparameters, setting the number of trees to 1,000 and the number of sampled features to 10. This choice aligns with the standard practice in machine learning literature. Specifically, the rule of thumb suggests the floor of the square root of the total number of features (James et al., 2013). This modified RF model is denoted as *RF2*. Finally, for the other two ANN algorithms, we explore a single-layer ANN (referred to as *NNET*) and another variant (*NNETpca*) where the algorithm incorporates Principal Component Analysis (PCA) before processing the data. The two additional ANN models are tuned using the same cross-validation procedure conducted for the other models studied in the main manuscript.

The results for SVR, RF2, NNET, and NNETpca are detailed in Tables IA.8, IA.9, IA.10, and IA.11, respectively. Several noteworthy observations emerge. First, there is no statistical evidence across the three models supporting the inclusion of analysts' data, i.e., we do have sufficient evidence to reject the claim that $\pi(M) - \pi(1) = 0$ for $M = 2, 3, 4$. In the case of the NNETpca model, we even observe a deterioration in the performance of the (H)-(L) portfolio when incorporating analysts' recommendations. Second, our analysis concludes that the RF algorithm with predetermined hyperparameters yields the highest performance

when evaluating the long-short portfolio within the baseline feature space, resulting in a substantial 21.9% annual risk-adjusted return.

Table IA.8: **Machine Learning Predictions and Future Returns Using Support Vector Regression (SVR)**
At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is nonlinear and utilizes support vector regression (SVR) with a radial kernel (Drucker et al., 1997). For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	0.005	0.021	0.002	−0.013	0.020	0.027**	0.025	0.025	0.065***	0.144***	0.138***	
	(0.042)	(0.026)	(0.020)	(0.013)	(0.017)	(0.012)	(0.016)	(0.020)	(0.024)	(0.041)	(0.050)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.010	−0.023	0.034	0.009	0.020	0.021	0.025*	0.028	0.060***	0.159***	0.168***	0.030
	(0.041)	(0.023)	(0.025)	(0.015)	(0.015)	(0.016)	(0.015)	(0.020)	(0.022)	(0.047)	(0.060)	(0.022)
						Market and Fundamentals plus Analyst Text						
3	0.020	0.009	0.005	0.029	0.028	−0.003	0.017	0.016	0.079***	0.124***	0.105**	−0.034*
	(0.045)	(0.023)	(0.022)	(0.020)	(0.017)	(0.014)	(0.016)	(0.015)	(0.029)	(0.038)	(0.052)	(0.019)
						Altogether						
4	0.007	−0.003	0.021	0.025	0.028*	0.017	0.011	0.011	0.054***	0.152***	0.145**	0.006
	(0.044)	(0.026)	(0.023)	(0.017)	(0.016)	(0.016)	(0.018)	(0.015)	(0.019)	(0.049)	(0.061)	(0.024)
						Text Alone						
5	0.035	0.009	0.038	0.046**	0.022	0.010	0.032**	0.037*	0.020	0.065***	0.030	−0.108**
	(0.035)	(0.022)	(0.026)	(0.019)	(0.020)	(0.017)	(0.015)	(0.021)	(0.018)	(0.024)	(0.033)	(0.050)
						Recommendation Data Alone						
6	−0.016	0.032	0.040*	0.052**	0.023	0.042**	0.035	0.039*	0.047**	0.020	0.036	−0.103*
	(0.027)	(0.020)	(0.022)	(0.023)	(0.018)	(0.020)	(0.021)	(0.020)	(0.019)	(0.020)	(0.023)	(0.053)
						Text and Recommendations Alone						
7	0.019	0.033	0.023	0.042**	0.010	0.058***	0.034*	0.045***	0.031*	0.017	−0.003	−0.141**
	(0.039)	(0.028)	(0.016)	(0.020)	(0.017)	(0.017)	(0.019)	(0.016)	(0.018)	(0.021)	(0.039)	(0.055)

Note: *p<0.1; **p<0.05; ***p<0.01

Table IA.9: **Machine Learning Predictions and Future Returns Using Random Forests (RF2) with Fixed Parameters**
At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is nonlinear and utilizes random forests (Breiman, 2001). This analysis replicates Table 4, with the key difference being that the algorithm is trained using fixed hyperparameters rather than cross-validation. For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	−0.043	−0.020	0.007	0.003	−0.010	−0.003	0.040**	0.055**	0.112***	0.176***	0.219***	
	(0.043)	(0.032)	(0.024)	(0.019)	(0.018)	(0.013)	(0.015)	(0.023)	(0.040)	(0.050)	(0.058)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.040	−0.030	0.015	0.008	−0.008	0.007	0.014	0.066***	0.101***	0.187***	0.227***	0.007
	(0.045)	(0.030)	(0.024)	(0.020)	(0.015)	(0.017)	(0.015)	(0.021)	(0.033)	(0.056)	(0.068)	(0.017)
						Market and Fundamentals plus Analyst Text						
3	−0.044	−0.027	0.004	0.009	0.019	0.023	0.022	0.034	0.098***	0.180***	0.224***	0.004
	(0.045)	(0.028)	(0.024)	(0.020)	(0.018)	(0.016)	(0.015)	(0.024)	(0.033)	(0.050)	(0.066)	(0.018)
						Altogether						
4	−0.057	−0.007	0.012	0.013	0.019	0.006	0.016	0.048**	0.099***	0.171***	0.228***	0.009
	(0.043)	(0.028)	(0.023)	(0.022)	(0.017)	(0.015)	(0.014)	(0.023)	(0.038)	(0.048)	(0.061)	(0.017)
						Text Alone						
5	0.023	0.033	0.024	0.067***	0.046**	0.021	0.010	0.024	0.023	0.042**	0.019	−0.201***
	(0.033)	(0.025)	(0.018)	(0.019)	(0.021)	(0.022)	(0.017)	(0.022)	(0.017)	(0.019)	(0.029)	(0.055)
						Recommendation Data Alone						
6	0.010	0.001	0.021	−0.005	0.038**	0.062**	0.073***	0.047**	0.041**	0.026	0.016	−0.204***
	(0.024)	(0.018)	(0.017)	(0.013)	(0.018)	(0.026)	(0.020)	(0.023)	(0.020)	(0.022)	(0.018)	(0.058)
						Text and Recommendations Alone						
7	0.027	0.028	0.040*	0.017	0.036**	0.015	0.046**	0.045**	0.031	0.025	−0.003	−0.222***
	(0.031)	(0.023)	(0.021)	(0.021)	(0.017)	(0.014)	(0.022)	(0.019)	(0.021)	(0.020)	(0.028)	(0.061)

Note: *p<0.1; **p<0.05; ***p<0.01

Table IA.10: **Machine Learning Predictions and Future Returns Using Single-Layer Neural Network (NNET)**
At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is nonlinear and utilizes a single-layer artificial neural network (ANN) (Rumelhart et al., 1985). For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	−0.009	−0.023	0.016	0.005	0.034*	0.022	0.049**	0.045*	0.043*	0.145***	0.153**	
	(0.038)	(0.029)	(0.025)	(0.019)	(0.020)	(0.018)	(0.023)	(0.024)	(0.025)	(0.047)	(0.067)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.023	−0.030	−0.014	−0.002	0.039**	0.027	0.041**	0.070***	0.070***	0.137***	0.160**	0.006
	(0.044)	(0.033)	(0.025)	(0.015)	(0.019)	(0.017)	(0.020)	(0.026)	(0.024)	(0.043)	(0.067)	(0.034)
						Market and Fundamentals plus Analyst Text						
3	0.011	0.007	−0.0001	0.003	0.012	0.016	0.033	0.031	0.065**	0.149***	0.138*	−0.015
	(0.047)	(0.034)	(0.023)	(0.019)	(0.018)	(0.017)	(0.022)	(0.025)	(0.031)	(0.045)	(0.075)	(0.041)
						Altogether						
4	−0.002	−0.014	0.0001	0.019	0.014	0.028	0.037	0.043	0.053*	0.142***	0.144***	−0.009
	(0.031)	(0.027)	(0.018)	(0.017)	(0.017)	(0.017)	(0.023)	(0.026)	(0.031)	(0.042)	(0.055)	(0.039)
						Text Alone						
5	0.043	0.041	0.051**	0.036*	0.023	0.021	0.004	0.001	0.033*	0.063**	0.020	−0.134**
	(0.030)	(0.029)	(0.023)	(0.018)	(0.015)	(0.020)	(0.019)	(0.017)	(0.020)	(0.029)	(0.035)	(0.057)
						Recommendation Data Alone						
6	0.022	−0.003	0.018	0.028	0.033	0.020	0.023	0.031	0.049**	0.097***	0.075**	−0.079
	(0.028)	(0.024)	(0.019)	(0.019)	(0.020)	(0.017)	(0.017)	(0.021)	(0.024)	(0.026)	(0.030)	(0.060)
						Text and Recommendations Alone						
7	0.013	−0.007	0.035*	0.037**	0.039*	0.043**	0.032*	0.032*	0.014	0.082***	0.069*	−0.085
	(0.029)	(0.023)	(0.018)	(0.018)	(0.023)	(0.022)	(0.018)	(0.019)	(0.019)	(0.027)	(0.037)	(0.054)

Note: *p<0.1; **p<0.05; ***p<0.01

Table IA.11: **Machine Learning Predictions and Future Returns Using pcaNNet**

At the end of each calendar year, firms are grouped into ten deciles based on the machine learning (ML) forecast. The ML algorithm is nonlinear and utilizes a single-layer artificial neural network (ANN) (Rumelhart et al., 1985). This algorithm is the same as NNET, with the main difference being the application of Principal Component Analysis (PCA) before feeding the full feature space into the mapping function. For more details on the ML mapping process, see Section 5.2. The M column denotes the data used to generate the ML forecast and, consequently, the portfolio formation. Baseline data capturing market and fundamentals are denoted by $M = 1$. Specifications $M = 2$ and $M = 3$ include baseline data plus analyst recommendations and report text, respectively. Specification $M = 4$ utilizes all data sources for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The table reports the risk-adjusted returns of equal-weighted portfolios, computed using the five risk factors from Fama and French (2015) and the momentum risk factor from Carhart (1997). The (H)-(L) column shows the return of a self-financing portfolio that goes long on the top decile of ML forecasts and shorts the bottom decile. The column $\pi(M) - \pi(1)$ evaluates the incremental value of data specifications $M = 2, 3, 4, 5, 6, 7$ relative to the baseline. In all cases, portfolios are rebalanced annually. The full period covers portfolio time series returns from January 1998 to December 2018, with the initial training sample spanning 1994 to 1997. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using the method by Newey and West (1986) with six lags.

M	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(H)	(H) - (L)	$\pi(M) - \pi(1)$
						Market and Fundamentals						
1	−0.002	−0.0004	−0.028	0.033*	0.015	0.028*	0.016	0.058**	0.065**	0.147***	0.150**	
	(0.046)	(0.037)	(0.024)	(0.019)	(0.017)	(0.017)	(0.016)	(0.027)	(0.032)	(0.047)	(0.076)	
						Market and Fundamentals plus Analysts Recommendation						
2	−0.015	−0.008	0.026	0.003	0.031*	0.021	0.034**	0.041	0.069**	0.124***	0.139*	−0.010
	(0.044)	(0.036)	(0.022)	(0.021)	(0.016)	(0.017)	(0.016)	(0.027)	(0.030)	(0.046)	(0.073)	(0.021)
						Market and Fundamentals plus Analyst Text						
3	0.022	−0.016	0.021	0.023	0.008	0.030	0.031	0.021	0.060**	0.127***	0.104	−0.045**
	(0.049)	(0.031)	(0.025)	(0.020)	(0.014)	(0.021)	(0.020)	(0.025)	(0.030)	(0.049)	(0.082)	(0.023)
						Altogether						
4	0.016	−0.019	−0.014	0.017	0.012	0.025	0.046**	0.033	0.076**	0.129**	0.113	−0.037
	(0.049)	(0.034)	(0.025)	(0.022)	(0.015)	(0.016)	(0.020)	(0.025)	(0.032)	(0.050)	(0.080)	(0.034)
						Text Alone						
5	0.059*	0.043**	0.066**	0.047**	0.021	0.034	−0.002	0.011	0.012	0.024	−0.035	−0.185**
	(0.033)	(0.021)	(0.026)	(0.020)	(0.019)	(0.024)	(0.017)	(0.020)	(0.020)	(0.023)	(0.040)	(0.076)
						Recommendation Data Alone						
6	−0.007	−0.005	0.017	0.025	0.036	0.023	0.046**	0.028	0.046**	0.108***	0.116***	−0.034
	(0.030)	(0.022)	(0.019)	(0.017)	(0.022)	(0.017)	(0.022)	(0.022)	(0.023)	(0.023)	(0.028)	(0.070)
						Text and Recommendations Alone						
7	0.024	0.024	0.009	0.021	0.033	0.042**	0.048***	0.036**	0.023	0.056***	0.033	−0.117
	(0.033)	(0.021)	(0.019)	(0.022)	(0.023)	(0.018)	(0.018)	(0.017)	(0.021)	(0.020)	(0.031)	(0.074)

Note: *p<0.1; **p<0.05; ***p<0.01

IA.4.3 Sensitivy Analysis

Table IA.12 reveals several important insights. First, out of all the features, we observe that market and fundamental data matter the most. Second, even though the MLP is the only algorithm that demonstrates marginal improvement of 4% per annum when adding both analysts' recommendations and text data, we note none of these features rank at the top 20, on average. Third, when including both analysts' content on top of the market and fundamental data in the RF model, we do not observe marginal improvement. However, as Table 4 demonstrates, we find some marginal value by adding recommendation data. In this regard, we observe that both the number of recommendations and average recommendation buy signal rank in the top 20 in terms of importance, as the sensitivity analysis from Table IA.12 indicates. Fourth, the majority of the market and fundamental features that matter the most for return predictions can be linked to the following attributes: volatility, prior returns (including momentum), illiquidity, systematic risk (beta), cash flows (valuation), and size. Finally, for the realized volatility predictions, among the reported features, prior realized volatility matters the most. This is not surprising, given the autoregressive nature of volatility.

Table IA.12: **Sensitivity Analysis of Important Features**
This table reports summary statistics of individual feature importance based on two nonlinear machine learning algorithms, random forests (RF in line with Table 4) and multilayer perceptron (MLP in line with Table 5). Importance is approximated using the sensitivity score, $S_{y,v}$, from Equation (6.1) of the manuscript. The importance scores are computed using the Permutation Feature Importance according to Biecek (2018). A larger sensitivity score implies higher importance for the corresponding variable. The scores are ordered with respect to the first column under *Returns*, and their values denote the average score over time. For further information, see Section 6.2. The columns correspond to two experiments: returns (first four columns) and realized volatility (last four columns). In the former, the response variable is next year's returns. In the latter, the response variable is the next year's realized volatility. Regardless of the experiment or algorithm, the sensitivity analysis corresponds to the full feature space that includes all four data sources (i.e., $M = 4$ feature space) in line Figure 4.

		Returns				Realized Volatility			
	Variable	Mean	Std	Min	Max	Mean	Std	Min	Max
		Random Forests							
1	RET_q	27.63	40.78	0.13	182.02	4.04	7.57	0.12	34.11
2	baspread	26.47	33.30	0.63	121.23	6.35	3.69	1.75	16.99
3	RV	24.49	25.28	0.60	89.83	13.84	7.76	1.25	33.01
4	retvol	17.13	22.24	0.21	91.37	2.22	1.00	0.65	4.05
5	MKTCAP	17.12	19.71	0.61	78.87	0.71	0.75	0.13	3.49
6	dy	16.82	19.07	0.01	57.54	0.47	0.78	0.00	2.45
7	mom12m	16.40	15.65	0.69	58.96	0.37	0.54	0.02	2.44
8	idiovol	15.71	17.66	0.04	64.79	1.25	1.12	0.05	3.77
9	sp	14.38	28.22	0.25	107.41	0.23	0.44	0.00	2.06
10	mvel1	11.96	12.77	0.54	55.16	0.78	0.84	0.02	3.14
11	ill	10.36	14.12	0.09	67.43	0.45	0.41	0.02	1.64
12	depr	10.17	15.98	0.03	68.07	0.18	0.37	0.00	1.71
13	cfp	9.75	15.58	0.20	69.45	0.21	0.29	0.02	1.27
14	NUMREC	9.42	11.69	0.44	42.33	0.34	0.69	0.00	2.87
15	mom1m	9.33	10.90	0.27	47.29	0.17	0.21	0.00	1.01
16	MEANREC	9.21	19.72	0.11	91.44	0.08	0.13	0.00	0.57
17	maxret	9.09	7.28	1.92	31.62	0.53	0.68	0.01	2.46
18	dolvol	9.04	14.07	0.62	67.79	0.41	0.40	0.01	1.50
19	mom6m	8.37	6.43	0.13	26.33	0.38	0.49	0.02	2.04
20	indmom	8.24	6.58	0.03	21.86	0.17	0.20	0.00	0.87
		Multilayer Perceptron							
1	retvol	62.55	129.78	1.01	515.04	3.18	3.95	0.03	14.18
2	RET_q	54.45	108.85	1.06	504.37	4.83	8.85	0.02	40.30
3	RV	53.88	105.08	4.92	488.69	65.22	32.71	0.20	98.82
4	baspread	43.83	68.99	0.00	273.13	18.48	23.93	0.01	87.23
5	zerotrade	37.28	65.65	0.75	249.99	1.14	1.13	0.01	4.86
6	betasq	36.12	34.26	1.08	110.72	8.32	8.41	0.11	29.71
7	mom12m	35.99	26.22	1.35	101.00	0.85	0.94	0.02	3.34
8	roaq	34.44	95.18	1.35	446.46	3.80	2.36	0.12	10.04
9	mvel1	31.76	31.02	1.11	143.72	14.34	14.73	0.76	48.63
10	mom6m	30.66	35.63	0.11	132.70	0.94	1.11	0.01	3.78
11	beta	28.94	35.46	0.16	138.39	2.40	2.49	0.02	8.27
12	cfp	27.82	30.09	1.24	118.42	0.28	0.38	0.00	1.37
13	cash	26.56	34.88	0.66	144.77	0.72	0.98	0.00	3.08
14	ep	26.01	50.40	0.68	193.80	0.78	0.86	0.02	3.58
15	std_turn	24.56	34.35	0.93	126.95	1.73	2.12	0.02	8.21
16	idiovol	24.54	23.05	0.16	88.99	4.13	4.82	0.01	19.80
17	dolvol	24.30	34.77	0.56	156.84	5.09	7.55	0.05	30.46
18	ill	23.17	46.08	1.80	214.47	4.62	4.74	0.19	16.46
19	turn	23.09	24.25	0.05	75.12	1.82	2.50	0.01	10.62
20	mom1m	22.62	39.11	1.47	179.12	0.55	0.90	0.03	3.95

IA.5 Mean-Variance Portfolio Analysis

By design, the ML experiment forms forecasts for stock return and realized volatility. While the findings of Table 5 demonstrate an advantage of utilizing return forecasts based on the analyst's data, it does not utilize realized volatility forecasts. Intuitively, if the model is accurate, a higher return forecast implies higher future portfolio returns. Nonetheless, the answer is less clear for volatility (see, e.g., Moreira and Muir (2017)). To utilize both return and volatility forecasts, we consider five mean/variance portfolio rules commonly studied in the literature (see, e.g., Kan et al. (2022)). These rules rely on mean returns as well as volatilities. We describe each rule briefly below.

IA.5.1 Mean-Variance Portfolio

The first rule is the mean-variance (henceforth MV) portfolio without short sales. We denote this rule by MV_NS. This relies on both the mean vector and covariance matrix of asset returns. The former, denoted by $\hat{\boldsymbol{\mu}}_{y+1}$, is determined by the ML return forecasts, whereas the diagonal elements of the covariance matrix, denoted by $\hat{\boldsymbol{\Sigma}}_{y+1}$, are estimated using the squared realized volatility forecast. For the off-diagonal covariance elements, we set their values to zero. We do so for brevity and to avoid the complication of forecasting pairwise correlation among the stock returns. Additionally, this simplification can be justified from a shrinkage point of view (Ledoit and Wolf, 2004).

Specifically, this decision rule is the vector of weights $\mathbf{w}_y$ at the end of year y that solves the following optimization problem:

$$\begin{aligned} \max_{\mathbf{w}_y} \quad & \mathbf{w}_y^\top \hat{\boldsymbol{\mu}}_{y+1} - \frac{\gamma}{2} \mathbf{w}_y^\top \hat{\boldsymbol{\Sigma}}_{y+1} \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_y^\top \mathbf{1} = 1 \\ & \mathbf{w}_{y,i} \geq 0 \, \forall i = 1, ..., d, \end{aligned} \tag{IA1}$$

where $\mathbf{1}$ denotes a column vector of ones, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ denote, respectively, the ML forecast for the stock returns and realized covariance matrix. The parameter γ denotes the risk aversion, which we set to $\gamma = 3$ in line with Kan et al. (2022). The first budget constraint states that the portfolio weights sum to one, whereas the second one assures that the portfolio weights take non-negative values, i.e., no short sales.

IA.5.2 Global Minimum Variance Portfolio

The global minimum variance (henceforth GMV) portfolio relies only on the covariance matrix. Hence, given our ML analysis, we can utilize the volatility forecasts alone. To compute the GMV portfolio, we solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}_y} \quad & \mathbf{w}_y^\top \hat{\mathbf{\Sigma}}_{y+1} \mathbf{w}_y, \\ \text{s.t.} \quad & \mathbf{w}_y^\top \mathbf{1} = 1, \\ & \mathbf{w}_{y,i} \geq 1 \, \forall i = 1, ..., d. \end{aligned} \tag{IA2}$$

Without short sales, the GMV portfolio is given by the following closed-form solution:

$$\mathbf{w}_y = \frac{\hat{\mathbf{\Sigma}}_{y+1}^{-1} \mathbf{1}}{\mathbf{1} \hat{\mathbf{\Sigma}}_{y+1}^{-1} \mathbf{1}}. \tag{IA3}$$

In our experiment, we consider two GMV rules. The first one is with short sales, given by Equation (IA3), which we simply refer to as GMV. The second one, which we denote by GMV_NS, is the GMV portfolio without short sales. The latter rule requires numerical optimization to solve the problem stated in Equation (IA2).

IA.5.3 Volatility and Reward-to-Risk Timing

The other two rules are related to volatility and risk-to-reward timing (Kirby and Ostdiek, 2012). These rules can be viewed as ad-hoc modifications of the above rules. Specifically, in line with Kan et al. (2022), the volatility timing rule is given by

$$\mathbf{w}_{y,i} = \frac{\left(\frac{1}{\hat{\sigma}_{y+1,i}^2} \right)^\eta}{\sum_{j=1}^d \left(\frac{1}{\hat{\sigma}_{y+1,j}^2} \right)^\eta} \tag{IA4}$$

For instance, for $\eta = 1$, the above rule is consistent with the GMV portfolio since we set the off-diagonal elements of the covariance matrix to zero. We denote this rule by KO_VT as an abbreviation for Kirby and Ostdiek (2012) and volatility timing. On the other hand, the risk-to-reward timing rule is given by

$$\mathbf{w}_{y,i} = \frac{\left(\frac{\hat{\mu}_{y+1,i}^+}{\hat{\sigma}_{y+1,i}^2} \right)^\eta}{\sum_{j=1}^d \left(\frac{\hat{\mu}_{y+1,j}^+}{\hat{\sigma}_{y+1,j}^2} \right)^\eta} \tag{IA5}$$

where $\hat{\mu}^{+}_{y+1,i}$ is the return forecasts that is non-negative. In other words, if the forecast is negative, the rule sets its values to zeros. This restriction assures that the portfolio rule does not involve short positions. We denote this rule by KO_RT. Finally, we set $\eta = 4$ for both rules in line with the empirical analysis by Kan et al. (2022).

IA.5.4 Evaluation

In line portfolio sorting, we follow a similar experiment with two major distinctions, however. First, rather than using an equally weighted portfolio, the following analysis value-weights the assets based on the above five rules. Second, we focus on a subset of the stocks. In particular, at the end of each year, we choose the largest 100 stocks in terms of market capitalization (see, e.g., DeMiguel et al. (2009); Kan et al. (2022)). For each stock, we utilize the ML forecasts to form the portfolio rules and evaluate their monthly returns over the next year. At the end of the following year, we repeat the same procedure where we choose the largest market cap, utilize their stock and volatility forecasts, and evaluate their performance over the next 12 months. We repeat this procedure until the end of the sample period. As a robustness check, the experiment is also conducted using the smallest 100 market cap stocks.

Given the out-of-sample portfolio returns, we evaluate the marginal contribution of each data. Specifically, we compute the certainty-equivalent return (CER) for different portfolio rules, ML algorithms, feature spaces, and both stock sizes. The CER for a risk-averse investor based on model M is given

$$CER(a, M) = \mu_p(a, M) - \frac{\gamma}{2}\sigma_p^2(a, M), \tag{IA6}$$

where $\mu_p(a, M)$ and $\sigma_p^2(a, M)$ denote the out-of-sample portfolio mean return and volatility based on feature space M and ML algorithm a. Similar to our main analysis, we compute $CER(M) - CER(1)$, where $CER(1)$ is the CER return of the baseline model that does not include any analyst data. This incremental value is similar to the metric used by Farboodi et al. (2024) in valuing financial data. Table IA.13 reports the marginal CER values for all specifications, where Panel (a) (respectively Panel (b)) corresponds to the largest (respectively smallest) 100 stocks. Note that the regression results presented in Table 7 of the manuscript are established using the values reported in Table IA.13.

Table IA.13: **Valuing Data using Certainty Equivalent Returns for Mean-Variance Portfolio Rules**
The following table reports the relative performance of portfolios that are value-weighted returns with respect to five different rules. The MV_NS denotes a mean-variance rule that does include short sales, whereas the GMV and GMV_NS denote the global minimum variance portfolio with and without short sales, respectively. The other two rules KO_VT and KO_RT denote the volatility and risk-to-reward timing rules by Kirby and Ostdiek (2012), respectively. At the end of each year, the largest 100 stocks in market capitalization are chosen. The mean return and volatility are estimated using a machine learning (ML) technology that maps data model M into either output. The procedure is repeated annually, resulting in 20×12 monthly returns for model-algorithm-rule. Further, the whole experiment is repeated using the smallest 100 stocks in market capitalization. Panels (a) and (b) report the results for the big and small stocks, respectively. The M column denotes the data used to form portfolios. The baseline data capturing market and fundamentals is $M = 1$. The specifications $M = 2$ and $M = 3$ denote the baseline data plus analyst recommendations and report text, respectively. $M = 4$ denotes the specification in which all data sources are utilized for grouping stocks into deciles. The final specifications, $M = 5$, $M = 6$, and $M = 7$, are added for robustness and correspond to cases where the ML model is trained solely on analysts' textual data, analysts' recommendation data, and both textual and recommendation data, respectively, without including market and fundamental data. The results below report the marginal contribution in terms of certainty equivalent return (CER) of model $M = 2, 3, 4, 5, 6, 7$ relative to the baseline model $M = 1$. The Algorithm column indicates which ML model (out of the seven ones considered in the investigation) is used. Additionally, the table reports statistical performance by computing the relative improvement in out-of-sample R^2 of data $M > 1$ over the baseline one ($M = 1$).

		Statistical Performance		CER for $\gamma = 3$					*CER for* $\gamma = 5$				
Algorithm	M	$R^2(RET)$	$R^2(RV)$	MV_NS	GMV	GMV_NS	KO_VT	KO_RT	MV_NS	GMV	GMV_NS	KO_VT	KO_RT
						Panel (a) 100 Large Cap Stocks							
glmnet	2	0.934	1.011	1.097	0.039	0.041	0.238	3.094	1.132	0.039	0.041	0.093	2.948
glmnet	3	2.056	1.039	0.228	0.001	-0.022	0.524	3.526	0.183	-0.007	-0.026	0.673	2.327
glmnet	4	1.690	1.043	1.642	0.071	0.071	1.002	4.974	1.611	0.063	0.064	1.067	3.528
glmnet	5	-0.237	0.012	1.745	0.232	0.240	5.387	4.901	2.070	0.321	0.329	7.215	6.197
glmnet	6	0.593	0.007	5.365	2.573	2.515	11.405	11.511	5.500	2.481	2.432	10.216	10.189
glmnet	7	1.481	0.034	4.380	2.841	2.765	6.393	8.092	4.686	2.502	2.445	4.031	6.358
svmRadial	2	0.572	1.016	-0.504	-0.057	-0.051	-1.017	0.356	-0.499	-0.064	-0.059	-0.738	0.354
svmRadial	3	0.815	1.009	-0.766	-0.010	-0.006	0.188	3.680	-0.806	-0.008	-0.005	0.182	7.725
svmRadial	4	0.627	1.017	-0.194	-0.037	-0.031	-0.379	3.676	-0.303	-0.039	-0.035	-0.193	7.030
svmRadial	5	1.571	0.121	-2.277	-1.202	-1.173	-9.415	-13.019	-2.063	-1.143	-1.115	-9.827	-9.283
svmRadial	6	-0.046	0.044	-2.114	0.376	0.389	5.152	-1.187	-3.459	0.516	0.528	7.858	3.108
svmRadial	7	2.615	0.177	-3.764	-0.104	-0.085	3.336	2.700	-4.123	0.031	0.048	5.331	8.693
rf	2	0.982	0.999	1.379	-0.019	-0.019	0.131	0.316	1.698	-0.022	-0.022	0.266	0.256
rf	3	1.292	0.999	-1.814	-0.116	-0.124	-0.523	-0.092	-1.920	-0.115	-0.124	-0.442	1.666
rf	4	1.426	1.018	-0.612	-0.032	-0.027	-0.511	-0.703	-0.374	-0.024	-0.020	-0.376	-0.067
rf	5	-0.133	0.055	2.388	-0.770	-0.755	1.086	-3.442	3.306	-0.775	-0.760	1.457	-0.724
rf	6	-0.178	0.019	0.083	0.103	0.116	3.758	-7.309	0.249	0.155	0.168	4.337	-6.087
rf	7	0.582	0.116	0.287	-0.439	-0.422	1.639	-6.942	1.074	-0.421	-0.404	2.323	-4.462
rf2	2	0.467	0.991	-0.201	0.035	0.050	0.564	0.008	-0.307	0.040	0.056	0.658	-0.429
rf2	3	1.010	1.002	-2.432	0.039	0.045	0.308	-1.983	-2.772	0.048	0.055	0.470	-1.210
rf2	4	1.044	0.997	-0.939	0.017	0.032	0.070	1.777	-1.021	0.026	0.041	0.235	1.066
rf2	5	-0.154	0.019	1.153	-0.466	-0.447	1.883	-4.665	0.989	-0.456	-0.436	2.240	-5.999
rf2	6	-0.078	0.019	-3.484	0.235	0.255	3.820	-5.947	-4.516	0.295	0.315	4.304	-7.329

rf2	7	0.133	0.125	0.151	-0.064	-0.042	3.323	-1.892	-0.226	-0.023	-0.001	4.031	-3.031
mlp	2	1.985	1.728	0.905	-2.294	0.289	9.016	5.757	0.990	-1.639	0.405	9.626	4.667
mlp	3	7.403	1.536	-1.250	-1.169	-0.312	2.192	1.887	-1.146	-0.696	-0.250	2.928	1.970
mlp	4	13.416	1.699	-1.438	-1.484	-0.520	1.244	2.590	-2.045	-1.362	-0.522	1.028	3.129
mlp	5	6.499	0.042	2.058	-1.077	-0.083	8.878	14.015	2.697	-0.026	0.037	12.655	19.126
mlp	6	-0.763	0.046	5.483	-0.415	0.580	10.734	14.465	6.023	0.652	0.716	14.573	19.668
mlp	7	-0.113	0.140	4.010	-0.334	0.649	5.793	1.092	3.979	0.593	0.649	5.436	-10.644
nnet	2	-1.730	0.778	2.131	0.399	0.177	0.488	-1.243	2.918	0.704	0.194	0.757	-0.797
nnet	3	-0.706	1.173	1.885	1.161	0.102	1.912	2.004	2.515	1.059	0.093	1.755	2.218
nnet	4	0.401	1.054	-1.575	-1.063	-0.165	-1.362	-1.580	-1.678	-0.989	-0.174	-1.355	-1.255
nnet	5	-0.814	0.148	5.392	-0.532	-0.284	-2.019	1.250	5.772	-0.635	-0.344	-4.443	-0.951
nnet	6	0.593	0.055	5.861	0.342	0.400	2.860	8.308	6.559	0.746	0.440	3.350	10.816
nnet	7	0.608	0.113	0.291	1.261	0.207	1.879	0.230	0.577	1.201	0.239	1.086	-1.603
pcaNNet	2	-0.221	2.470	-0.451	1.611	-0.050	0.317	2.482	-0.650	1.214	-0.059	-0.213	1.644
pcaNNet	3	-0.586	1.540	1.049	0.291	0.029	0.098	0.378	0.862	0.477	0.041	0.236	0.152
pcaNNet	4	-0.035	2.136	3.513	1.021	-0.132	0.265	5.369	4.535	0.758	-0.142	-0.102	5.431
pcaNNet	5	-4.190	0.162	4.399	1.212	-0.453	2.189	8.349	5.625	0.829	-0.448	1.005	9.974
pcaNNet	6	-10.049	0.617	6.164	3.856	0.087	5.236	14.836	7.209	4.303	0.098	5.710	17.921
pcaNNet	7	1.077	0.217	1.602	2.022	0.357	5.484	7.624	1.653	1.664	0.386	4.367	7.313
						Panel (b) 100 Small Cap Stocks						
glmnet	2	0.897	1.019	0.239	-0.022	-0.022	-0.369	0.653	0.620	-0.024	-0.025	-0.431	1.410
glmnet	3	0.609	1.021	-0.372	0.148	0.142	0.688	-0.562	-0.349	0.139	0.133	0.714	-0.596
glmnet	4	0.532	1.037	0.663	0.103	0.103	0.594	0.943	0.938	0.095	0.095	0.652	1.646
glmnet	5	0.095	0.001	4.012	1.193	1.185	5.574	7.640	5.211	1.218	1.209	7.364	10.465
glmnet	6	0.006	0.080	-0.255	-1.638	-1.622	-6.571	0.969	-0.468	-2.504	-2.475	-10.636	-1.762
glmnet	7	-0.093	0.001	-2.680	-1.425	-1.410	-9.359	0.438	-3.372	-2.415	-2.379	-13.526	-2.415
svmRadial	2	1.456	1.033	-1.164	0.168	0.160	1.139	-0.188	-1.068	0.155	0.146	1.024	0.089
svmRadial	3	0.458	1.059	0.047	-0.012	-0.023	0.486	0.359	0.120	-0.001	-0.012	0.629	-0.037
svmRadial	4	0.564	1.095	0.032	0.082	0.079	0.970	-1.354	0.272	0.086	0.082	0.963	-1.706
svmRadial	5	0.193	0.013	1.565	1.522	1.500	10.093	1.524	1.433	1.504	1.481	10.659	-0.192
svmRadial	6	-0.098	0.061	-4.357	1.067	1.049	7.236	-0.997	-4.907	1.102	1.083	8.512	-0.993
svmRadial	7	0.366	0.019	1.473	1.263	1.246	6.892	2.294	1.505	1.287	1.270	6.940	2.922
rf	2	0.965	1.049	-0.348	0.020	0.020	0.097	-1.024	-0.265	0.008	0.007	0.066	-0.681
rf	3	0.381	1.032	-0.574	-0.023	-0.029	0.171	-2.408	-0.522	-0.024	-0.030	0.323	-2.636
rf	4	0.671	1.052	-1.426	-0.014	-0.014	-0.065	-2.344	-1.437	-0.014	-0.015	-0.122	-2.745
rf	5	-0.025	0.051	2.560	0.929	0.924	3.545	1.022	2.755	0.909	0.903	4.388	1.373
rf	6	0.146	0.045	0.057	0.466	0.462	1.414	-1.711	0.202	0.479	0.474	2.246	-2.084
rf	7	-0.059	0.063	0.618	1.007	1.000	3.544	1.307	0.779	1.009	1.002	4.629	1.964
rf2	2	0.858	1.032	0.772	0.035	0.035	0.190	0.779	0.835	0.023	0.023	0.078	0.074
rf2	3	0.741	0.988	0.600	-0.009	-0.007	0.255	0.538	0.612	-0.001	0.001	0.374	0.064
rf2	4	0.728	1.013	0.434	0.019	0.020	0.107	0.538	0.433	0.018	0.020	0.168	0.079
rf2	5	-0.005	0.063	0.359	0.894	0.891	2.972	0.703	-0.456	0.883	0.879	3.640	-0.559
rf2	6	0.060	0.036	-0.421	0.490	0.487	1.394	0.130	-0.856	0.514	0.509	2.188	-0.995
rf2	7	-0.024	0.088	-0.437	0.794	0.792	2.440	1.822	-1.295	0.815	0.812	3.425	1.333
mlp	2	1.146	1.354	3.431	-1.692	0.608	1.256	-0.227	3.362	-1.679	0.759	3.783	0.364

mlp	3	-0.331	1.681	1.820	-3.811	0.122	-3.374	-3.194	1.806	-4.361	0.233	-1.946	-3.514
mlp	4	4.559	1.492	3.808	-2.668	0.620	1.174	1.009	4.334	-2.705	0.670	1.593	2.735
mlp	5	6.112	0.002	6.363	-1.084	1.627	7.123	8.141	7.705	-0.609	1.869	11.656	12.674
mlp	6	0.099	0.197	3.872	-2.993	-0.270	3.001	1.558	5.115	-2.656	-0.164	6.980	6.400
mlp	7	0.564	0.020	7.861	0.151	2.854	10.823	8.218	8.599	0.521	2.994	13.945	10.307
nnet	2	0.993	0.789	2.675	-0.211	0.027	-0.101	2.330	2.515	0.017	0.004	-0.009	1.855
nnet	3	0.945	2.577	-0.097	1.145	0.096	1.448	1.905	-0.805	1.057	0.097	1.423	1.601
nnet	4	0.492	0.602	0.513	-0.745	0.261	-0.265	-1.391	-0.293	-0.654	0.251	-0.367	-3.095
nnet	5	0.025	0.203	6.414	1.005	1.282	4.232	9.673	6.974	1.009	1.328	3.834	10.849
nnet	6	0.331	0.004	4.084	0.279	0.609	0.994	5.433	4.373	0.630	0.632	0.896	7.371
nnet	7	-0.027	0.076	4.339	1.832	0.783	0.910	3.211	4.203	1.763	0.803	0.452	2.940
pcaNNet	2	0.494	11.944	-0.495	1.828	0.083	1.331	3.334	-0.831	1.354	0.074	0.496	2.071
pcaNNet	3	0.494	1.837	-0.139	-0.224	0.145	-0.293	-1.703	-0.044	-0.168	0.122	-0.281	-1.530
pcaNNet	4	0.682	3.588	-4.070	1.374	0.252	1.437	-3.201	-4.013	1.048	0.241	0.820	-3.795
pcaNNet	5	-0.021	0.183	6.019	3.063	1.314	4.077	12.272	7.250	2.561	1.277	2.200	12.968
pcaNNet	6	0.128	1.082	4.532	4.559	0.777	5.956	12.713	5.877	4.859	0.761	6.428	15.411
pcaNNet	7	0.064	0.430	2.075	2.830	1.083	3.187	10.213	2.554	2.361	1.078	2.476	10.889

Appendix Bibliography

Biecek, P., 2018. Dalex: explainers for complex predictive models in r. The Journal of Machine Learning Research 19, 3245–3249.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Brochet, F., Loumioti, M., Serafeim, G., 2015. Speaking of the short-term: Disclosure horizon and managerial myopia. Review of Accounting Studies 20, 1122–1163.

Brown, L. D., Call, A. C., Clement, M. B., Sharp, N. Y., 2015. Inside the "black box" of sell-side financial analysts. Journal of Accounting Research 53, 1–47.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The Review of Financial studies 22, 1915–1953.

Dessaint, O., Foucault, T., Frésard, L., 2024. Does alternative data improve financial forecasting? the horizon effect. The Journal of Finance 79, 2237–2287.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al., 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9, 155–161.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Farboodi, M., Singal, D., Veldkamp, L., Venkateswaran, V., 2024. Valuing financial data. The Review of Financial Studies .

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning, vol. 112. Springer.

Kan, R., Wang, X., Zhou, G., 2022. Optimal portfolio choice with estimation risk: No risk-free asset case. Management Science 68, 2047–2068.

Kirby, C., Ostdiek, B., 2012. It's all in the timing: simple active portfolio strategies that outperform naive diversification. Journal of Financial and Quantitative Analysis 47, 437–467.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Matsumoto, D., Pronk, M., Roelofsen, E., 2011. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. The Accounting Review 86, 1383–1414.

Moreira, A., Muir, T., 2017. Volatility-managed portfolios. The Journal of Finance 72, 1611–1644.

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

Painter, M., 2018. Unlevelling the playing field: The investment value and capital market consequences of alternative data. Available at SSRN 3222741 .

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al., 1985. Learning internal representations by error propagation.