Valuation ratios, surprises, uncertainty or sentiment: How does financial machine learning predict returns from earnings announcements?

Matthias Schnaubelt[a,1,*], Oleg Seifert[a,1]

[a] *University of Erlangen-Nürnberg, School of Business, Economics and Society, Lange Gasse 20, 90403 Nürnberg, Germany*

Abstract

We apply state-of-the-art financial machine learning to assess the return-predictive value of more than 45,000 earnings announcements on a majority of S&P1500 constituents. To represent the diverse information content of earnings announcements, we generate predictor variables based on various sources such as analyst forecasts, earnings press releases and analyst conference call transcripts. We sort announcements into decile portfolios based on the model's abnormal return prediction. In comparison to three benchmark models, we find that random forests yield superior abnormal returns which tend to increase with the forecast horizon for up to 60 days after the announcement. We subject the model's learning and out-of-sample performance to further analysis. First, we find larger abnormal returns for small-cap stocks and a delayed return drift for growth stocks. Second, while revenue and earnings surprises are the main predictors for the contemporary reaction, we find that a larger range of variables, mostly fundamental ratios and forecast errors, is used to predict post-announcement returns. Third, we analyze variable contributions and find the model to recover non-linear patterns of common capital markets effects such as the value premium. Leveraging the model's predictions in a zero-investment trading strategy yields annualized returns of 11.63 percent at a Sharpe ratio of 1.39 after transaction costs.

Keywords: Earnings announcements; Asset pricing; Machine learning; Natural language processing

*Corresponding author

Email addresses: matthias.schnaubelt@fau.de (Matthias Schnaubelt), oleg.seifert@fau.de (Oleg Seifert)

[1]The authors have benefited from many helpful discussions with Ingo Klein, Michael Amberg, Jonas Dovern, Adrian Rott, Christopher Krauss, Thomas Fischer, Alexander Glas, Daniel Perico Ortiz, Benjamin Hübel and Maximilian Glück. We would further like to express our gratitude to Tino Zepeck and Stephanie Kolbe.

1. Introduction

Voluntary earnings announcements, i.e., the preliminary communication of business results in earnings press releases and dedicated analyst conference calls, convey important information to investors. Compared to other information sources such as 10-K and 10-Q filings, "earnings announcements are an important source of new information in the equity market" (Basu et al., 2013, p. 221) as they represent the first public release of accounting and other business-related information to financial markets. The information content of earnings press releases has significantly increased over time (Landsman and Maydew, 2002; Collins et al., 2009; Beaver et al., 2018) and executives direct elevated attention towards earnings announcements (Graham et al., 2005). With the present paper, we study the price-predictive power of the various information components of earnings press releases and conference calls by means of a flexible state-of-the-art machine learning model and a large – both in terms of included variables and observations – sample.

Our study relates to the large strand of literature relating earnings announcements to stock price behavior. Beginning with the seminal works of Ball and Brown (1968) and Beaver (1968), researchers have studied the link between forecast errors in market expectations and asset prices, such as the positive relation between unexpected earnings and stock returns. Ball and Brown (1968) find stock returns to drift in the direction of earnings surprises for up to 60 trading days after the announcement, which has been replicated in several subsequent studies (Foster et al., 1984; Bernard and Thomas, 1989) and for non-US samples (compare, e.g., the survey in Kothari and Wasley, 2019). Yet other authors find evidence for a non-linear dependence of returns on earnings surprises (Freeman and Tse, 1992; Cheng et al., 1992; Kothari, 2001). Similar to earnings surprises, the role of additional forecast errors, such as revenue or cashflow surprises has been examined (Livnat and Zarowin, 1990; Swaminathan and Weintrop, 1991; Ertimur et al., 2003; Jegadeesh and Livnat, 2006a; Livnat and Santicchia, 2006). Related to errors in market expectations, other studies use differences in opinion of institutional analysts as a predictor of abnormal stock returns (Stickel, 1991; Diether et al., 2002; Gleason and Lee, 2003; Zhang, 2006; Hirshleifer et al., 2009). Apart from errors in market expectations, a large strand of asset pricing literature focuses on the relation between fundamentals-based stock characteristics and stock returns.[2] Specific examples are works on the value premium, which state that value stocks, characterized for instance by a high book-to-market ratio, outperform growth stocks (see, for example, Lakonishok et al., 1994; Chan et al., 1995; Fama and French, 1998; Desai et al., 2004).

[2]Surveys of further related works can be found for example in Kothari (2001), Jacobs (2015) and Kothari and Wasley (2019).

Our study is likewise related to the accelerating use of advanced textual analysis techniques and machine learning in financial and accounting research. For example, recent corporate disclosure research in the context of earnings announcements has applied sentiment analysis to earnings press releases (Li, 2008; Henry, 2008; Demers and Vega, 2008; Sadique et al., 2008; Davis et al., 2012) or earnings conference call transcripts (Price et al., 2012; Davis et al., 2015; Brockman et al., 2015) to study the relation of spoken content and stock return behavior.[3] The use of machine learning in finance is commonly motivated with its ability to choose the best from a large number of possible predictors and its capability to model non-linearities and interactions present in financial data. An early and directly related example is the work of Henry (2006), who applies classification and regression trees to analyze the reaction of contemporary returns to quantitative accounting information and text-based variables, and finds additional value in the verbal components of a small sample of earnings press releases. More recent empirical studies document a superior return-predictive performance of machine learning over linear models, for example, in event studies of financial news (Ke et al., 2019; Schnaubelt et al., 2020), in applied momentum strategies (Krauss et al., 2017; Fischer and Krauss, 2018), or in evaluating large numbers of potential factors for asset pricing (Gu et al., 2018; Sak et al., 2018).

To our knowledge, this study presents the first large-scale empirical study of financial machine learning in the context of earnings announcements, thereby covering the wealth of complex information in analyst expectations, earnings press releases and earnings conference calls from more than 45,000 announcements on a majority of S&P1500 constituents. Specifically, we make the following contributions to the literature:

- First, we introduce a machine learning framework that predicts contemporary and post-announcement abnormal returns from earnings announcement data. Based on the literature, we group features into four categories to obtain a holistic representation of the diverse information components of both the earnings press release and the conference call transcript, i.e., valuations ratios, forecast errors, uncertainty and information quality features as well as textual sentiment polarity features. Using these features as input, we apply random forests and several benchmark models to predict abnormal returns for different forecast horizons. To study events with the largest expected abnormal returns in further detail, we extend conventional portfolio sort methodology to sort events into decile portfolios based on the rolling rank of the model's return prediction.

[3] Compare Kearney and Liu (2014) and Loughran and McDonald (2016) for detailed surveys of textual analysis in finance and accounting.

- Second, we assess the model's out-of-sample performance and learning. Compared to our benchmarks, the random forest's predictions lead to the highest abnormal returns. In resemblance to the post-earnings-announcement drift, we find that returns tend to increase with the forecast horizon. Specifically, mean abnormal return in the top or flop decile increase from 95.4 bp for a forecast horizon of 5 trading days to 193.6 bp for 60 days. We perform further analyses to assess the economic rationale behind the predictions of the random forest. (1) We analyze the influence of size and value-growth characterization on the model's performance. We find larger abnormal returns for firms with smaller market capitalization. Compared to value stocks, growth stocks exhibit a delayed return drift. (2) While forecast errors pertaining to earnings and revenue are the main predictors for contemporary returns, we find that a larger number of variables, mostly valuation ratios and forecast errors, contributes to post-announcement predictions. (3) We leverage accumulated local effects plots as a state-of-the-art model inspection technique to inspect the non-linear influence of features on the model's prediction. We find that the model recovers well-known capital markets patterns, such as earnings surprise and size effects as well as the value premium, for its predictions. Overall, we interpret our results in terms of the gradual diffusion of information at different speeds, with a slower diffusion of value-relevant information in a large number of complex and costly predictors.

- Third, we introduce a straightforward event-based trading strategy based on earnings announcements to evaluate the economic significance of the model's predictions. The strategy's financial performance demonstrates that an investor could have earned statistically and economically significant daily returns over the period from 2013 to 2019. Specifically, we find annualized returns of 11.63 percent at a Sharpe ratio of 1.39 after conservative transaction costs.

The remainder of this paper is organized as follows: In Section 2, we describe our data sources and the data preparation process. Section 3 details all building blocks of our methodology, i.e., the generation of features and targets, the training of predictive models, our portfolio sort methodology and the announcement-based trading strategy. Our results are presented in Section 4, and we conclude in Section 5.

2. Data

We retrieve our data set on S&P1500 earnings announcements from two sources, i.e., Thomson Reuters and Seeking Alpha. In the following, we describe the retrieval of I/B/E/S analyst estimates

and corresponding actual values (Section 2.1), price data (Section 2.2) as well as earnings conference call transcript data (Section 2.3). Section 2.4 provides details on the subsequent data cleaning and consolidation steps. Finally, Section 2.5 presents summary statistics.

2.1. I/B/E/S data

We obtain analyst estimate data and corresponding actual values from the Institutional Brokers Estimate System (I/B/E/S) database of Thomson Reuters. Reported actual earnings and other items[4] are collected by Thomson Reuters and come with the timestamp of the value's first recording. Reported actual values are adjusted before entering the I/B/E/S database to match the majority accounting basis, i.e., the accounting basis used by the majority of analysts (Thomson Reuters, 2018). Apart from the reported actual values, we retrieve time series for the mean and standard deviation of analyst forecasts for a given firm-quarter. In these time series, every update in the contributing forecasts leads to a new entry that states the updated mean and standard deviation values as well as the timestamp of the update. The Thomson Reuters data collection process requires estimate contributors to confirm the up-to-dateness of their estimates on a regular basis, and excludes any estimate from the mean forecast that has not been confirmed for 180 days (Thomson Reuters, 2018). All I/B/E/S data are adjusted for corporate actions, such as stock splits, in the same way as price data.

2.2. Price data

We obtain price data from the Thomson Reuters Eikon database for all stocks that have ever been a constituent of the S&P1500 index during our sample period. The data contain daily open and close prices and are adjusted for stock splits and other corporate actions. For return computation, we additionally adjust price data for dividend payments following standard methodology (Woolridge, 1983; Center for Research in Security Prices, 2018). Specifically, we multiply all prices before the respective ex-dividend date with an adjustment factor, which is one minus the stock dividend divided by the close price one day before the ex-dividend date.

2.3. Earnings conference call transcripts

Finally, we complement our data set with earnings call transcript data from Thomson Reuters Eikon and the financial news provider Seeking Alpha[5]. The Thomson Reuters database does not offer a bulk download functionality for earnings conference call transcripts, which is a clear

[4] In addition to earnings-per-share, we retrieve revenue, book-value-per-share, cash-flow-per-share and dividend-per-share values. Section 3.3 provides details about all retrieved variables.

[5] seekingalpha.com

bottleneck for our study as we aim to create an extensive data set. We validate transcript data from Seeking Alpha for a random sample of calls with manually downloaded transcripts from Thomson Reuters Eikon. For our random sample, we see that both structure and content retrieved from these two different sources show no relevant deviations, and we hence proceed with Seeking Alpha to create a larger data base. We retrieve roughly 150,000 transcripts of earnings conference calls from Seeking Alpha and convert relevant content of the transcripts from HTML files to plain text. We complement our data set with an additional 1,900 calls from Thomson Reuters Eikon that have not been available from Seeking Alpha. For each transcript, we extract meta data, information on the participants of the conference call, and the spoken content. Meta data includes the company ticker, the respective fiscal quarter as well as the date and time of the call. Participating persons on the part of the company are recorded with their name and position, and participating analysts with their name and affiliation. We separate the spoken content of the call into the prepared remarks section and the subsequent Q&A session, if available. Further, we subdivide the content of the Q&A session into the individual statements of each participant. We preprocess the textual content by splitting sections into single sentences and creating word tokens after removing stop words, i.e., common words with little or no information content. Word tokens are given by the lemmata (base forms) of the respective terms.

2.4. Data preparation and consolidation

We merge data from our two sources, Seeking Alpha and Thomson Reuters, by matching ticker symbols and company names with the Reuters ID Code (RIC). We then apply thorough data cleaning and sanity checks: (1) We validate the date and fiscal period of the earnings call with data from Thomson Reuters Street Events, a database of stock-related events. Observations where information could not be precisely matched with Street Event data are discarded. (2) We keep observations only if the earnings report date from I/B/E/S is at or before the date of the earnings call to ensure that all information has been available at the time of call, as we use the call's date as our event date. Observations where the report date is more than one week before the call are discarded as well. (3) We remove observations with missing call transcripts, earnings-per-share or revenue estimates as well as respective actual values from I/B/E/S. This ensures that every observation is based on an earnings forecast from at least one analyst. (4) We only keep observations with sufficient price data to compute all relevant returns. (5) We purge any duplicate firm-quarter observations. We restrict our sample to constituents of the S&P1500 index. First, this ensures that our sample is based on survivor-bias-free data on common stocks of U.S. companies, and excludes American depositary receipts, closed-end funds and other share types as detailed in the index's eligibility criteria (S&P Dow Jones Indices, 2019). Second, this focuses our study on

	Year of the earnings call													
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	**Total**
Panel A: Total number of calls														
S&P500	662	672	1369	1188	1491	1646	1759	1810	1834	1835	1867	1851	(444)	18432
S&P400	187	434	732	504	555	837	1327	1374	1423	1444	1444	1455	(352)	12068
S&P600	144	488	741	571	336	587	1655	1836	1991	2050	2075	2080	(506)	15060
S&P1500	993	1594	2842	2263	2382	3070	4741	5024	5248	5329	5386	5386	(1302)	45560
Panel B: Percentage coverage of index segment														
S&P500	33.1	33.6	68.5	59.4	74.6	82.3	87.9	90.7	91.7	91.8	93.3	92.5	(22.2)	
S&P400	11.7	27.1	45.8	31.5	34.7	52.3	82.9	85.9	88.9	90.2	90.2	90.9	(22.0)	
S&P600	6.0	20.3	30.9	23.8	14.0	24.5	69.0	76.5	83.0	85.4	86.5	86.7	(21.1)	
S&P1500	16.6	26.6	47.4	37.7	39.7	51.2	79.0	83.7	87.5	88.8	89.8	89.8	(21.7)	

Table 1: **Sample size.** This table reports earnings event counts for the final sample, i.e., after all data cleaning and validation procedures have been carried out. Panel A states the total number of events conditional on the market index segment and year of the earnings call. S&P500, S&P400, and S&P600 refer to the large-, mid- and small-cap segments of the S&P1500 index, respectively. Panel B states the corresponding percentage coverage of earnings calls. These frequencies are relative to the upper bound of expected calls given by four times the nominal index constituent count. Numbers in parentheses refer to earnings calls that occurred in 2019, which are only partially covered by our sample.

the most liquid market segment of the 1500 largest companies listed on U.S. exchanges, which cover about 90% of U.S. market capitalization. Third, concentrating on data from the S&P1500 index ensures exhaustive call transcript and analyst forecast coverage. To determine whether a given earnings announcement should be included in our final data set, we first obtain month-end constituent lists of the S&P500 (large-cap), S&P400 (mid-cap) and S&P600 (small-cap) indices from Thomson Reuters. In a second step, we determine whether an earnings event has been a constituent of one of these indices at the date of the earnings call, and remove all other observations.

2.5. Summary statistics

Panel A of Table 1 provides details on the size of our sample. The first and last earnings events are on January 16, 2007 and March 26, 2019, respectively. In total, our final sample contains 45,560 earnings events on a majority of S&P500 constituents. Panel B of the same table relates these numbers to the expected event count assuming that every constituent of the S&P1500 index conducts four conference calls per year, which is a lower bound of the actual event coverage, as a minority of firms do not conduct conference calls.[6] We observe that event coverage is generally higher for market segments with high market capitalization, i.e., highest for the S&P500 large-cap index and lowest for the S&P600 small-cap index. Average event coverage for the S&P1500 index is close to 90 percent for years 2013 to 2018. Table 2 presents statistics on the time difference between

[6]Because the SEC allows a choice of channel for the dissemination of public disclosures, it can be assumed that some companies do not hold earnings conference calls (www.sec.gov/rules/final/33-7881.htm#P169_65201). For their 2004-2007 sample, Price et al. (2015) find that 78 percent of the examined real estate investment trusts held conference calls.

Number of days after earnings press release	Number of earnings calls			
	S&P500	S&P400	S&P600	S&P1500
0	15278	8016	10494	33788
1	3075	3987	4496	11558
2	42	37	37	116
3	11	13	22	46
4	4	4	3	11
5	1	4	2	7
6	10	5	4	19
7	11	2	2	15

Table 2: **Relative timing of earnings press release and conference call.** This table reports count statistics conditional on the relative timing between the earnings press release and earnings call, expressed as the difference in days between the date of the earnings press release and the respective conference call. Data are shown for the final sample, which means that calls occurring before the earnings press release have already been removed.

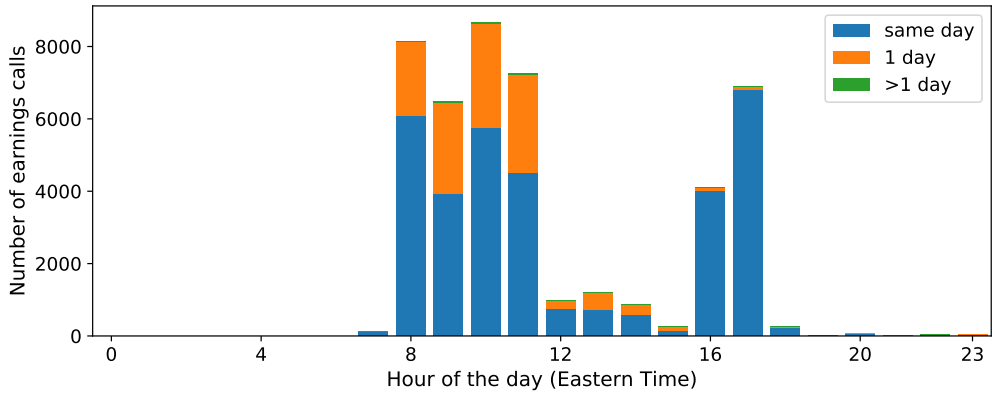

Figure 1: **Distribution of call start times over the day.** This figure depicts the number of earnings calls by starting hour of the day in Eastern Time in our final sample. Call counts are further subdivided based on the number of days the earnings press release occurs prior to the earnings call, which is shown by the colored bar segments.

earnings press releases and conference calls. For the full sample, three quarters of calls are held at the day of the earnings press release, and about one quarter of calls take place one day after the press release. Delays exceeding one day are very rare. With 82.89 percent same-day calls, firms of the S&P500 stage calls considerably faster than smaller firms (on average, 68.23 percent are same-day calls for the S&P400 and S&P600). Figure 1 depicts the distribution of events conditional on the starting hour of the earnings call in Eastern Time. Most calls take place either around market open or market close of the New York Stock Exchange.[7] Most calls in the afternoon are held on the day of the earnings press release. By contrast, a larger number of morning calls occurs one day after the respective earnings press release.

[7]Please note that a few companies such as Walmart or Comcast hold earnings calls at 7:00 am or 7:30 am. We checked these cases on the websites of the individual companies (compare, e.g, https://corporate.walmart.com/newsroom/events/ and https://www.cmcsa.com/events-and-presentations/).

3. Methodology

Our methodology consists of five steps. First, we split our data into consecutive training and testing periods (Section 3.1). Second, we calculate buy-and-hold abnormal returns as a prediction target (Section 3.2) and several features (i.e., input variables, Section 3.3). Third, we train different predictive models (Section 3.4). Fourth, we sort earnings events into decile portfolios based on the rolling rank of predicted abnormal returns (Section 3.5). Fifth, we introduce a straightforward announcement-based trading strategy to assess the economic significance of our model's predictions (Section 3.6).

3.1. Study design

In line with several similar studies at the intersection of finance and machine learning (see, for example, Pesaran and Timmermann, 1995; Gu et al., 2018), we use a growing-window forward-validation scheme (Tashman, 2000; Schnaubelt, 2019) to split data into a series training and out-of-sample validation sets. Specifically, we employ a period of one year as out-of-sample validation and trading period, which is rolled forward by one year for every split. The first test set comprises all earnings events with a call date in 2013 and the last all events from 2019, which yields seven data splits in total. To keep the temporal order of observations, we only use data preceding a split's out-of-sample validation window for training. Hence, the first model run uses training data from years 2007 to 2012, and the last model is trained on all data from 2007 to 2018.

3.2. Abnormal return calculation and target generation

Next, we calculate buy-and-hold abnormal returns to isolate the stock's unexpected price reaction. To calculate returns, we assume that an investor opens a position at market open on the day following the earnings event (defined as the day of the earnings call), and holds the position for T trading days. We define the raw return following an earnings event (at trading day $t = 0$) covering stock s for fiscal quarter q as

$$R_{s,q}^{T} = \frac{P_{s,T}^{C}}{P_{s,1}^{O}} - 1, \tag{1}$$

where $P_{s,1}^{O}$ denotes the open price of stock s on the day following the call, and $P_{s,T}^{C}$ is the close price T trading days after the call. Both prices are adjusted for dividends and other distributions as well as stock splits and other corporate actions. In line with several prior studies (see, for example, Balakrishnan et al., 2010; Battalio and Mendenhall, 2011), we measure abnormal returns for a period of T trading days as buy-and-hold abnormal returns given by

$$Y_{s,q}^{T} = R_{s,q}^{T} - ER_{s,q}^{T}, \tag{2}$$

where $ER^T_{s,q}$ denotes the expected return over the same period. We follow the approach of Battalio and Mendenhall (2011) and compound intraday and interday return components into expected returns:

$$ER^T_{s,q} = \left(ER^{intra}_{s,q} + 1\right) \cdot \left(ER^{T,inter}_{s,q} + 1\right) - 1. \tag{3}$$

The expected intraday return $ER^{intra}_{s,p}$ is measured from market open to market close at day $t = 1$, and is approximated by the return of an S&P1500 ETF.[8] Interday returns $ER^{T,inter}_{s,q}$ are given by the returns of ten size-matched, equally-weighted decile portfolios (Foster et al., 1984; Bernard and Thomas, 1989, 1990; Price et al., 2012).[9] For most results, we study one contemporaneous return period with a forecast horizon of $T = -5$ trading days, and three post-event drift windows with forecast horizons of $T = 5$, $T = 20$, and $T = 60$ trading days, as research suggests that the post-earnings-announcement drift persists for up to 60 days following the earnings announcement (Foster et al., 1984; Campbell et al., 2009; Price et al., 2012). To train our models, we use abnormal returns of the respective forecast horizon, i.e., $Y^T_{s,q}$, as prediction target.

3.3. Feature generation

To represent the information content of earnings announcements, we generate a feature vector $X_{s,q}$ with a total of 54 elements for every earnings event. Building on existing literature, we group potential drivers of earnings-announcement-related returns into the following four categories to obtain a holistic view:

1. *Valuation ratios (VR):* The first category comprises common stock valuation ratios. Based on the literature, we identify six fundamental measures with potential value relevance, i.e., book value, size, dividends, earnings, sales and cashflows. Prior research has identified that stock-by-stock differences in these measures relative to stock prices possess return-predictive power (see, for example, Fama and French, 1992; Basu, 1977, 1983; Barbee et al., 1996; Porta, 1996; Keim, 1985). In all features, we use the most recent fundamental values, as reported in I/B/E/S data on the fiscal quarter covered in the earnings event. We choose to compute ratios from individual items (as opposed to using pre-computed time-averaged trailing ratios, e.g., from Thomson Reuters) as we are interested in the short-term price reaction related to the newly disseminated information from the earnings event.

2. *Forecast errors (FE):* Variables in the second category reflect unexpected deviations between prior market expectation and actually reported accounting information. Ball and Brown

[8] Specifically, we use the iShares Core S&P Total U.S. Stock Market ETF (BlackRock Inc., 2020).

[9] We thank Kenneth R. French for providing daily size-adjusted portfolio returns on his website. Size deciles are defined by New York Stock Exchange market capitalization deciles as given on the website.

(1968) empirically find that unexpected corporate earnings lead to subsequent abnormal stock price movements. Building upon these findings, various researches have examined the relationship between surprises regarding other accounting information, such as cashflows (Livnat and Zarowin, 1990) or sales (Jegadeesh and Livnat, 2006b), and abnormal stock price performance. We compute forecast errors as the difference between actual values and matched mean analyst estimates, both from I/B/E/S. In line with Fried and Givoly (1982), we expect analyst estimates to be better surrogates of market expectations than predictions from time-series models, as analyst forecasts may also incorporate the broad range of additional information emitted between two consecutive earnings announcement events. Supporting empirical evidence finds that the return drift following earnings announcements is stronger and persists longer when using analyst estimates instead of predictions from time-series models (Brown and Rozeff, 1978; Fried and Givoly, 1982; Doyle et al., 2006; Livnat and Mendenhall, 2006). Following previous studies (Hirshleifer et al., 2009; Dellavigna and Pollet, 2009; Battalio and Mendenhall, 2011), we scale forecast errors by price to avoid issues arising from dividing by zero earnings values.[10] In this form, forecast errors can be interpreted as a correction to the associated valuation ratio, which is also reflected in the naming convention of our variables.

3. *Uncertainty and information quality (UIQ)*: The third category is based on literature pertaining to the relation of information uncertainty and information quality of corporate disclosures on abnormal post-earnings-announcement returns. Givoly and Lakonishok (1979) and Hawkins et al. (1984) study the effect of analyst forecast revisions on stock prices by examining the number and changes of revisions in certain time periods. In line with the behavioral argument of Hirshleifer (2001), Zhang (2006) considers the dispersion of I/B/E/S analyst earnings forecasts as a proxy for information uncertainty and finds empirical evidence that information uncertainty may intensify stock price reactions. Li (2008) links the quality of corporate disclosures to firm performance using the length of annual reports as a proxy for their readability. Price et al. (2012) apply a similar measure in the context of earnings conference calls. Based on these studies, we calculate several proxies of information quality and uncertainty related to earnings, revenues and dividends.

4. *Textual sentiment polarity (POL):* The fourth category comprises variables capturing the sentiment of the earnings conference call, and builds upon the vast literature studying the relation between abnormal stock returns and the sentiment of corporate disclosures (Loughran and McDonald, 2011), earnings releases (Henry, 2008) and conference calls (Matsumoto et al.,

[10]The price data used for scaling are adjusted in the same way as I/B/E/S data (see, e.g., Hirshleifer et al., 2009).

2011; Price et al., 2012). To calculate sentiment scores, we use positivity and negativity word lists from the finance-specific sentiment dictionary of Loughran and McDonald (2011), which have specifically been created to study the effect of SEC filings and earnings call transcripts on stock returns.[11] We separate positivity and negativity scores for different sections of the transcript to consider different aspects of the call. First, following Price et al. (2012), we calculate separate sentiment scores for the prepared remarks section, which largely restates the earnings press release, and the Q&A session. Second, inspired by Mayew (2008) and Cen et al. (2020), we consider separate sentiment scores for the first and second halves of analyst statements in the Q&A section of the earnings call. Third, we compute topic-specific sentiment scores. We determine topic weights for all sentences of the earnings call and then aggregate sentence-level sentiment scores according to their respective topic weights into topic-specific sentiment scores. Details on the calculation of topic-specific sentiment variables are given in Appendix A.

Table 3 lists the definitions of all features along with relevant references. Descriptive statistics of variables are given in Table C.10 in Appendix C.

3.4. Model training

We model abnormal returns $Y_{s,q}^T$ with an additive error model. Given a vector of predictors $x_{s,q}$, the prediction of the model, i.e., the expected abnormal return, $\hat{y}_{s,q}^T$, is denoted by

$$\hat{y}_{s,q}^T = \mathbb{E}\left[Y_{s,q}^T | X_{s,q} = x_{s,q}\right] = f(x_{s,q}; \hat{\theta}), \tag{4}$$

where $f(x_{s,q}; \hat{\theta})$ is the regression model and $\hat{\theta}$ denotes all parameters learned during model training. In the following, n denotes the number of observations in the training sample used to estimate $\hat{\theta}$, and d is the dimensionality of the feature space.

3.4.1. Random forest model

We employ random forests as our main model. Introduced by Breiman (2001), random forests are a very popular tree-based machine learning model. They use an ensemble of decorrelated binary decision trees on bootstrapped samples of the training data, and aggregate predictions from this ensemble. Random forests have several favorable properties: (1) As other tree-based methods, random forests are able to model variable interactions and non-linear relationships between variables

[11] Specifically, we use the L&M-MasterDictionary as updated in 2018. We want to thank Tim Loughran and Bill McDonald for providing these resources on the web (`https://sraf.nd.edu`).

Variable	Description	Reference(s)
Panel A: Valuation ratios (VR)		
$EP_{s,q}$	*Earnings-to-price ratio*: Earnings per share from the earnings event for quarter q, divided by price	Basu (1977, 1983); Bernard et al. (1997); Fama and French (1992)
$SP_{s,q}$	*Sales-to-price ratio*: Net revenue per share from the earnings event for quarter q, divided by price	Barbee et al. (1996, 2008)
$CP_{s,q}$	*Cashflow-to-price ratio*: Most recent cashflow per share (operating cashflow before investing and financing) divided by price[12]	Wilson (1986); Bernard and Stober (1989); Chan et al. (1991)
$DY_{s,q}$	*Dividend yield*: Dividend per share from the earnings event, divided by price	Litzenberger and Ramaswamy (1979); Lewellen (2004)
$DR_{s,q}$	*Dividend payout ratio*: Dividend per share divided by earnings per share, each from the earnings event	Kallapur (1994)
$BM_{s,q}$	*Book-to-market ratio:* Most recent book value per share divided by price	Stattman (1980); Rosenberg et al. (1985); Fama and French (1992)
$MV_{s,q}$	*Market value:* Common logarithm of the firm's market capitalization at the day of the earnings event	Banz (1981); Fama and French (1992)
Panel B: Forecast errors (FE)		
$EP\text{-}FE_{s,q}$	*Earnings surprise*: Earnings-per-share actual value from the earnings event covering quarter q less respective mean analyst forecast, divided by price	Ball and Brown (1968); Dellavigna and Pollet (2009); Battalio and Mendenhall (2011)
$SP\text{-}FE_{s,q}$	*Sales surprise*: Net revenue-per-share actual value from the earnings announcement less respective mean analyst forecast, scaled by price	Swaminathan and Weintrop (1991); Ertimur et al. (2003)
$CP\text{-}FE_{s,q}$	*Cashflow surprise*: Cashflow-per-share (operating cashflow before investing and financing) actual value less corresponding mean analyst forecast, divided by price[13]	Bernard and Stober (1989); Melendrez et al. (2008)
$DY\text{-}FE_{s,q}$	*Dividend surprise*: Actual value of quarterly dividend less corresponding mean analyst forecast, scaled by price	Pettit (1972); Aharony and Swary (1980)
$BM\text{-}FE_{s,q}$	*Book value surprise*: Book value per share (total assets minus liabilities, preferred stock and intangible assets) actual value less respective mean analyst forecast, scaled by price[13]	
Panel C: Uncertainty/Information quality (UIQ)		
$N\text{-}EPS_{s,q}$	*Number of earnings forecasts*: Number of analyst forecasts contributing to the mean earnings-per-share estimate	Givoly and Lakonishok (1979); Hawkins et al. (1984); Hong et al. (2000); Hirshleifer et al. (2009)
$N\text{-}REV_{s,q}$	*Number of revenue forecasts*: Number of analyst forecasts contributing to the mean revenue estimate	
$C\text{-}EPS_{s,q}$	*Number of earnings forecast changes*: Number of changes to the mean earnings-per-share forecast prior to the earnings event	Gleason and Lee (2003)

[12]Note that this substitution does not occur for market value, earnings-per-share, revenue and dividend data, as data coverage is sufficiently high in our sample.

[13]In some cases pertaining to book value and cashflows, I/B/E/S does not contain analyst estimates or matched actual values. In these cases, we replace forecast errors with a default value of zero.

Variable	Description	Reference(s)
$V\text{-}EPS_{s,q}$	*Variance of the consensus earnings forecast*: Standard deviation of the time-series of mean analyst estimates prior to the earnings event, divided by price	Gleason and Lee (2003)
$D\text{-}EPS_{s,q}$	*Earnings forecast dispersion*: Standard deviation of analyst's earnings-per-share forecasts, scaled by price	Diether et al. (2002); Demers and Vega (2010); Zhang (2006)
$D\text{-}REV_{s,q}$	*Revenue forecast dispersion*: Standard deviation of analyst's net-revenue-per-share forecasts, scaled by price	
$D\text{-}DIV_{s,q}$	*Dividend forecast dispersion*: Standard deviation of analyst's dividend forecasts, divided by price	
$N\text{-}ANA_{s,q}$	*Number of analysts in call*: Number of named analysts present in the conference call as listed in the call's transcript	Li (2008); Price et al. (2012)
$I\text{-}LEN_{s,q}$	*Prepared remarks length*: Common logarithm of the number of words in the prepared remarks section of the earnings call	Li (2008); Price et al. (2012)
$Q\text{-}LEN_{s,q}$	*Q&A length*: Common logarithm of the number of words in the Q&A section of the earnings call	Price et al. (2012); Ji and Rozenbaum (2018)
Panel D: Textual sentiment polarity (POL)		
$I\text{-}P/N_{s,q}$	*Sentiment of the prepared remarks*: Positivity/Negativity scores of the prepared remarks section	Price et al. (2012); Brockman et al. (2015)
$Q\text{-}P/N_{s,q}$	*Sentiment of the Q&A session*: Positivity/Negativity scores of the Q&A section	Price et al. (2012); Brockman et al. (2015)
$FA\text{-}P/N_{s,q}$	*Sentiment of analysts in the first or second half of the Q&A session*:	
$SA\text{-}P/N_{s,q}$	Positivity/Negativity scores of analyst statements in the first and second half of the Q&A sections	
$I\text{-}EA\text{-}P/N_{s,q}$	*Earnings sentiment*: Topic-specific positivity/negativity scores for	Henry (2006); Tetlock et al. (2008)
$Q\text{-}EA\text{-}P/N_{s,q}$	the "earnings" topic, for the prepared remarks and Q&A section	
$I\text{-}RE\text{-}P/N_{s,q}$	*Revenue sentiment*: Topic-specific positivity/negativity scores for the	Engelberg (2008)
$Q\text{-}RE\text{-}P/N_{s,q}$	"revenue" topic, for the prepared remarks and Q&A section	
$I\text{-}LI\text{-}P/N_{s,q}$	*Liquidity sentiment*: Topic-specific positivity/negativity scores for the	Henry (2006)
$Q\text{-}LI\text{-}P/N_{s,q}$	"liquidity" topic, for the prepared remarks and Q&A section	
$I\text{-}EN\text{-}P/N_{s,q}$	*Environment sentiment*: Topic-specific positivity/negativity scores	Henry (2006)
$Q\text{-}EN\text{-}P/N_{s,q}$	for the "environment" topic, i.e., from statements related to the business environment, for the prepared remarks and Q&A section	
$I\text{-}OU\text{-}P/N_{s,q}$	*Outlook sentiment*: Topic-specific positivity/negativity scores for the	Henry (2006); Engelberg (2008);
$Q\text{-}OU\text{-}P/N_{s,q}$	"outlook" topic, i.e., for forward-looking statements, for the prepared remarks and Q&A section	Henry and Leone (2015)
$I\text{-}D\text{-}P/N_{s,q}$	*Changes sentiment*: Topic-specific positivity/negativity scores for the	Henry (2008)
$Q\text{-}D\text{-}P/N_{s,q}$	"changes" topic, i.e., statements expressing changes or differences, for the prepared remarks and Q&A section	

Table 3: **Description of features.** This table provides details on the calculation of variables used as inputs of the predictive model. The last column lists the first authors to report an effect of the respective variable on abnormal returns, as well as those references closest to our specific implementation. Revision counts and mean estimates are computed from all analyst forecasts updated no more than 180 days prior to the respective earnings event. When we scale values by price, we use the stock's closing price five trading days prior to the day of the earnings call (Dellavigna and Pollet, 2009), adjusted in the same way as the I/B/E/S data. We calculate positivity and negativity scores as the word count of positive and negative terms, respectively, divided by the number of words.

and the target. (2) They cope well with high-dimensional feature spaces, are not affected by multicollinearity and select the most influential features from a number of candidates. (3) Random forests are not prone to overfitting (Breiman, 2001) and are fairly robust to noise when compared to boosting techniques (Khoshgoftaar et al., 2011). (4) For the specific task addressed in our paper, i.e., stock return prediction, they have empirically been found to have very good predictive performance, and perform similar or better than neural networks (Krauss et al., 2017; Gu et al., 2018; Schnaubelt et al., 2020). (5) Compared to neural networks and popular deep learning architectures, random forests have far less hyperparameters. Consequently, they require very little or no tuning.

Next, we provide a short description of random forests which loosely follows Hastie et al. (2009) and the implementation in Pedregosa et al. (2011). A random forest is trained by fitting B binary decision trees to bootstrapped samples of size n, which are drawn with replacement from the initial training data. Every tree b is grown by considering the information gain of m random feature candidates and all possible split points. Typical choices for regression tasks are to compute the information gain from the mean squared error, and to set $m = d$ (Pedregosa et al., 2011). The tree is grown until all leaf nodes are pure, or until some stopping criterion such as a maximum tree depth J is met. The prediction of the trained random forest is obtained from the ensemble of trees by averaging single-tree estimates, i.e.,

$$f(x; \hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x), \tag{5}$$

where $\hat{f}_b(x)$ denotes the prediction of tree b, given by the average target value of all observations in the leaf node reached when following the tree for feature vector x. To average over a large number of bootstrapped training samples, we use $B = 5000$ trees. Following Krauss et al. (2017), we set the maximum tree depth to $J = 20$ to allow for a larger degree of feature interaction.

3.4.2. Earnings surprise model

As first benchmark model, we follow the literature on the post-earnings-announcement-drift and consider a naïve model based on earnings surprises. Specifically, we construct a univariate model from earnings surprises, i.e.,

$$f(x; \beta) = \beta \cdot x^{EP\text{-}FE}, \tag{6}$$

where $x^{EP\text{-}FE}$ is the earnings-to-price forecast error, and β is a positive constant estimated from the training data using ordinary least squares (Hastie et al., 2009). The model replicates a univariate portfolio sort for which the specific choice of the positive constant β is irrelevant, as stocks are sorted based on the rank of the model's prediction (compare Section 3.5).

3.4.3. Linear regression model

As second benchmark model, we employ the standard linear model of form

$$f(x;\beta) = \beta_0 + \sum_{i=1}^{d} \beta_i \cdot x_i, \tag{7}$$

where β_0 denotes the intercept and β is the coefficient vector. We estimate the model using ordinary least squares (Hastie et al., 2009). To mitigate possible adverse effects of outliers, we winsorize feature values at the one and 99 percent quantiles of the training data.[14]

3.4.4. Binned linear regression model

To study the effect of non-linear responses to feature values, we loosely follow Hirshleifer et al. (2009) and substitute every feature with decile indicator variables, i.e.,

$$f(x;\beta) = \beta_0 + \sum_{i=1}^{d} \sum_{k=1}^{9} \tilde{\beta}_{i,k} \cdot c_{i,k}(x_i), \tag{8}$$

where β_0 denotes the intercept and $\tilde{\beta}$ is now a matrix of coefficients. Decile bin edges are determined from training data only and are applied also to out-of-sample data for prediction. The $c_{i,k} : \mathbb{R} \mapsto \{0, 1\}$ denote one-hot encoding functions that take a value of one if the i-th element of the feature vector x_i falls into the k-th decile, and zero otherwise. As before, we estimate the model using ordinary least squares (Hastie et al., 2009).

3.5. Portfolio sorting

In the spirit of conventional portfolio sort methodology, we sort earnings events into decile portfolios based on the model's abnormal return prediction, $\hat{y}^T_{s,q}$. For each event, we compute a rolling percentage rank $r^T_{s,q}$ by comparing the event's prediction $\hat{y}^T_{s,q}$ to the predictions of the $W = 1500$ preceding events.[15] This approach has a number of advantages: (1) It is look-ahead free, i.e., the portfolio sort is performed with information available at the time of the sorted earnings event. (2) As we are sorting based on a rolling window of previous earnings events, we do not have to restrict our sample to firms for which the fiscal year equals the calendar year or grouping of firms into quarter-cohorts. (3) As we relate every earnings event to the most recent preceding

[14] See Price et al. (2012) and Kausar (2017) as examples of using winsorization in the earnings-surprise literature.

[15] The window size W is chosen such that it corresponds to roughly one quarter of the yearly number of earnings announcements covering S&P1500 constituents. The use of the previous quarter's distribution of earnings surprises is common in portfolio sorts employed by studies on the post-earnings-announcement drift (see, e.g., Bernard and Thomas, 1989; Livnat and Mendenhall, 2006). To initialize the rolling window of events, we retain the W last events from the training data and apply the model to obtain W out-of-sample predictions.

observations, the portfolio sort is able to quickly adjust to potential shifts in the cross-section of firms, caused for example by altered macroeconomic conditions. Based on the rolling percentage rank $r^T_{s,q}$, we then assign earnings events to one of ten decile portfolios to study the extremes of the distribution of expected abnormal returns. To assess out-of-sample model performance, we primarily use two metrics: First, we calculate return statistics for all earnings events. Specifically, we assign earnings events to either the long portfolio ($r^T_{s,q} \geq 0.5$), or the short portfolio ($r^T_{s,q} < 0.5$). Before computing the mean and other statistics of the return distribution, we change signs of the abnormal returns for events in each short portfolio. Second, we calculate top-flop return statistics. To this end, we consider earnings events from the top ($r^T_{s,q} > 0.9$) or flop decile ($r^T_{s,q} \leq 0.1$). For events in the flop decile, we again change the sign of abnormal returns.

3.6. Announcement-based trading strategy

In their work on the post-earnings-announcement drift, Bernard and Thomas (1989) present a zero-investment trading strategy as further check of the apparent abnormal returns. In a similar manner, we assess the financial performance of the model's predictions in a straightforward, zero-investment trading strategy that would have been relatively easy to implement for an investor. Our trading strategy proceeds as follows: For every earnings event, we apply the trained model and calculate the rolling percentage rank $r^5_{s,q}$ as described in Section 3.5. We use a forecast horizon of $T = 5$ days as we expect the largest part of the price reaction to occur within the first few days (compare, for example, Bernard and Thomas, 1989). If the event falls into the top decile (i.e., $r^5_{s,q} > 0.9$), we generate a trading signal for a long position in stock s. Similarly, we generate short signals for events in the flop decile ($r^5_{s,q} \leq 0.1$). No trading signal is generated for events with intermediate ranks. For long trading signals, we initiate a position in the respective stock at the day following the earnings call by buying at the open price $P^O_{s,1}$. Similarly, we short-sell stocks with a short trading signal. We use the following strategy for capital allocation. Each day, we allocate one fifth of the total available capital for potential investments in trading signals. This daily amount of capital is then divided equally between trading signals, irrespective of the direction of the signal. We further assume that all non-invested capital is held as cash position with no interest. This allocation strategy ensures that no more than 20 percent of total capital is invested in one stock, and that the leverage on invested capital is always smaller than two. Keeping the ease of implementation in mind, we offset positions with an opposing investment in the iShares Core Total U.S. Stock Market ETF (ticker symbol ITOT, BlackRock Inc., 2020). Instead of replicating the size-decile portfolios used to calculate abnormal returns, this approach is common in the statistical arbitrage literature to achieve market-neutrality (see, for example, Avellaneda and Lee, 2010; Schnaubelt et al., 2020). We assume transaction costs of 10 bp, 15 bp and 20 bp for stocks in the S&P500, S&P400 and

S&P600, respectively. These assumptions are well in line with the empirical estimates of de Groot et al. (2012) for the period from 2000 to 2009, and are conservative with regard to the range of 2 bp to 10 bp for the complete U.S. stock universe for the more recent period from 2000 to 2015 (Jha, 2016).

4. Results

The presentation of our results proceeds in five steps: First, Section 4.1 compares the performance of the random forest model with our benchmarks. To better understand the model's decision making, we perform further in-depth analyses of our results: In a second step, we therefore break down results in terms of industry sector, size and value-growth classification of stocks (Section 4.2). Third, we investigate the importance of feature groups and single features (Section 4.3). Fourth, we further analyze the non-linear functional relationship between features and the prediction learned by the model (Section 4.4). Fifth, Section 4.5 discusses the financial performance of our trading strategy that leverages the model's predictions.

4.1. Overview

4.1.1. Comparison of models by out-of-sample performance

First, we compare our predictive models based on buy-and-hold abnormal returns in the out-out-sample period, i.e., for all earnings events from 2013 to 2019. Table 4 reports return statistics for forecast horizons of 5, 20 and 60 days and all models. Panel A depicts results for all earnings events. By contrast, Panel B considers only events in the top or flop deciles according to the rolling rank of predicted returns. Looking first at the results from all earnings events (Panel A) for the two simplest benchmark models, i.e., the earnings surprise model (ES) and the linear regression (LR), we find that the linear regression model yields little or no improvement over the earnings surprise model. For a forecast horizon of 5 days, the linear model improves the mean return from 14.1 bp to 21.8 bp. For the two longer forecast horizons, we find that the linear model performs slightly worse than the earnings surprise model. Turning to the subset of earnings events in the top and flop deciles (Panel B), we find that the linear model does somewhat improve out-of-sample mean abnormal returns. We find the largest improvement for a forecast horizon of 60 days, with a mean return of 94.5 bp for the linear model compared to 45.8 bp for the earnings surprise model.

Next, we substitute predictors with respective decile indicator variables with the binned linear regression model (LR-B) to allow for non-linear feature impact, and find a considerable improvement in results. In terms of mean abnormal return for all earnings events, we observe a nearly twofold increase. For a forecast horizon of 60 days, the mean return increases from 22.8 bp for the earnings

	Forecast horizon											
	5 days				20 days				60 days			
	ES	**LR**	**LR-B**	**RF**	**ES**	**LR**	**LR-B**	**RF**	**ES**	**LR**	**LR-B**	**RF**
Panel A: All out-of-sample earnings events												
Count	32416	32416	32416	32416	32416	32416	32416	32416	32416	32416	32416	32416
Mean	0.141	0.218	0.286	0.319	0.122	0.079	0.273	0.408	0.228	0.213	0.411	0.567
Mean (long)	0.348	0.422	0.490	0.522	0.286	0.243	0.437	0.566	0.726	0.711	0.902	1.059
Mean (short)	-0.064	0.012	0.080	0.111	-0.041	-0.085	0.109	0.243	-0.268	-0.282	-0.086	0.071
t-statistic	5.030	7.751	10.206	11.357	2.756	1.789	6.181	9.252	2.993	2.794	5.382	7.430
First quart.	-2.249	-2.212	-2.148	-2.102	-3.692	-3.811	-3.640	-3.464	-6.679	-6.733	-6.524	-6.332
Median	0.078	0.091	0.164	0.197	0.079	-0.037	0.129	0.274	0.072	0.073	0.224	0.344
Third quart.	2.430	2.467	2.547	2.572	3.927	3.796	3.970	4.146	6.826	6.767	6.962	7.125
Std. dev.	5.060	5.057	5.054	5.052	7.949	7.950	7.946	7.940	13.745	13.745	13.740	13.735
Dir. acc.	50.775	50.888	52.129	52.376	50.479	49.785	50.896	51.373	50.243	50.182	50.663	50.678
Panel B: Earnings events from the flop and top deciles												
Count	6447	6516	6525	6529	6447	6475	6473	6482	6447	6471	6479	6464
Mean	0.531	0.643	0.874	0.954	0.397	0.539	0.790	1.142	0.458	0.945	1.542	1.936
Mean (long)	1.188	1.077	1.361	1.496	1.062	0.976	1.433	2.083	1.759	1.970	2.694	3.305
Mean (short)	-0.121	0.213	0.385	0.408	-0.268	0.085	0.138	0.201	-0.854	-0.093	0.371	0.541
t-statistic	5.883	8.307	10.870	11.140	2.809	4.423	6.460	8.492	1.913	4.338	7.000	8.078
First quart.	-2.663	-2.175	-2.056	-2.312	-4.678	-3.867	-3.714	-4.107	-8.382	-7.292	-6.701	-7.284
Median	0.408	0.225	0.448	0.584	0.290	0.073	0.351	0.459	0.327	0.171	0.791	1.135
Third quart.	3.613	2.830	3.125	3.799	5.487	4.314	4.604	5.413	9.200	7.744	8.448	9.914
Std. dev.	7.253	6.244	6.492	6.921	11.359	9.810	9.842	10.830	19.219	17.521	17.736	19.265
Dir. acc.	53.732	52.516	54.603	55.416	51.619	50.657	52.263	52.701	51.265	50.658	53.057	53.573

Table 4: **Abnormal post-announcement return characteristics by model and forecast horizon.** This table depicts out-of-sample return characteristics for three different forecast horizons and the earnings surprise (ES), linear regression (LR), linear regression with categorical feature encoding (LR-B) and random forest (RF) model. Panel A shows mean buy-and-hold abnormal returns for all out-of-sample events from 2013 to 2019. By contrast, Panel B shows mean buy-and-hold abnormal returns for events in the flop or top deciles. Both panels exhibit mean abnormal returns and corresponding t-statistics adjusted for heteroscedasticity and serial correlation. Return values are given in percent. The values in row *Dir. acc.* are directional, balanced accuracies calculated from the signs of predicted and actual abnormal returns.

surprise model to 41.1 bp for the binned linear model. A comparable improvement is apparent in mean and median returns for events in the top and flop deciles. We carefully infer that the observed abnormal return after earnings events can be attributed to non-linear effects. This finding is not surprising: First, there is ample evidence of a non-linear behavior of abnormal returns on earnings surprises (see, for example, Freeman and Tse, 1992; Cheng et al., 1992; Kothari, 2001; Dellavigna and Pollet, 2009). Second, other studies applying machine learning to the prediction of asset returns report similar improvements over linear models (Fischer and Krauss, 2018; Gu et al., 2018).

We observe a further improvement in out-of-sample returns for the random forest model, which allows for a large range of non-linear responses and additionally permits feature interactions. For all forecast horizons, we find that the random forest yields the largest mean abnormal return. For the exemplary forecast horizon of 5 days, the mean abnormal return from the random forest (31.9 bp) surpasses corresponding results from the earnings surprise and binned linear models (14.1 bp and 28.6 bp, respectively). This improvement is even more pronounced for longer forecast horizons. We can reject the null hypothesis of zero mean abnormal returns at the 1 percent level for the random forest model and all forecast horizons. Abnormal returns for events in the extremes of the ranking, i.e., in the top and flop deciles, are considerably higher than for the average event. For a forecast horizon of 5 days, the mean top-flop abnormal return triples from 31.9 bp to 95.4 bp. Compared to all other models, mean abnormal top-flop returns from the random forest are by far largest. We find that the random forest also yields the highest median returns. For example, median top-flop abnormal returns for a forecast horizon of 60 days are at 113.5 bp, compared to the second-best binned linear model with 79.1 bp. We find larger mean abnormal returns in the top decile than in the flop decile for all forecast horizons. For a forecast horizon of 60 days, we find mean abnormal returns of 3.305 percent in the top decile, but only 0.51 percent in the flop decile (Panel B). When looking at results from all out-of-sample events (Panel A), we similarly find mean returns from long events (1.059 percent) to be significantly larger than from short events (0.071 percent). The directional accuracy, i.e., the balanced accuracy to predict the correct sign of the abnormal return, is generally highest for the random forest model. Take, as an example, the directional accuracy across all out-of-sample events for a forecast horizon of 20 days, which is at 51.37 percent for the random forest compared to 50.90 percent for the binned linear model. We also find that the directional accuracy of the predictions increases for events in the top-flop deciles: For the 5-day horizon, the random forest achieves a balanced accuracy of 55.42 percent, compared to 52.38 percent for all earnings events.

To evaluate the statistical significance of abnormal return differences between models, we apply the dependent two-sample t-test. The p-values given in Panel A of Table 5 refer to the null

	Forecast horizon								
	5 days			20 days			60 days		
	ES	LR	LR-B	ES	LR	LR-B	ES	LR	LR-B
Panel A: Paired two-sample t-test									
LR	0.019			0.230			0.441		
LR-B	0.000	0.012		0.005	0.000		0.042	0.012	
RF	0.000	0.001	0.153	0.000	0.000	0.003	0.000	0.000	0.035
Panel A: Wilcoxon signed-rank test on all earnings events									
LR	0.258			0.995			0.624		
LR-B	0.001	0.003		0.179	0.000		0.105	0.047	
RF	0.000	0.000	0.054	0.000	0.000	0.000	0.002	0.000	0.050

Table 5: **Comparing models by out-of-sample abnormal return.** In this table, we list p-values for the pairwise comparison of our predictive models. Panel A reports results from a paired two-sided t-test applied to all out-of sample earnings events with the null hypothesis that the mean abnormal return of the row model is smaller than the one of the column model. Panel B reports results from the Wilcoxon signed-rank test applied to all out-of sample earnings events with the null hypothesis that median abnormal return of the row model is smaller than the one of the column model. Abnormal returns with a below-median rank $r_{s,q}$ are negated. The table reports results for forecast horizons of 5, 20 and 60 days.

hypothesis of a positive difference in mean abnormal return between the column and the row model. We can reject this null hypothesis for the earnings surprise and the linear model only for a forecast horizon of five trading days. The increase in abnormal returns observed with the binned linear model is however statistically significant for all forecast horizons. Similarly, we can reject the null hypothesis for the random forest model and the earnings surprise or linear model for all horizons. As discussed before, switching from the binned linear regression to the random forest leads to a relatively small increase in mean abnormal returns, and consequently, we find it to be statistically significant for forecast horizons of 20 and 60 days only. We also apply the Wilcoxon signed-rank test (Wilcoxon, 1945) with Pratt's treatment of zeros (Pratt, 1959). The results (Panel B of Table 5) generally lead to similar conclusions, however the median return of the random forest model is now statistically significantly larger than for the binned linear regression also for a forecast horizon of five days.

4.1.2. Evolution of post-announcement returns

Abnormal returns increase with the forecast horizon for both the binned linear and the random forest models: For forecast horizons of 5, 20 and 60 days, mean abnormal returns for all earnings events are at 31.9 bp, 40.8 bp and 56.7 bp, respectively (Panel A of Table 4). Similarly, mean top-flop returns are at 95.4 bp, 114.2 bp and 193.6 bp, respectively (Panel B of the same table). From the literature on the post-earnings-announcement drift (PEAD), we would have expected to find such a drift also for the model based on the earnings surprise only (see, for example, Ball and Brown, 1968; Foster et al., 1984; Bernard and Thomas, 1989). We conjecture that the PEAD has weakened considerably for our sample period, and that the drift observed with the random forest

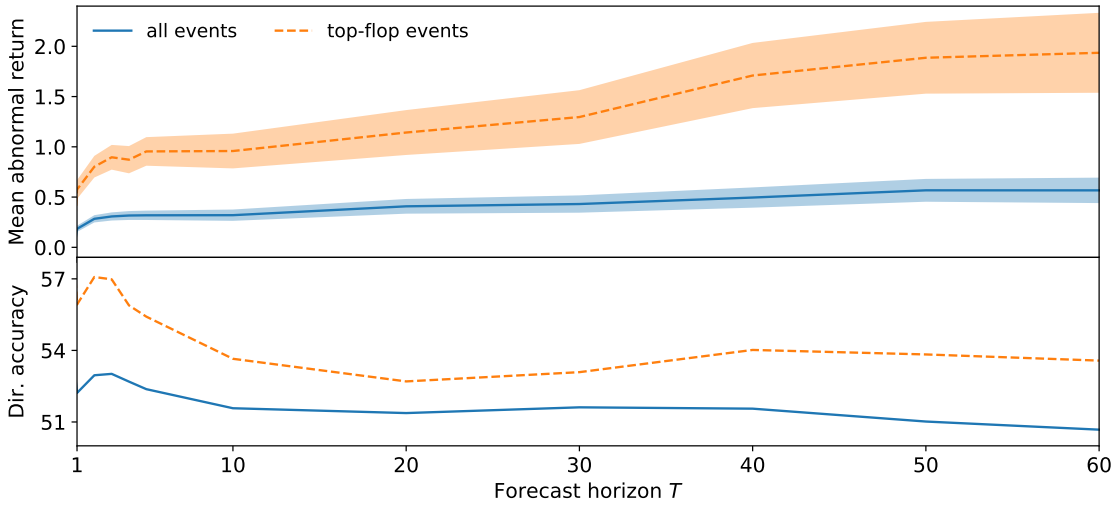

Figure 2: **Evolution of mean abnormal return and directional accuracy with forecast horizon.** This figure depicts mean abnormal returns (upper plot) and balanced directional accuracies (lower plot) by forecast horizon for the random forest model. The solid and dashed lines show results for all earnings events and events from top and flop deciles, respectively. Shaded regions in the upper plot indicate 90% confidence bands.

model might not be due to the earnings surprise alone. Figure 2 shows that abnormal returns from the random forest model are monotonically increasing also for intermediate forecast horizons. We find that a large part of the top-flop drift (57.5 bp) originates from the open-to-close return of the first day following the call (forecast horizon of one day). It takes the following nineteen days to accumulate a similar magnitude of abnormal return (56.7 bp). By contrast, we see that the directional accuracy of predictions is at the maximum shortly after the earnings event, and declines afterwards (lower plot of Figure 2).

4.1.3. Robustness checks

These results are very robust with respect to variations of the seed value of the random number generator[16] and the choice of tuning parameters of the random forest model. Table D.11 in Appendix D shows how results change when we vary the seed of the random number generator used with the random forest. We find that the relative variation of results between seed values amounts to only a few percent. Table D.12 presents results from the random forest model using alternative parameter settings. The results show that the maximum tree depth J, the number of trees B and the length of the rolling window W can be changed over large ranges without materially affecting results. Also, we find returns to be fairly stable over our sample period (compare Figure E.12 in Appendix E).

[16]The random number generator influences the model's growth of trees due to randomly chosen bootstrap samples and split variables (compare Section 3.4.1).

	Selected top-flop events			All events			Count ratio
	Count	**Mean ret.**	**t-stat.**	**Count**	**Mean ret.**	**t-stat.**	
Panel A: Forecast horizon 5 days							
Basic Materials	492	1.158	4.336	2056	0.526	4.926	0.239
Consumer Cyclicals	1185	1.153	7.024	4748	0.426	6.174	0.250
Consumer Non-Cyclicals	343	1.109	3.015	1772	0.297	2.518	0.194
Energy	534	1.990	5.460	1889	0.434	3.158	0.283
Financials	1080	0.437	3.229	6973	0.194	4.427	0.155
Healthcare	858	0.793	3.222	3566	0.405	2.867	0.241
Industrials	841	1.108	5.346	5234	0.249	4.207	0.161
Technology	992	0.763	3.560	4680	0.381	4.660	0.212
Telecom. Services	72	-0.536	-0.635	320	-0.169	-0.520	0.225
Utilities	132	0.361	1.200	1178	0.048	0.539	0.112
Panel B: Forecast horizon 60 days							
Basic Materials	486	2.894	3.327	2056	1.008	3.140	0.236
Consumer Cyclicals	1106	1.889	3.946	4748	0.766	3.902	0.233
Consumer Non-Cyclicals	333	2.847	2.914	1772	1.081	3.633	0.188
Energy	768	2.562	3.230	1889	0.522	1.092	0.407
Financials	1065	0.917	2.209	6973	0.106	0.867	0.153
Healthcare	802	2.665	3.641	3566	0.821	3.082	0.225
Industrials	831	1.565	3.406	5234	0.602	4.124	0.159
Technology	885	2.248	4.402	4680	0.693	3.531	0.189
Telecom. Services	59	-4.399	-2.025	320	0.511	0.649	0.184
Utilities	129	-0.343	-0.351	1178	-0.393	-1.438	0.110

Table 6: **Contribution analysis by industry classification and forecast horizon.** This table reports event counts and mean abnormal returns conditional on industry classification for forecast horizons of 5 and 60 days and for the random forest model. Industry classifications are according to the Thomson Reuters Business Classification scheme. The table shows statistics for all events as well as those selected in the top-flop deciles. The last column provides the fraction of selected top-flop events in the respective sector.

4.2. Detailed breakdown by sector, size and value-growth classification

Next, we calculate out-of-sample mean abnormal returns from the random forest model conditional on industry sector, firm size and value-growth classification to analyze whether results are driven by specific stock characteristics.

Industry sector: First, we investigate whether results are primarily influenced by specific industries. Table 6 compares industry-specific counts and mean abnormal returns separately for all out-of-sample earnings events and those in the top and flop deciles. First, we see that the selection of earnings events in the top and flop deciles roughly follows the distribution of industries in the S&P1500 index. The shares of selected top-flop events (last column of Table 6) in specific industries are relatively similar. Second, average abnormal returns for financial, telecommunication and utility stocks are lowest or even negative. As the nature of business of these stock is often different from other sectors, also their accounting information has to be interpreted differently.[17] It

[17]Some researchers therefore exclude these industries from their sample in empirical studies, see, for example, Jegadeesh and Livnat (2006b).

	Forecast horizon								
	5 days			20 days			60 days		
	S&P 500	S&P 400	S&P 600	S&P 500	S&P 400	S&P 600	S&P 500	S&P 400	S&P 600
Panel A: All out-of-sample earnings events									
Count	11404	8819	12193	11404	8819	12193	11404	8819	12193
Mean	0.157	0.307	0.478	0.333	0.215	0.618	0.236	0.464	0.950
t-statistic	5.035	6.406	8.073	6.223	2.665	6.983	2.494	3.221	6.350
Median	0.134	0.174	0.303	0.239	0.214	0.376	0.216	0.175	0.641
Std. dev.	3.336	4.505	6.535	5.715	7.582	9.767	10.125	13.538	16.518
Dir. acc.	52.90	51.77	52.42	51.93	50.67	51.38	50.03	50.52	51.40
Panel B: Earnings events from the flop and top deciles									
Count	1333	1548	3648	1248	1533	3701	1445	1622	3397
Mean	0.579	0.872	1.126	0.439	0.732	1.549	1.246	1.699	2.341
t-statistic	5.055	5.559	8.563	2.090	2.916	7.789	3.598	3.575	6.384
Median	0.494	0.434	0.730	0.113	0.401	0.705	1.211	0.501	1.460
Std. dev.	4.187	6.171	7.944	7.426	9.831	12.102	13.174	19.150	21.380
Dir. acc.	51.84	54.50	55.15	50.90	51.74	52.70	54.83	51.50	54.29
Panel C: Fraction of earnings events in the top or flop deciles									
Count ratio	0.117	0.176	0.299	0.109	0.174	0.304	0.127	0.184	0.279

Table 7: **Return characteristics by index segment and forecast horizon.** In this table, we report return characteristics conditional on the S&P1500 sub-index and the forecast horizon for the random forest model. Panel A reports statistics for all earnings events. Panel B depicts corresponding statistics for those earnings events selected in the top or flop decile. Panel C compares the number of top-flop events to the total number of events in the out-of-sample period.

is therefore not surprising that our model, which is trained on the cross-section of stocks, performs worse for these industries. Third, we conclude that these findings are fairly independent of the considered forecast horizon.

Firm size: Next, we break down our results by firm size, which we approximate by the three market capitalization segments of the S&P1500 index. Table 7 compares return statistics for the three sub-indices of the S&P1500, i.e., the S&P500 large-cap, the S&P400 mid-cap and the S&P600 small-cap indices. First, we find that abnormal returns are larger for smaller firms: For example, five-day mean top-flop returns for the small-cap S&P600 index are at 112.6 bp, which is considerably higher than for the S&P400 (87.2 bp) and S&P500 (57.9 bp) indices (Panel B). Similar patterns are evident for the other forecast horizons and for all out-of-sample earnings events (Panel A). These findings are well in line with research on the size effect, first reported by Banz (1981), as well as with studies on the PEAD (see, for example, Bernard and Thomas, 1989; Garfinkel and Sokobin, 2006). Following the argument of Zhang (2006), investors have higher costs when acquiring information on small stocks, which might increase the potential for misvaluation.

Second, there is a selection preference for stocks with lower market capitalization. Panel C depicts the fraction of earnings events in the top-flop deciles relative to the overall number of

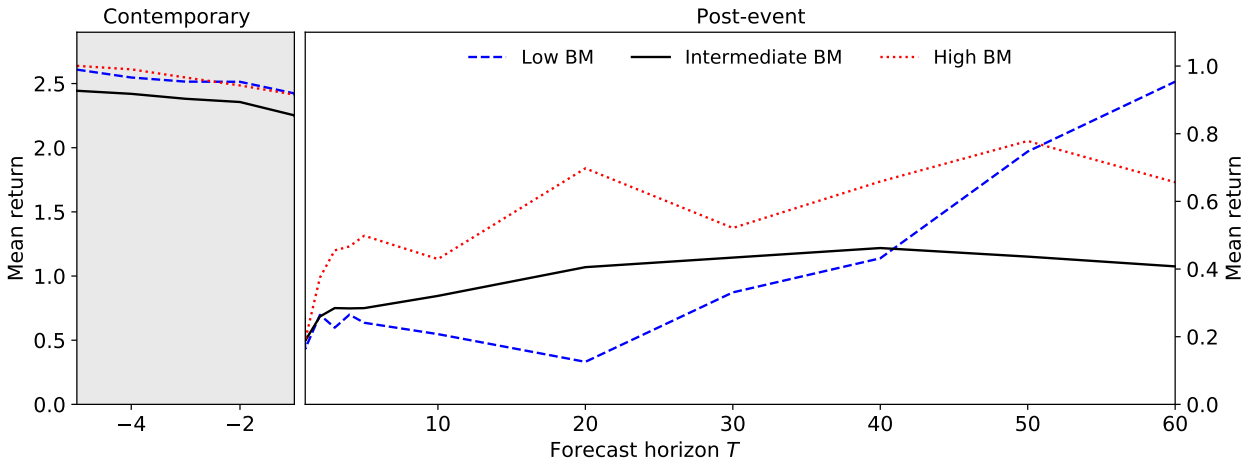

Figure 3: **Evolution of mean abnormal returns by book-to-market ratio.** This figure compares the evolution of mean buy-and-hold abnormal returns for all earnings events, conditional on the book-to-market ratio (BM). The dashed blue (dotted red) lines depict the evolution of mean returns for events with a book-to-market ratio in the lowest (highest) quintile. The solid black line shows mean returns for the intermediate three quintiles.

events in the corresponding index segment. For a forecast horizon of 60 days, we observe that 27.9 percent of the S&P600 stocks are in a top or flop decile, but only 12.7 percent of the S&P500 stocks. With a selection frequency of 18.4 percent, the S&P400 constituents are about as frequently selected as one would expect from an independent selection (20 percent). We conjecture that this selection preference might be due to two possible reasons: First, the model might have learned an announcement-specific size effect, i.e., the market capitalization variable interacts with other variables such that predictions for smaller firms are more extreme. Second, there might be a higher variance in feature values for smaller firms, leading to a larger magnitude in abnormal return predictions. In both cases, more extreme return predictions for smaller firms would lead to their more frequent selection into top or flop deciles.

Value-growth classification: Finally, we analyze the influence of a stock's value-growth classification on out-of-sample abnormal returns. Various authors in finance and accounting often use the terms value or growth/glamour to differentiate stocks and use the book-to-market (BM) ratio for classification (see, for example, Lakonishok et al., 1994; Fama and French, 1998; Desai et al., 2004; Campbell et al., 2010). Desai et al. (2004) argue that the BM ratio subsumes information from other fundamental ratios that are also frequently used to distinguish value and growth stocks, such as the earnings-to-price and the cashflow-to-price ratio. We apply the same characterization criteria for an ex-post inspection of mean abnormal returns by assigning stocks to three classes with low, intermediate and high values of the book-to-market ratio, and depict results for all earnings events in Figure 3.[18] We observe that returns for the contemporary return window are similar

[18] We apply three cutoffs for the value-growth classification, i.e., 10/80/10, 20/60/20 and 30/40/30. All three

for the three stock types. By contrast, for post-announcement windows, the behavior of abnormal returns for value or growth stocks is noticeably different from stocks in the intermediate class. In line with various studies on the value premium (see, for example, Basu, 1977; Fama and French, 1989; Daniel and Titman, 1997), we find that abnormal returns for value stocks are larger than for growth stocks for short to medium forecast horizons. Taking into account the previous observation that a large fraction of the value premium originates from a short time horizon after earnings announcements (Porta et al., 1997), it is unsurprising that we find – similar to the value premium effect – larger returns for value stocks than growth stocks immediately after the earnings announcement. However, returns from growth stocks seem to drift from around 20 bp to nearly 95 bp over the course of the last 30 trading days, and yield higher returns than value stocks in the long run. Figure E.13 in the appendix shows results from a similar value-growth analysis for top-flop mean abnormal returns. We find that mean abnormal top-flop returns for growth stocks exhibit a late drift, starting from approximately 30 trading days after the earnings announcement. In contrast to our findings from all earnings events, returns for value stocks are larger than for growth stocks for all forecast horizons, which might be related to a preferred selection of value stocks into top or flop deciles. As seen from the analysis on firm size, the model preferably selects stocks from smaller firms. Chan and Lakonishok (2004) find that the value premium is higher for small-cap stocks, which might explain the generally higher level of mean abnormal top-flop returns for value stocks than for growth stocks.

Overall, the results from the value-growth analysis seem consistent with the following explanation: For value stocks, there is little stock-specific uncertainty that the equity is currently undervalued, as information pertaining to fundamental ratios is sufficient to determine the current mispricing. On the other hand, a common conception about growth stocks is that they are valued above their fundamental value in consequence of a series of good past news (Hong and Stein, 1999). To determine whether this overvaluation is justified, information from valuation ratios is insufficient. Hence, investors tend to use more complex information, such as analyst sentiment, to arrive at their valuation, which potentially delays the pricing process.[19] In consequence, the model's performance for growth stocks increases at longer forecast horizons.

4.3. Analyzing the importance of features

Next, we analyze the out-of-sample predictive ability of feature groups and the in-sample importance of single features to shed further light into the machine learning black box.

divisions show a similar evolution of returns.

[19] In Sections 4.3 and 4.4, we find that other variables such as textual sentiment play a pronounced role for long-term returns.

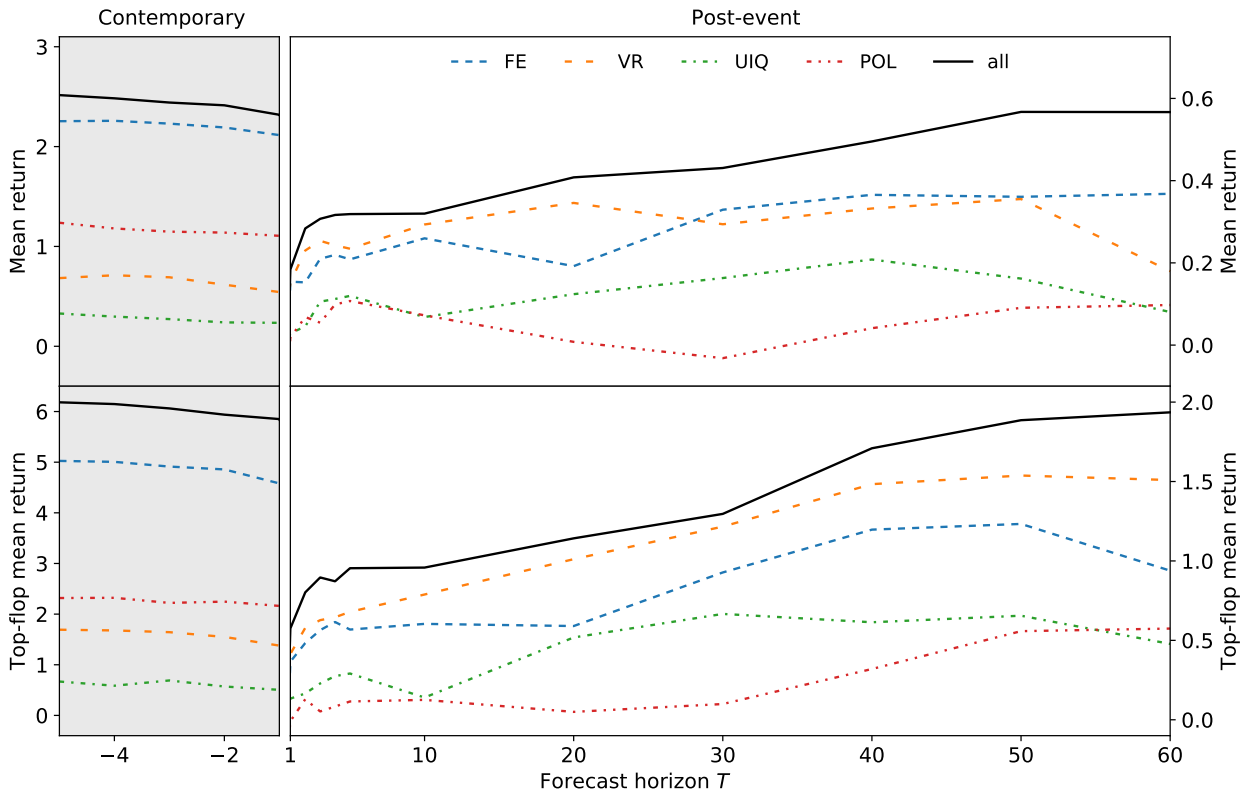

Figure 4: **Evolution of mean abnormal returns by feature group.** This figure compares mean buy-and-hold abnormal returns for different feature groups. We retrain the random forest model with features from single feature groups for different forecast horizons T, and evaluate mean abnormal returns from all earnings events (upper figure) as well as earnings events in the top and flop deciles (lower figure). The left plots display results for contemporary return periods, i.e., abnormal returns are calculated such that the earnings event is enclosed by the return window. The right plots show results for post-event return windows. The figure shows results for all four feature groups, i.e., forecast errors *(FE)*, valuations ratios *(VR)*, uncertainty/information quality *(UIQ)* and textual sentiment polarity *(POL)*, as well as for all features *(all)*.

Predictive ability of feature groups: To assess the predictive power of our four feature groups, we first retrain the random forest model using only features from single feature groups, and contrast out-of-sample abnormal returns to those obtained with all features. We depict results in Figure 4, which separates results for contemporary return periods (negative forecast horizons, left plots) and post-event drift periods (positive forecast horizons, right plots). The upper (lower) plots show abnormal returns for all (top-flop) earnings events. The model based on all features generally yields higher out-of-sample mean abnormal returns than models based on single feature groups. For contemporary return windows, we find that the model based on forecast errors achieves by far the best performance (abnormal top-flop returns of 4.58 to 5.02 percent). The feature group with the second-best predictive ability are textual sentiment polarity features (POL), which alone procure contemporary abnormal returns of 2.16 to 2.32 percent. Similarly, Henry (2006) reports that verbal information from earnings press releases improves the prediction of contemporary returns.

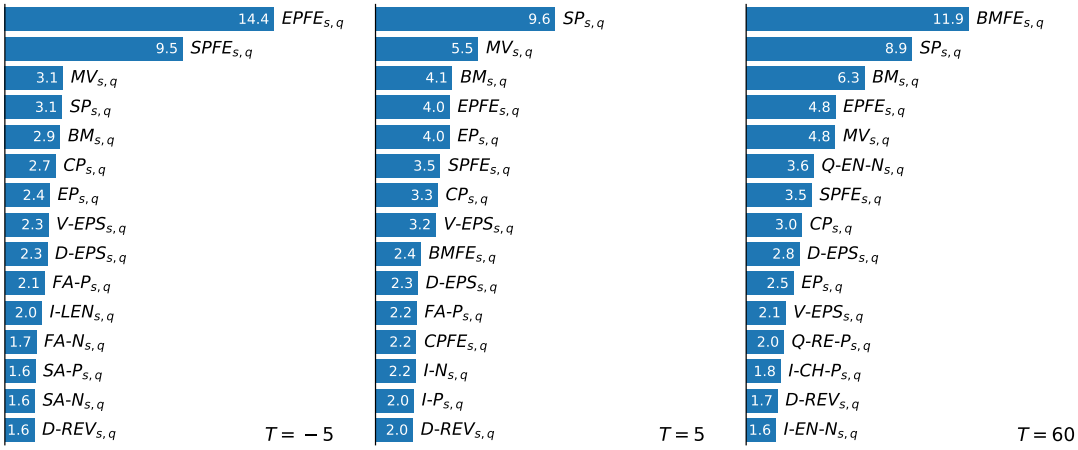

Figure 5: **Feature importance by forecast horizon.** This figure depicts feature importances in terms of the mean decrease in impurity for the 15 most important features. The left plot displays results for the contemporary return period ($T = -5$), the central and right plots for drift windows with a forecast horizon of $T = 5$ and $T = 60$ trading days, respectively.

Generally, we find that most of the contemporary price reaction occurs in the shortest contemporary return window, i.e., from market close on the day before the earnings call to market open on the day after.

For post-announcement forecast horizons, models based on forecast errors (FE) and valuation ratios (VR) achieve the largest abnormal returns for the entirety of earnings events. Also, abnormal returns from these two feature groups are on a similar level for most forecast horizons. In comparison, abnormal returns from models based solely on uncertainty/information quality and textual sentiment polarity features have inferior standing. We find that abnormal returns from the model based on variables on uncertainty and information quality (UIQ) peak for intermediate forecast horizons of 20 to 40 days. Abnormal returns based on sentiment features tend to increase for long-term horizons. Regarding top-flop abnormal returns, we find that models based on forecast errors and valuation ratios are similar in performance for short forecast horizons of up to five days. For longer return windows, features in the valuation ratio group alone achieve a better top-flop performance than forecast error features. Valuation ratios generate close to 80 percent of overall post-event abnormal top-flop returns. For example, we find that 67.9 bp of 95.4 bp could be due to valuation rations alone for a forecast horizon of 5 days, and 150.9 bp of 193.6 bp for a forecast horizon of 60 days.

Evolution of in-sample feature importance: Next, we analyze which features are predominantly selected by the model to predict abnormal returns. Following Krauss et al. (2017) and Schnaubelt et al. (2020), we calculate feature importance as the mean decrease in impurity (MDI) for each feature during training, and average results from all study periods. Figure 6 depicts the evolution

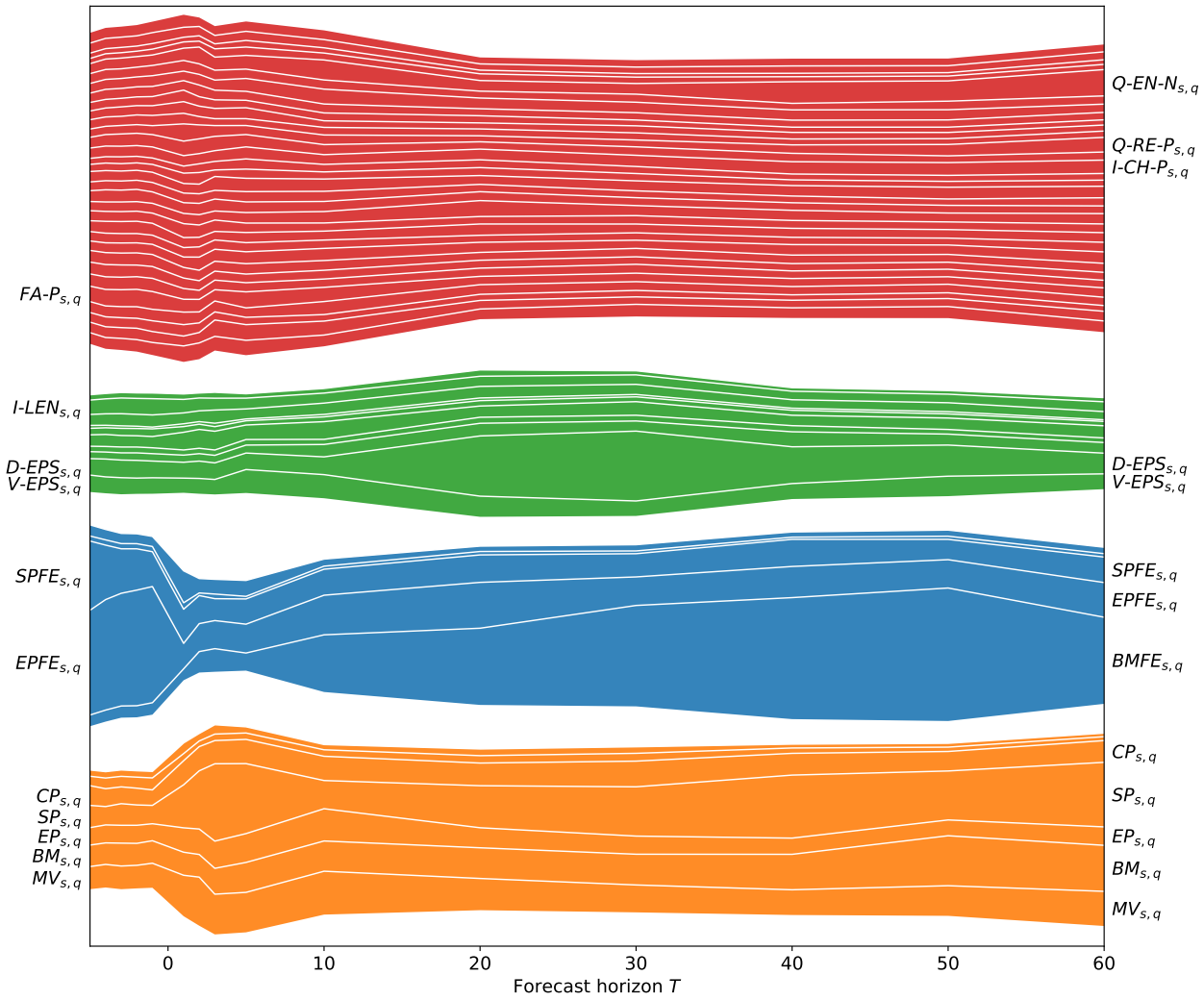

Figure 6: **Evolution of feature importance.** This figure depicts feature importances of the random forest model in terms of the mean decrease in impurity as a function of forecast horizon. We average feature importances for every study period during model training. Negative (positive) forecast horizons correspond to contemporary returns (post-event drift returns). The height of the colored band shows the total feature importances for the following feature groups: orange: valuation ratios (VR), blue: forecast errors (FE), green: uncertainty/information quality (UIQ), red: textual sentiment polarity (POL). Single-feature importances are reflected by the further subdivision of bands.

of mean feature importances by forecast horizon. The widths of the four colored bands illustrate the total feature importance in the respective feature group. The contribution of single features is shown by the further division of bands. Figure 5 displays the same data for the 15 most important features for the contemporary window and forecast horizons of 5 and 60 trading days. Variables commonly used in asset pricing and studies of earnings announcements are most frequent in all time horizons. For the contemporary return window, we observe that by far the most important predictors of the price reaction are the earnings surprise ($EP\text{-}FE_{s,q}$) and the revenue surprise ($SP\text{-}FE_{s,q}$) variables, with relative feature importances of 14.4 and 9.5 percent (left chart of Figure 5).

Next in feature importance are valuation ratios. In this group, cashflow-to-price ($CP_{s,q}$), sales-to-price ($SP_{s,q}$), earnings-to-price ($EP_{s,q}$) and book-to-market ($BM_{s,q}$) ratios as well as the market value ($MV_{s,q}$) exhibit similar feature importances.

By contrast, the model selects a wider range of features to predict post-announcement returns. The previously dominating earnings and revenue surprise variables lose in feature importance and are supplemented by a variety of predictors. Consistent with our previous observation that a model based on valuation ratios alone yields the largest mean abnormal returns, features with the highest feature importances are valuation ratios. For a forecast horizon of 5 trading days, the five most important features are the sales-to-price ratio ($SP_{s,q}$) with 9.6 percent, market value ($MV_{s,q}$, 5.5 percent), book-to-market ratio ($BM_{s,q}$, 4.1 percent), earnings forecast error ($EP\text{-}FE_{s,q}$, 4.0 percent) and earnings-to-price ratio ($EP_{s,q}$, 4.0 percent). For forecast horizons beyond 10 trading days, we find that the most important feature is the forecast error related to the book-to-market ratio.[20] In comparison, variables from the uncertainty and information quality feature group are less frequently selected. We find that the earnings forecast dispersion ($D\text{-}EPS_{s,q}$) and variance of the consensus earnings forecast ($V\text{-}EPS_{s,q}$), both proxies for the uncertainty of the earnings forecast, exhibit the largest feature importances within this feature group. For intermediate post-announcement forecast horizons, the earnings forecast dispersion ($D\text{-}EPS_{s,q}$) is among the most important variables. While the group of textual sentiment features has the largest feature importance shortly before and after the earnings event, it also seems to gain slightly in importance for the long-term drift. For example, for a forecast horizon of 60 days, we find that environment-related negativity in the Q&A session ($Q\text{-}EN\text{-}N_{s,q}$) has the sixth-largest feature importance (rightmost chart in Figure 5, 3.6 percent). For the contemporary return window, the positivity of those analysts first asking questions in the Q&A session ($FA\text{-}P_{s,q}$) has the largest feature importance within the group of textual sentiment features. The observation that those polarity variables with the largest feature importance are related to the Q&A section of the earnings call rather than to the prepared remarks section is consistent with results of Matsumoto et al. (2011) and Price et al. (2012), who find that the Q&A section is the more informative part of the earnings call.

Summarizing our observations of this section, we find that earnings and revenue forecast errors have the highest ability to predict the contemporary price reaction. This appears consistent with intraday studies reporting that the price reaction following earnings surprises occurs within a few hours (Patell and Wolfson, 1984; Lee, 1992; Kothari, 2001). For post-announcement abnormal returns, we make two central observations: First, valuation ratios based on the most recent

[20] We will further investigate the role of the book-to-market forecast error in Section 4.4.

accounting information tend to possess the highest predictive power, especially for the top and flop prediction deciles. Second, a larger number of variables are jointly responsible for the model's prediction.

An interpretation in terms of gradual information diffusion: A plausible interpretation of these results relates to the literature on delayed information diffusion and information processing costs.[21] The model of Hong and Stein (1999) suggests that a slow diffusion of information causes an under-reaction of stock prices. Following this picture, Engelberg (2008) examine how soft (qualitative) and hard (quantitative) information is related to the post-earnings announcement drift. He finds that qualitative information in form of the sentiment of the earnings announcement has greater predictive power than quantitative information (the earnings surprise). He suggests that higher information processing costs of qualitative information compared to quantitative information leads to a slower diffusion of soft information. In a related study of earnings press releases, Demers and Vega (2008) observe that "it takes relatively longer for soft information to be incorporated into asset prices than for hard information". Our observations seem well in line with this strand of literature: For the immediate intraday price reaction to the earnings announcement, the price seems to react primarily based on presumably readily-available and simple-to-process quantitative information. By contrast, to predict post-announcement abnormal returns, a larger number of more complex, less common and harder-to-process features are chosen by the model. For example, the forecast error related to the book-to-market ratio ($BMFE_{s,q}$), presumably a less common forecast error and therefore less readily available, exhibits the largest feature importance for long-term predictions. Also, variables on the sentiment polarity of the earnings call, which are typical examples of soft information (Demers and Vega, 2008; Price et al., 2012), slightly increase in feature importance for long-term predictions. Similarly, higher information processing costs might explain the increase in importance for variables on information uncertainty, such as analyst forecast dispersion (Stickel, 1991), for intermediate forecast horizons.

4.4. Inspecting the non-linear contribution of individual features

Next, we analyze the effects of individual features on the random forest's prediction in terms of accumulated local effect (ALE) plots (Apley and Zhu, 2019). ALE plots visualize a variable's influence on the prediction of a model by accumulating local prediction changes of small data intervals. Compared to common partial dependence plots (Friedman, 2001) or plots of the conditional density, ALE plots are computationally less demanding and do not produce corrupted results with

[21] Another possible interpretation could be the limited attention of investors, who are unable to simultaneously pay attention to all public information. However, as Hong and Stein (2007) note in their survey, "for many practical purposes, [limited attention] boils down to almost the same thing".

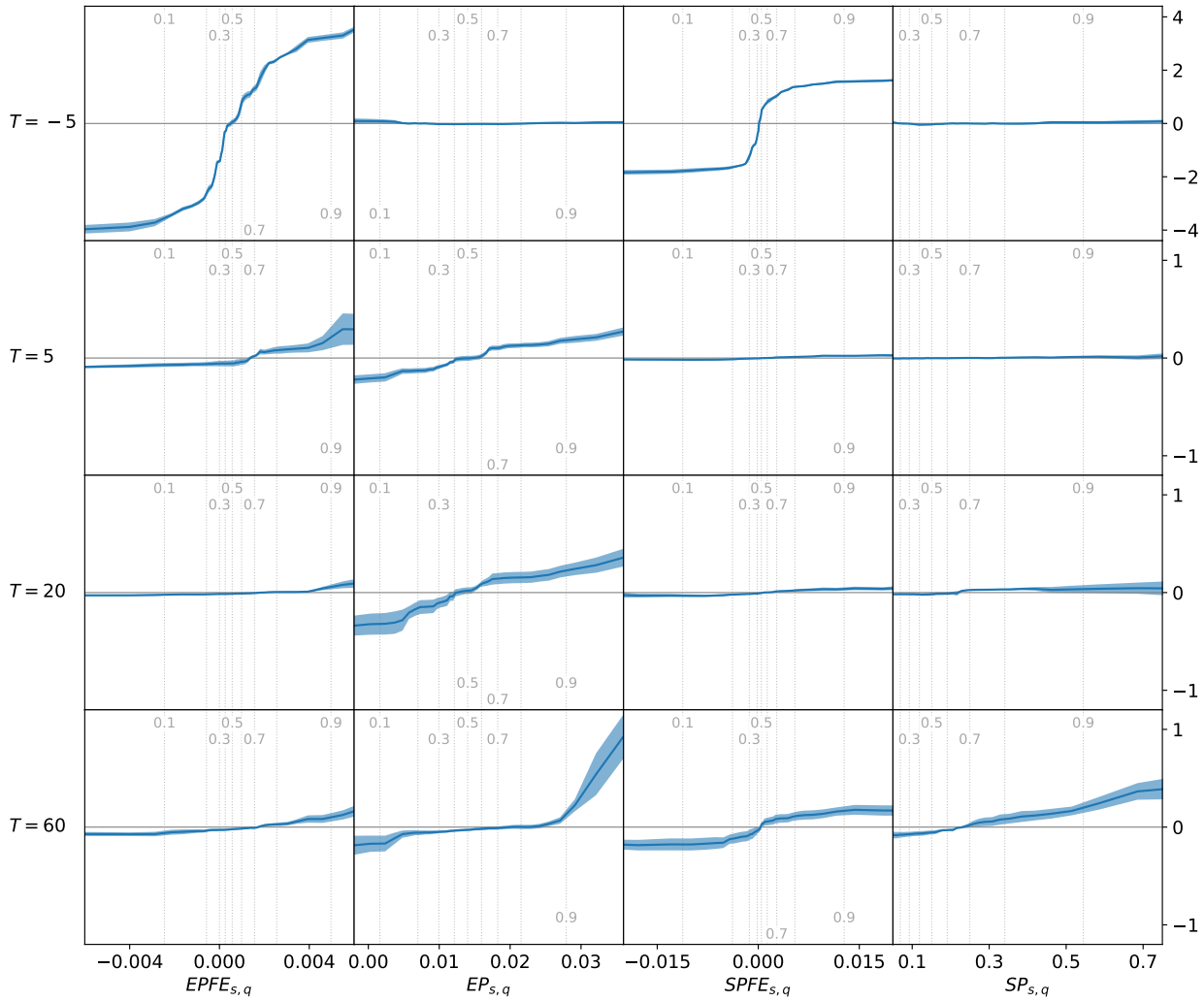

Figure 7: **Accumulated local effects plots for earnings- and revenue-based forecast errors and valuation ratios.** These plots show the accumulated local effects for earnings- and revenue-related forecast errors and valuation ratios for the contemporary return window ($T = -5$ days; top row) and three different drift periods ($T = 5, 20, 60$), in percent. Accumulated local effects show the cumulative changes in the random forest's abnormal return prediction (vertical axis) when varying the indicated feature's value (horizontal axis). We plot the mean ALE curve from all model runs, and display the point-wise standard deviation between ALE curves from different runs as shaded bands to provide some indication of the curve's variation between models trained with different data sets. Dotted vertical lines depict quantiles of the respective feature's values.

correlated predictors. Appendix B describes the computation of ALE plots in further detail. In the following, we focus on the most important features according to our analysis of feature importance (Section 4.3) and on common stock characteristics from the literature.

Contrasting earnings- and sales-related forecast errors to valuation ratios: In a first analysis, we compare the roles of earnings- and sales-related forecast errors with the related valuation ratios for different time horizons. According to our analysis of feature importance (Section 4.3), these forecast errors are by far the most important features for contemporary returns. Figure 7 depicts the

resulting ALE plots. The plots show the cumulative changes in the random forest's abnormal return prediction when varying the indicated feature's value, averaged over model runs. The shaded bands display the standard deviation between model runs to provide some indication on the variation of the curve's shape between model runs. We see that the model is learning several well-known effects of abnormal return behavior around earnings announcements. For the contemporary return period (top row of Figure 7), we observe a symmetric S-shaped ALE curve in both the earnings ($EP\text{-}FE_{s,q}$) and the revenue forecast error ($SP\text{-}FE_{s,q}$) with a zero crossing very close to a forecast error of zero. We find that the learned price reaction for extreme earnings surprises is about twice as strong as for extreme revenue surprises. The immediate S-shaped reaction of prices to earnings surprises is well-documented (see, for example, Kothari, 2001; Dellavigna and Pollet, 2009). The observation of an S-shaped reaction is typically explained by the hypothesis that investors do not expect extreme unexpected changes in earnings to be permanent, which results in smaller price adjustments for exceptionally large forecast errors (Kothari, 2001). By contrast, for post-announcement return periods, the learned price reaction is much weaker, and appears to be stronger for positive earnings surprises than for negative ones.

Turning to the ALE plots for valuation ratios related to earnings ($EP_{s,q}$) and sales ($SP_{s,q}$), we find that the model's predictions for the contemporary period are only weakly based on these valuation ratios. However, their influence in predicting post-announcement returns is comparably higher. We observe an almost linear relation of abnormal returns on the earnings-to-price ratio for a forecast horizon of 5 days, while for the longest horizon (60 days), primarily the top and bottom deciles of the earnings-to-price ratio affect the model's predictions. The relation between the earnings-to-price ratio and abnormal returns is positive in all cases. This observation in well in line with previous empirical findings on the return-predictive power of the earnings-to-price ratio (Basu, 1977, 1983), as well as its common use in the dissociation of value and growth stocks (Desai et al., 2004). As seen from our analysis on feature importance (Section 4.3), the relative contributions of forecast errors and valuation ratios is reversed. For the contemporary horizon (top row of Figure 7), both forecast errors have a much larger effect on the model's prediction than the valuation ratios. By contrast, earnings-to-price and sale-to-price ratios contribute more to the model's prediction than the corresponding forecast errors for post-announcement returns, especially for 60-day forecasts (bottom row of Figure 7).

Earnings-revenue interaction: We also find evidence that the model learns plausible interactions between features. Figure 8 depicts the accumulated local effects for the interaction between earnings and sales forecast errors. The heat map depicts the additional accumulated effect from the interaction of the two features, i.e., the effect that is not due to the superposition of the two

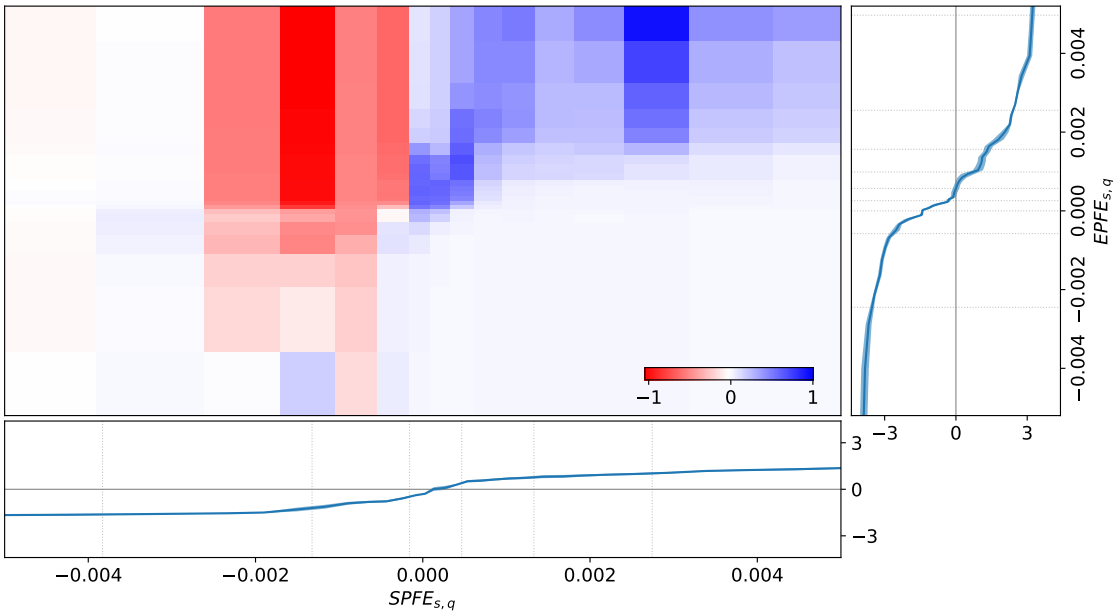

Figure 8: **Interaction between earnings- and sales forecast errors.** This heat map depicts the additional accumulated local effect for the interaction between the earnings and sales forecast errors for the contemporary return period. Blue (red) values indicate that the interaction between earnings and sales forecast errors increases (decreases) the model's prediction beyond the combined effects of the single variables. The smaller plots at the bottom and left side of the figure show the respective single-variable accumulated local effects.

single-feature effects alone. The blue area indicates that if both revenue and earnings surprise are (weakly) positive, the predicted abnormal return is amplified above the level what would be expected from the added effects of the single features alone. Similarly, the red area indicates that a negative revenue surprise reduces the positive return prediction expected from the positive earnings surprise alone. If both earnings and revenue surprise are negative, we find an additional reduction in the abnormal return prediction. These findings are well in line with the results of Jegadeesh and Livnat (2006a,b) and Chen et al. (2014) which indicate a stronger post-earnings-announcement drift when earnings and revenue surprises have the same sign.

Recovery of common capital markets effects: The ALE plots depicted in Figure 9 show how the model is identifying patterns related to size and value-growth effects. Looking first at the market capitalization variable ($MV_{s,q}$), we see that the effect on the model's predictions changes from a weakly positive relation for the contemporary return window to a negative relation for post-announcement horizons. The learned effect of firm size on abnormal returns for post-announcement windows is congruent with the findings of Banz (1981) and our previous observation in Section 4.2. We find market capitalization to notably increase the model's return prediction in the two lowest deciles, which is well in line with results from Vassalou and Xing (2004). For the smallest firms, the model learns an upward shift in the order of 0.5 to 1.0 percent for forecast horizons of 5 to 60 trading days, respectively. We also see that the effect for medium and large firms is not

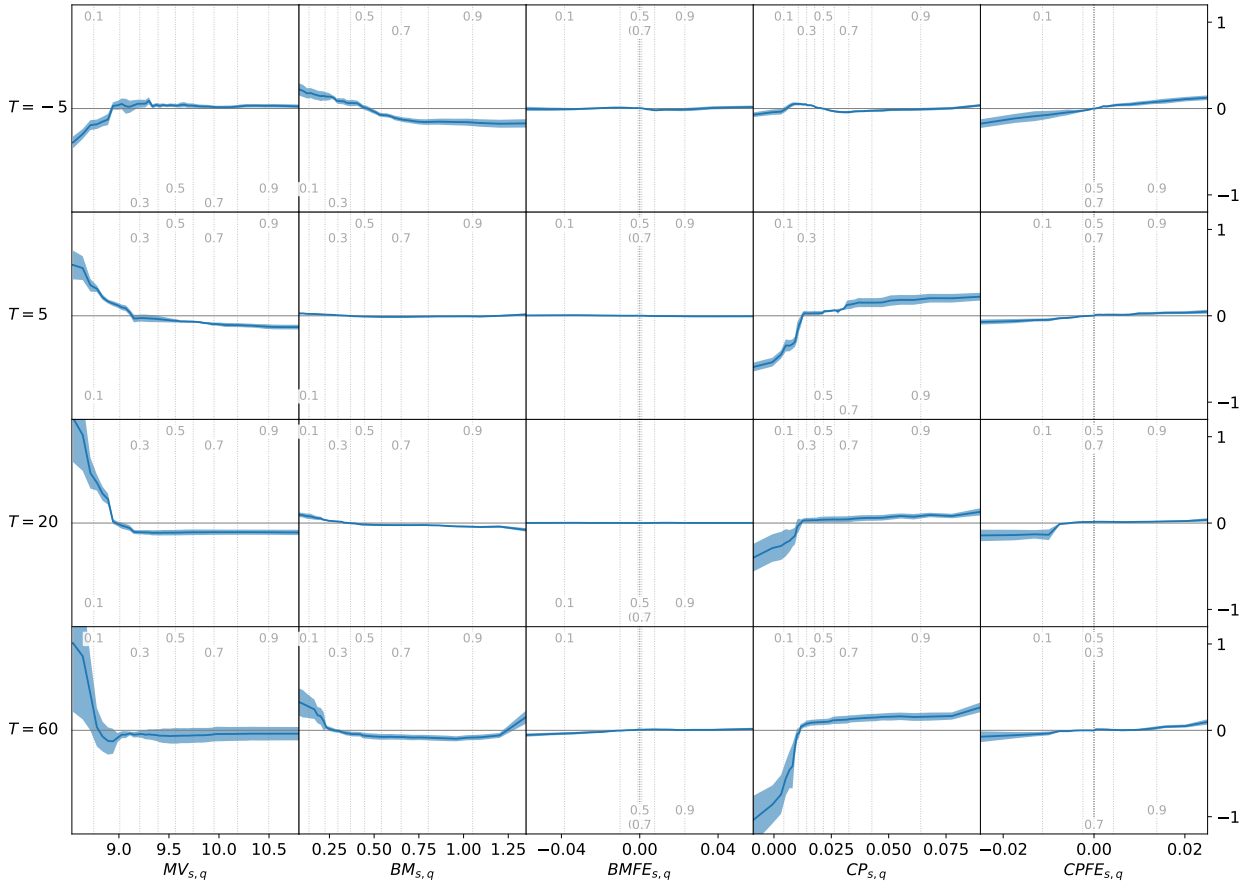

Figure 9: **Accumulated local effects plots for value-driving variables.** These plots show the accumulated local effects for call-related features for the contemporary return window ($T = -5$ days; top row) and three different drift periods ($T = 5, 20, 60$). Results are shown for the prepared remarks length ($I\text{-}LEN_{s,q}$), positivity of analysts in the first or second half of the Q&A session ($FA\text{-}P_{s,q}$, $SA\text{-}P_{s,q}$) and the environment-related negativity in the Q&A session ($Q\text{-}EN\text{-}N_{s,q}$). Accumulated local effects show the cumulative changes in the random forest's abnormal return prediction (vertical axis) when varying the indicated feature's value (horizontal axis). We plot the mean ALE curve from all model runs, and display the point-wise standard deviation between ALE curves from different runs as shaded bands to provide some indication of the curve's variation between models trained with different data sets. Dotted vertical lines depict quantiles of the respective feature's values.

distinct, which might be due to the fact the size-adjustment of abnormal returns adjusts for the effect in this range. Regarding the book-to-market ratio ($BM_{s,q}$), we see that the effect is not as prominent as the results of ex-post returns of value and growth stocks (Figure 3 in Section 4.2). The model's prediction is significantly affected only in the lower and upper deciles. For a forecast horizon of 60 days, we find a U-shaped pattern, i.e., stocks with extreme BM ratios tend to yield higher abnormal returns. This is in line with the finding of Baker and Wurgler (2006), who hypothesize that companies with extreme BM ratios are more difficult to value, which results in a risk premium. Interestingly, the effect of the book-to-market forecast error ($BMFE_{s,q}$) is fairly weak in terms of its accumulated local effect. Comparing this result to the variable's high feature importance (Figure 6), we note a stark contrast, which could be explained by the following: First,

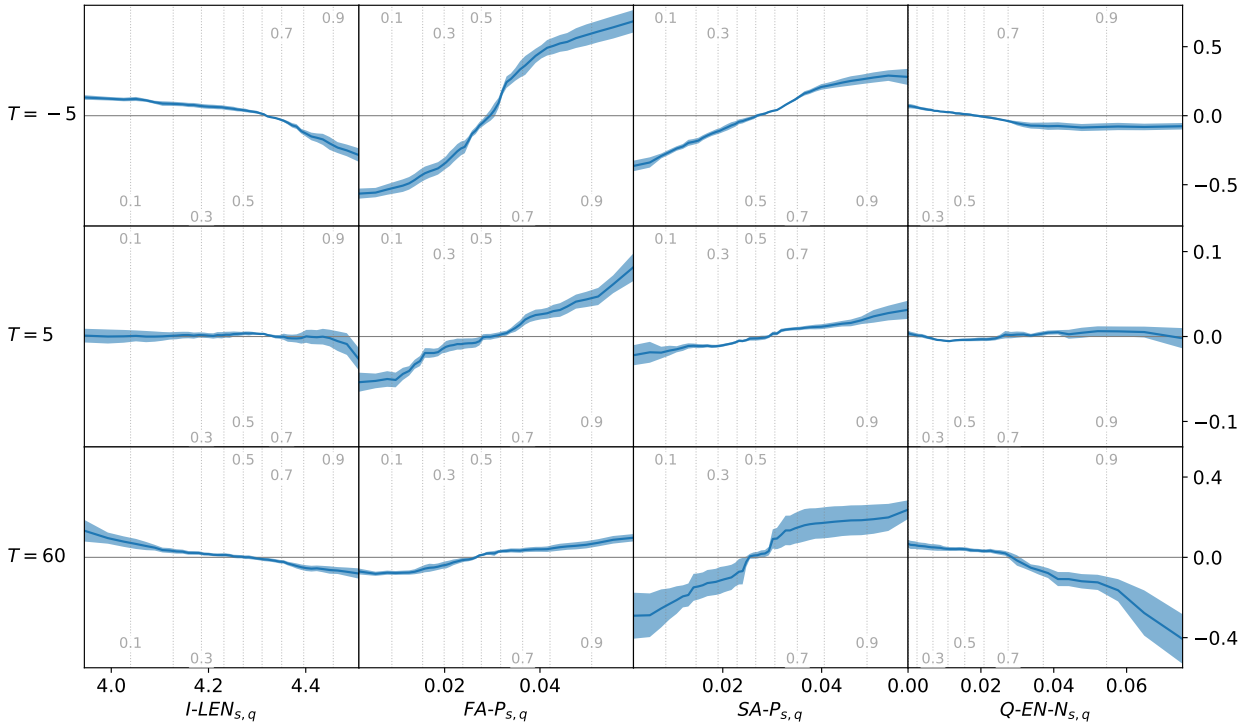

Figure 10: **Accumulated local effects plots for call-related features.** These plots show the accumulated local effects for call-related features for the contemporary return window ($T = -5$ days; top row) and three different drift periods ($T = 5, 20, 60$), in percent. Results are shown for the prepared remarks length ($I\text{-}LEN_{s,q}$), the positivity of analysts in the first or second half of the Q&A session ($FA\text{-}P_{s,q}$, $SA\text{-}P_{s,q}$) and the environment-related negativity in the Q&A session ($Q\text{-}EN\text{-}N_{s,q}$). Accumulated local effects show the cumulative changes in the random forest's abnormal return prediction (vertical axis) when varying the indicated feature's value (horizontal axis). We plot the mean ALE curve from all model runs, and display the point-wise standard deviation between ALE curves from different runs as shaded bands to provide some indication of the curve's variation between models trained with different data sets. Dotted vertical lines depict quantiles of the respective feature's values.

the random forest might preferably select the variable during training because of its very significant relation to abnormal returns, but at the same time low effect size. Second, the variable could gain in feature importance because of interactions with other variables. The effect of the cashflow-to-price ratio ($CP_{s,q}$) on the model's prediction is clearly visible for post-announcement return windows. Low cashflow-to-price ratios, i.e., in the first two deciles, decrease the model's abnormal return prediction in a range from about 0.4 to 1.0 percent. For cashflow-to-price rations above the second decile, the model learns a weakly positive relation to abnormal returns. This observation supports the argument of Wilson (1986) that cashflows from operations provide incremental information on post-earnings-announcement returns compared to earnings alone. In fact, we find that the model learns to predict positive abnormal returns for the largest earnings-to-price ratios, and negative returns for the lowest cashflow-to-price ratios. The return decrease due to the cashflow-to-price ratio and the return increase due to the earnings-to-price ratio have a similar magnitude.

Influence of call-related features: Figure 10 shows ALE plots for variables related to earnings

conference calls. The length of the prepared remarks section ($I\text{-}LEN_{s,q}$) shows a negative trend in the top three deciles. We hypothesize that prepared remarks of above-average length might often deal with rather negative business topics. The two variables capturing analyst positivity ($FA\text{-}P_{s,q}$ and $SA\text{-}P_{s,q}$) influence the model's prediction as expected, as a higher level of analyst positivity results in a more positive abnormal return prediction. Regarding positivity measures for analyst statements in the first half of the Q&A session ($FA\text{-}P_{s,q}$), we see the largest effect in the contemporary window, where model's prediction varies over a range of one percent, and exhibits an S-curved shape similar to the one for the earnings forecast error (compare Figure 7). By comparison, the post-announcement behavior is linear and ranges between -5 bp and 10 bp. The second-half analyst positivity ($SA\text{-}P_{s,q}$) exhibits a weaker contemporary effect when compared to the effect of first half analyst positivity. On the other hand, the effect size in the 60 days window is comparatively larger. A possible explanation might be that information from the Q&A session differs in nature and that clearly positive aspects of business performance are more likely to be highlighted at the beginning of each session. The effect on predicted return of the topic-specific negativity related to the business environment ($Q\text{-}EN\text{-}N_{s,q}$) increases with time horizon from close to zero to about -0.4 percent for the top three deciles. This indicates that frequent negative statements regarding business conditions in Q&A sessions are related to negative abnormal returns by the model, yet this effect is most pronounced for long forecast horizons. Similarly, and in line with expectation, we find positive (negative) effects of the remaining positivity (negativity) variables on predicted abnormal returns.

Revisiting the gradual information diffusion interpretation: Following our interpretation in terms of the gradual information diffusion hypothesis of Hong and Stein (1999) from Section 4.3, we may interpret this section's results as a delayed adjustment to updated information on a stock's fundamental value. Differences in predictive ability between variables may be caused by differences in the speed of information diffusion, which could for example occur due to the variability of the cost of information retrieval and processing (Engelberg, 2008).

The adjustment pertaining to the correction of market expectations, expressed in terms of forecast errors, appears to occur relatively fast. We observe the largest effect of earnings and revenue surprises for the contemporary time window, and a relatively small effect in the post-event period. As information on errors in market expectations are relatively easy to obtain – and therefore cheap – it is not surprising to see prices react quickly to this kind of information component.[22] Consequently, our model's predictions for contemporary return windows are primarily based on

[22]These variables are commonly the first present in a wide range of media sources and databases shortly after earnings announcements, which likely reduces information retrieval and processing costs.

	Before transaction costs				After transaction costs			
	ES	**LR**	**LR-B**	**RF**	**ES**	**LR**	**LR-B**	**RF**
Panel A: Daily return statistics								
Mean return	0.00023	0.00055	0.00056	0.00064	-0.00017	0.00035	0.00036	0.00045
Standard error	0.00013	0.00011	0.00011	0.00011	0.00018	0.00014	0.00013	0.00013
t-statistic	1.74783	4.81061	4.99277	5.81949	-0.94912	2.52641	2.71319	3.46725
Minimum	-0.04812	-0.03060	-0.03490	-0.03416	-0.07469	-0.03676	-0.04261	-0.04014
First quartile	-0.00183	-0.00133	-0.00115	-0.00123	-0.00291	-0.00199	-0.00169	-0.00183
Median	0.00000	0.00014	0.00033	0.00026	0.00000	0.00000	0.00009	0.00002
Third quartile	0.00253	0.00219	0.00214	0.00228	0.00293	0.00242	0.00227	0.00241
Maximum	0.03348	0.04423	0.04534	0.04911	0.05552	0.05608	0.05561	0.06064
Share ≥ 0	0.57416	0.58943	0.59198	0.59898	0.54297	0.55506	0.55570	0.56906
Standard dev.	0.00522	0.00455	0.00449	0.00435	0.00704	0.00554	0.00532	0.00515
Skewness	-0.89308	0.78373	0.93425	1.25281	-1.22650	0.94138	1.11005	1.40553
Kurtosis	10.70347	11.64969	20.12479	19.79053	17.07616	12.49011	20.40821	21.39488
Panel B: Risk characteristics								
1-percent VaR	-0.01552	-0.01129	-0.01141	-0.01028	-0.02219	-0.01454	-0.01439	-0.01257
1-percent CVaR	-0.00755	-0.00576	-0.00525	-0.00527	-0.01052	-0.00750	-0.00677	-0.00652
5-percent VaR	-0.02312	-0.01613	-0.01738	-0.01513	-0.03117	-0.01964	-0.02026	-0.01813
5-percent CVaR	-0.01309	-0.00947	-0.00938	-0.00839	-0.01821	-0.01184	-0.01153	-0.01043
Max. drawdown	-0.18389	-0.07744	-0.06773	-0.05445	-0.39082	-0.11683	-0.07982	-0.07745
Panel C: Annualized risk-return metrics								
Return	0.05610	0.14620	0.15006	0.17163	-0.04761	0.08879	0.09214	0.11634
Volatility	0.08293	0.07221	0.07122	0.06902	0.11181	0.08792	0.08442	0.08168
Sharpe ratio	0.69980	1.92608	1.99901	2.33001	-0.38001	1.01153	1.08631	1.38822
Sortino ratio	0.96681	3.16623	3.25941	4.03399	-0.49687	1.60005	1.71009	2.28816

Table 8: **Daily and annualized risk-return metrics.** This table provides summary statistics on the performance of the earnings-event-based trading strategy. We depict results for the earnings surprise model, the standard (LR) and binned linear regression (LR-B) and the random forest (RF), both before and after transaction costs. Panel A depicts daily return statistics. Panel B exhibits risk metrics, and Panel C shows annualized risk-return metrics.

earnings and revenue surprises. By contrast, the post-announcement adjustment can be related to a gradual diffusion of updated value-growth characteristics: Most of the valuation ratios that we find to substantially drive our model's post-announcement predictions are commonly used to differentiate between value and growth stocks, i.e., book-to-market, earnings-to-price or cashflow-to-price ratios. Similar to the long-term value premium (see, for example, Fama and French, 1989; Desai et al., 2004), the accumulated local effects of these features show that the model predicts higher abnormal returns for value stocks. Similar to our observation from feature importances, variables with potentially higher information processing costs, for example textual sentiment, tend to increasingly impact the model's prediction for longer forecast horizons.

4.5. Financial performance of an announcement-based trading strategy

We conclude with results on the financial performance of our zero-investment announcement-based trading strategy to assess the economic significance of our model's predictions. Table 8 reports daily and annualized risk-return metrics, both before and after trading costs, for the random forest

model, the simple and binned linear models as well as the earnings surprise model. Figure 11 depicts the evolution of cumulative profits for all models after transactions costs.

Return characteristics: We find that the trading strategy based on the random forest exhibits the highest average daily return, with 6.4 bp before and 4.5 bp after transaction costs (Panel A of Table 8). With Newey-West t-statistics of 5.82 and 3.48 (compared to a critical value of 1.96 at the 5 percent level), average daily returns are statistically significant for the null hypothesis of a zero mean return (Newey and West, 1987). The distribution of daily returns is skewed to the right and leptokurtic, which is caused by large outliers also visible in the minimum and maximum statistics. In comparison, the strategies based on the linear and binned linear models exhibit lower daily returns. With mean returns at 3.5 bp and 3.6 bp after transaction costs, we find these models to exhibit a similar performance. Again, these returns are statistically different from zero. When comparing these average daily returns to abnormal event returns for a forecast horizon of 5 days (Panel B of Table 4), this finding surprises: In terms of the average abnormal return following earnings events, the binned linear model outperforms the linear regression (87.4 bp compared to 64.3 bp). As the number of top-flop earnings events selected by these models is similar (6516 and 6525), we conjecture that the relative attenuation of daily return is due to an unfavorable distribution of earnings events onto trading days.[23] By contrast, we find that the earnings surprise model performs considerably worse than the other models, as the small average daily return before costs (2.3 bp) is insufficient for positive returns after transaction costs (-1.7 bp). The development of the portfolio value depicted in Figure 11 illustrates that the strategy accumulates returns quite uniformly over time. Further analysis of the apparent jump in portfolio value at the beginning of March, 2016 shows that a series of high-return announcements, closely clustered in time, gives rise to a sequence of very successful trading days.[24] It is noticeable that the jump occurs only with models which employ the full set of features, and is not apparent for the earnings surprise model.

Risk characteristics: Panel B of Table 8 shows that the strategy based on the random forest model has favorable risk characteristics. With a historical one-percent value at risk (1-percent VaR) of 1.267 percent, we find that the random forest exhibits the lowest daily risk. Risk characteristics of the simple linear and the binned linear model are similar (1-percent VaR of 1.454 and 1.439 percent). In comparison, the earnings surprise model has a nearly twice as high 1-percent VaR (2.219). Results for the 5-percent VaR and the conditional value at risk (CVaR) lead to similar conclusions. The random forest also exhibits the lowest maximum drawdown (-7.745 percent). By

[23] For a further discussion on a timing-related attenuation of performance in event-based trading strategies, we refer the reader to Schnaubelt et al. (2020).

[24] To exclude data errors, we have validated the returns of the respective stocks with two independent sources.

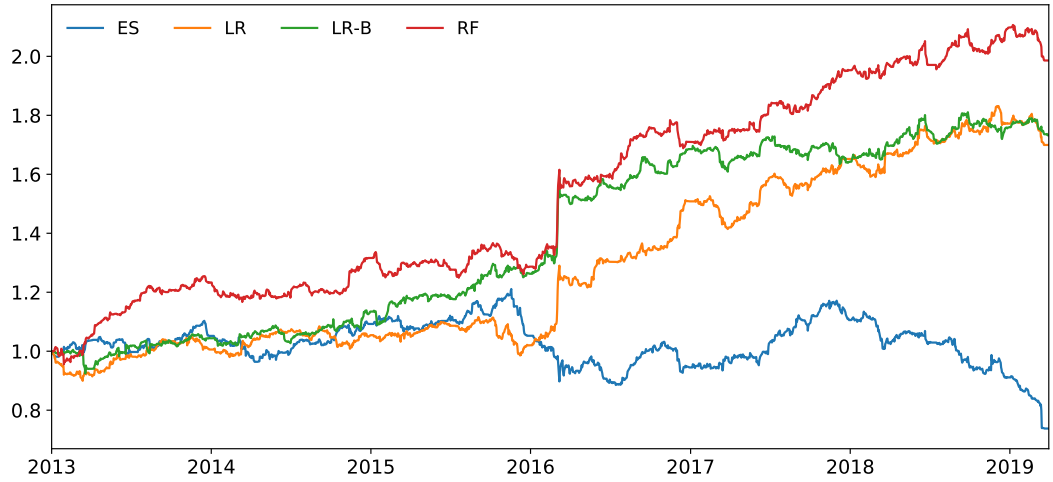

Figure 11: **Development of portfolio value for the earnings-based trading strategy.** This figure depicts the value of a portfolio allocated according to the earnings announcement-based trading strategy, after transaction costs, for the earnings surprise (ES), the linear regression (LR), the binned linear regression (LR-B) and the random forest model (RF).

contrast, the linear model (-11.683 percent) and the binned linear model (-7.982 percent) show higher levels of drawdown risk.

Annualized risk-return characteristics: Annualized risk-return metrics are summarized in Panel C of Table 8. With annualized returns of 15.001 and 11.634 percent before and after transaction costs, the random forest performs significantly better than the other models. The second-best binned linear model achieves an annualized mean return of 9.214 percent after transaction costs. The better risk characteristics and higher mean returns of the random forest model are also reflected in the annualized risk-return metrics. The Sharpe ratio for the random forest is at 2.33 and 1.39 before and after transaction costs, which exceeds the values of the binned linear model (2.00 and 1.09, respectively).

5. Conclusion

Earnings announcements are the first source of new information on a firm's quarterly financial performance, and, as a major capital market event, receive elevated attention from both investors and executives. In this paper, we apply state-of-the-art financial machine learning to assess the return-predictive value of information conveyed in earnings announcements. Our empirical analysis is based on I/B/E/S analyst estimates, corresponding actual values and conference call transcripts for more than 45,000 earnings announcements from 2007 to 2019 on a majority of S&P1500 constituents.

We contribute to the existing literature in the following ways: First, we describe how machine learning can be applied to abnormal return prediction based on the various information components

relayed in earnings announcements. Leveraging domain knowledge and existing literature, we identify four categories of potentially informative variables, i.e., forecast errors, valuation ratios, proxies for information quality and quantity as well as topic- and speaker-specific sentiment polarity. We then apply random forests and several benchmark models to predict abnormal returns for different forecast horizons. Finally, earnings announcements are assigned to decile portfolios based on the rolling rank of the model's prediction.

Second, we assess the model's out-of-sample performance and learning. We find that the random forest exhibits a return-predictive performance that is superior to our benchmark models. Similar to the well-known post-earnings-announcement drift, we find that post-announcement abnormal returns increase with the forecast horizon. Specifically, mean abnormal return in the top or flop decile increase from 95.4 bp for a forecast horizon of 5 trading days to 193.6 bp for 60 days. We perform several in-depth analyses on the superior performance of the random forest. In a first analysis, we find larger abnormal returns for small firms and a delayed abnormal return drift for growth stocks. A second set of analyses inspects the contribution of variables. While forecast errors pertaining to earnings and revenue are the main predictors for contemporary returns, we find that a larger number of variables, mostly valuation ratios and forecast errors, is used to predict post-announcement abnormal returns. We find that valuation ratios alone contribute the largest part of the observed abnormal post-announcement returns. The third analysis leverages accumulated local effects plots to analyze the influence of individual variables on the random forest's prediction. We find that the model recovers non-linear patterns in several common capital market effects such as the value premium. Our findings are consistent with the hypothesis that differences in the speed of information diffusion entail differences in predictive ability between variables. Such differences in diffusion speed may be related to variable costs of information retrieval and processing.

Third, we assess the economic significance of the model's predictions in an announcement-based zero-investment trading strategy. The underlying trading simulation integrates several real-world constraints such as transaction costs and limited capital leverage. Despite of such conservative constraints, we find the trading strategy to yield statistically significant annualized returns of 11.63 percent at a Sharpe ratio of 1.39.

Overall, we have, to our knowledge, presented the first large-scale empirical study of financial machine learning in the context of earnings announcements, thereby covering the wealth of complex information in analyst expectations, earnings press releases and earnings conference calls. We have successfully demonstrated that the model leverages several well-known capital market effects to achieve superior financial performance and may offer new perspectives on the diffusion of valuation-related information from earnings announcements.

References

Aharony, J., Swary, I., 1980. Quarterly dividend and earnings announcements and stockholders' returns: An empirical analysis. The Journal of Finance 35, 1–12.

Apley, D.W., Zhu, J., 2019. Visualizing the effects of predictor variables in black box supervised learning models. arXiv 1612.08468.

Avellaneda, M., Lee, J.H., 2010. Statistical arbitrage in the US equities market. Quantitative Finance 10, 761–782.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The Journal of Finance 61, 1645–1680.

Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. Journal of Accounting and Economics 50, 20–41.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159.

Banz, R.W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9, 3–18.

Barbee, W.C., Jeong, J.G., Mukherji, S., 2008. Relations between portfolio returns and market multiples. Global Finance Journal 19, 1–10.

Barbee, W.C.J., Mukherji, S., Raines, G.A., 1996. Do sales–price and debt–equity explain stock returns better than book–market and firm size? Financial Analysts Journal 52, 56–60.

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. The Journal of Finance 32, 663.

Basu, S., 1983. The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. Journal of Financial Economics 12, 129–156.

Basu, S., Duong, T.X., Markov, S., Tan, E.J., 2013. How important are earnings announcements as an information source? European Accounting Review 22, 221–256.

Battalio, R.H., Mendenhall, R.R., 2011. Post-earnings announcement drift: Bounds on profitability for the marginal investor. Financial Review 46, 513–539.

Beaver, W.H., 1968. The information content of annual earnings announcements. Journal of Accounting Research 6, 67–92.

Beaver, W.H., McNichols, M.F., Wang, Z.Z., 2018. The information content of earnings announce-

ments: New insights from intertemporal and cross-sectional behavior. Review of Accounting Studies 23, 95–135.

Bernard, V., Thomas, J., Wahlen, J., 1997. Accounting-based stock price anomalies: Separating market inefficiencies from risk. Contemporary Accounting Research 14, 89–136.

Bernard, V.L., Stober, T.L., 1989. The nature and amount of information in cash flows and accruals. The Accounting Review 64, 624–652.

Bernard, V.L., Thomas, J.K., 1989. Post-earnings-announcement drift: Delayed price response or risk premium? Journal of Accounting Research 27, 1–36.

Bernard, V.L., Thomas, J.K., 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics 13, 305–340.

BlackRock Inc., 2020. iShares Core S&P Total U.S. Stock Market ETF. URL: https://www.ishares.com/us/products/239724/ishares-core-sp-total-us-stock-market-etf.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

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

Brockman, P., Li, X., Price, S.M., 2015. Differences in conference call tones: Managers vs. analysts. Financial Analysts Journal 71, 24–42.

Brown, L.D., Rozeff, M.S., 1978. The superiority of analyst forecasts as measures of expectations: Evidence from earnings. The Journal of Finance 33, 1–16.

Campbell, J.Y., Polk, C., Vuolteenaho, T., 2010. Growth or glamour? Fundamentals and systematic risk in stock returns. The Review of Financial Studies 23, 305–344.

Campbell, J.Y., Ramadorai, T., Schwartz, A., 2009. Caught on tape: Institutional trading, stock returns, and earnings announcements. Journal of Financial Economics 92, 66–91.

Cen, L., Chen, J., Dasgupta, S., Ragunathan, V., 2020. Do analysts and their employers value access to management? Evidence from earnings conference call participation. Journal of Financial and Quantitative Analysis (forthcoming).

Center for Research in Security Prices, 2018. Data descriptions guide – CRSP US stock & US index databases. URL: http://www.crsp.com/products/documentation/crsp-us-stock-and-index-databases-data-descriptions-guide.

Chan, L.K.C., Hamao, Y., Lakonishok, J., 1991. Fundamentals and stock returns in Japan. The

Journal of Finance 46, 1739–1764.

Chan, L.K.C., Jegadeesh, N., Lakonishok, J., 1995. Evaluating the performance of value versus glamour stocks: The impact of selection bias. Journal of Financial Economics 38, 269–296.

Chan, L.K.C., Lakonishok, J., 2004. Value and growth investing: Review and update. Financial Analysts Journal 60, 71–86.

Chen, H.Y., Chen, S.S., Hsin, C.W., Lee, C.F., 2014. Does revenue momentum drive or ride earnings or price momentum? Journal of Banking & Finance 38, 166–185.

Cheng, C.S.A., Hopwood, W.S., McKeown, J.C., 1992. Non-linearity and specification problems in unexpected earnings response regression model. The Accounting Review 67, 579–598.

Collins, D.W., Li, O.Z., Xie, H., 2009. What drives the increased informativeness of earnings announcements over time? Review of Accounting Studies 14, 1–30.

Daniel, K., Titman, S., 1997. Evidence on the characteristics of cross sectional variation in stock returns. The Journal of Finance 52, 1–33.

Davis, A.K., Ge, W., Matsumoto, D., Zhang, J.L., 2015. The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies 20, 639–673.

Davis, A.K., Piger, J.M., Sedor, L.M., 2012. Beyond the numbers: Measuring the information content of earnings press release language. Contemporary Accounting Research 29, 845–868.

Dellavigna, S., Pollet, J.M., 2009. Investor inattention and friday earnings announcements. The Journal of Finance 64, 709–749.

Demers, E., Vega, C., 2008. Soft information in earnings announcements: News or noise? Board of Governors of the Federal Reserve System, International Finance Discussion Papers 951.

Demers, E.A., Vega, C., 2010. Soft information in earnings announcements: News or noise? INSEAD Working Paper 2010/33/AC. Fontainebleau, France. URL: http://www.ssrn.com/abstract=1153450.

Desai, H., Rajgopal, S., Venkatachalam, M., 2004. Value-glamour and accruals mispricing: One anomaly or two? The Accounting Review 79, 355–385.

Diether, K.B., Malloy, C.J., Scherbina, A., 2002. Differences of opinion and the cross section of stock returns. The Journal of Finance 57, 2113–2141.

Doyle, J.T., Lundholm, R.J., Soliman, M.T., 2006. The extreme future stock returns following I/B/E/S earnings surprises. Journal of Accounting Research 44, 849–887.

Engelberg, J., 2008. Costly information processing: Evidence from earnings announcements. SSRN Scholarly Paper ID 1107998. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=1107998.

Ertimur, Y., Livnat, J., Martikainen, M., 2003. Differential market reactions to revenue and expense surprises. Review of Accounting Studies 8, 185–211.

Fama, E.F., French, K.R., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25, 23–49.

Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. The Journal of Finance 47, 427–465.

Fama, E.F., French, K.R., 1998. Value versus growth: The international evidence. The Journal of Finance 53, 1975–1999.

Fischer, T.G., Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270, 654–669.

Foster, G., Olsen, C., Shevlin, T., 1984. Earnings releases, anomalies, and the behavior of security returns. The Accounting Review 59, 574–603.

Freeman, R.N., Tse, S.Y., 1992. A nonlinear model of security price responses to unexpected earnings. Journal of Accounting Research 30, 185–209.

Fried, D., Givoly, D., 1982. Financial analysts' forecasts of earnings: A better surrogate for market expectations. Journal of Accounting and Economics 4, 85–107.

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

Garfinkel, J.A., Sokobin, J., 2006. Volume, opinion divergence, and returns: A study of post-earnings announcement drift. Journal of Accounting Research 44, 85–112.

Givoly, D., Lakonishok, J., 1979. The information content of financial analysts' forecasts of earnings: Some evidence on semi-strong inefficiency. Journal of Accounting and Economics 1, 165–185.

Gleason, C.A., Lee, C.M.C., 2003. Analyst forecast revisions and market price discovery. The Accounting Review 78, 193–225.

Graham, J.R., Harvey, C.R., Rajgopal, S., 2005. The economic implications of corporate financial reporting. Journal of Accounting and Economics 40, 3–73.

de Groot, W., Huij, J., Zhou, W., 2012. Another look at trading costs and short-term reversal

profits. Journal of Banking & Finance 36, 371–382.

Gu, S., Kelly, B., Xiu, D., 2018. Empirical asset pricing via machine learning. Working Paper 25398. National Bureau of Economic Research. URL: http://www.nber.org/papers/w25398.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics. 2nd ed., Springer-Verlag, New York.

Hawkins, E.H., Chamberlin, S.C., Daniel, W.E., 1984. Earnings expectations and security prices. Financial Analysts Journal 40, 24–74.

Henry, E., 2006. Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. Journal of Emerging Technologies in Accounting 3, 1–19.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? The Journal of Business Communication 45, 363–407.

Henry, E., Leone, A.J., 2015. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. The Accounting Review 91, 153–178.

Hirshleifer, D., 2001. Investor psychology and asset pricing. The Journal of Finance 56, 1533–1597.

Hirshleifer, D., Lim, S.S., Teoh, S.H., 2009. Driven to distraction: Extraneous events and underreaction to earnings news. The Journal of Finance 64, 2289–2325.

Hong, H., Lim, T., Stein, J.C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. The Journal of Finance 55, 265–295.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. The Journal of Finance 54, 2143–2184.

Hong, H., Stein, J.C., 2007. Disagreement and the stock market. Journal of Economic Perspectives 21, 109–128.

Jacobs, H., 2015. What explains the dynamics of 100 anomalies? Journal of Banking & Finance 57, 65–85.

Jegadeesh, N., Livnat, J., 2006a. Post-earnings-announcement drift: The role of revenue surprises. Financial Analysts Journal 62, 22–34.

Jegadeesh, N., Livnat, J., 2006b. Revenue surprises and stock returns. Journal of Accounting and Economics 41, 147–171.

Jha, V., 2016. Timing equity quant positions with short-horizon alphas. The Journal of Trading

11, 53–59.

Ji, Y., Rozenbaum, O., 2018. Analysts' suspicions of earnings management and conference call narratives. Working Paper. URL: http://scholarspace.manoa.hawaii.edu/handle/10125/59320.

Kallapur, S., 1994. Dividend payout ratios as determinants of earnings response coefficients: A test of the free cash flow theory. Journal of Accounting and Economics 17, 359–375.

Kausar, A., 2017. Post-earnings-announcement drift and the return predictability of earnings levels: One effect or two? Management Science 64, 4877–4892.

Ke, Z.T., Kelly, B., Xiu, D., 2019. Predicting returns with text data. University of Chicago, Becker Friedman Institute for Economics Working Paper 2019-69.

Kearney, C., Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis 33, 171–185.

Keim, D.B., 1985. Dividend yields and stock returns: Implications of abnormal January returns. Journal of Financial Economics 14, 473–489.

Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2011. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 41, 552–568.

Kothari, S.P., 2001. Capital markets research in accounting. Journal of Accounting and Economics 31, 105–231.

Kothari, S.P., Wasley, C., 2019. Commemorating the 50-year anniversary of Ball and Brown (1968): The evolution of capital market research over the past 50 years. Journal of Accounting Research 57, 1117–1159.

Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research 259, 689–702.

Lakonishok, J., Shleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation, and risk. The Journal of Finance 49, 1541–1578.

Landsman, W.R., Maydew, E.L., 2002. Has the information content of quarterly earnings announcements declined in the past three decades? Journal of Accounting Research 40, 797–808.

Lee, C.M.C., 1992. Earnings news and small traders: An intraday analysis. Journal of Accounting and Economics 15, 265–302.

Lewellen, J., 2004. Predicting returns with financial ratios. Journal of Financial Economics 74, 209–235.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247.

Litzenberger, R.H., Ramaswamy, K., 1979. The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. Journal of Financial Economics 7, 163–195.

Livnat, J., Mendenhall, R.R., 2006. Comparing the post-earnings announcement drift for surprises calculated from analyst and time series forecasts. Journal of Accounting Research 44, 177–205.

Livnat, J., Santicchia, M., 2006. Cash flows, accruals, and future returns. Financial Analysts Journal 62, 48–61.

Livnat, J., Zarowin, P., 1990. The incremental information content of cash-flow components. Journal of Accounting and Economics 13, 25–46.

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.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54, 1187–1230.

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.

Mayew, W.J., 2008. Evidence of management discrimination among analysts during earnings conference calls. Journal of Accounting Research 46, 627–659.

Melendrez, K.D., Schwartz, W.C., Trombley, M.A., 2008. Cash flow and accrual surprises: Persistence and return implications. Journal of Accounting, Auditing & Finance 23, 573–592.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv 1301.3781.

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

Patell, J.M., Wolfson, M.A., 1984. The intraday speed of adjustment of stock prices to earnings and dividend announcements. Journal of Financial Economics 13, 223–252.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,

Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pesaran, M.H., Timmermann, A., 1995. Predictability of stock returns: Robustness and economic significance. The Journal of Finance 50, 1201–1228.

Pettit, R.R., 1972. Dividend announcements, security performance, and capital market efficiency. The Journal of Finance 27, 993–1007.

Porta, R.L., 1996. Expectations and the cross-section of stock returns. The Journal of Finance 51, 1715.

Porta, R.L., Lakonishok, J., Shleifer, A., Vishny, R., 1997. Good news for value stocks: Further evidence on market efficiency. The Journal of Finance 52, 859–874.

Pratt, J.W., 1959. Remarks on zeros and ties in the Wilcoxon signed rank procedures. Journal of the American Statistical Association 54, 655–667.

Price, S.M., Doran, J.S., Peterson, D.R., Bliss, B.A., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance 36, 992–1011.

Price, S.M., Salas, J.M., Sirmans, C.F., 2015. Governance, conference calls and CEO compensation. The Journal of Real Estate Finance and Economics 50, 181–206.

Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. The Journal of Portfolio Management 11, 9–16.

Sadique, S., In, F.H., Veeraraghavan, M., 2008. The impact of spin and tone on stock returns and volatility: Evidence from firm-issued earnings announcements and the related press coverage. SSRN Scholarly Paper ID 1121231. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=1121231.

Sak, H., Huang, T., Chng, M., 2018. Exploring the factor zoo with a machine-learning portfolio. SSRN Scholarly Paper ID 3202277. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=3202277.

Schnaubelt, M., 2019. A comparison of machine learning model validation schemes for non-stationary time series data. FAU Discussion Papers in Economics No. 11/2019.

Schnaubelt, M., Fischer, T.G., Krauss, C., 2020. Separating the signal from the noise – financial machine learning for Twitter. Journal of Economic Dynamics and Control 114, 103895.

S&P Dow Jones Indices, 2019. S&P U.S. Indices Methodology. Technical Report. S&P Dow Jones Indices. URL: https://us.spindices.com/indices/equity/sp-composite-1500.

Stattman, D., 1980. Book values and stock returns. The Chicago MBA: A journal of selected papers 4, 25–45.

Stickel, S.E., 1991. Common stock returns surrounding earnings forecast revisions: More puzzling evidence. The Accounting Review 66, 402–416.

Swaminathan, S., Weintrop, J., 1991. The information content of earnings, revenues, and expenses. Journal of Accounting Research 29, 418–427.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting 16, 437–450.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance 63, 1437–1467.

Thomson Reuters, 2018. Methodology for estimates – a guide to understanding Thomson Reuters methodologies, terms and policies for I/B/E/S estimates databases. Technical Report.

Vassalou, M., Xing, Y., 2004. Default risk in equity returns. The Journal of Finance 59, 831–868.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

Wilson, G.P., 1986. The relative information content of accruals and cash flows: Combined evidence at the earnings announcement and annual report release date. Journal of Accounting Research 24, 165–200.

Woolridge, J.R., 1983. Ex-date stock price adjustment to stock dividends: A note. The Journal of Finance 38, 247–255.

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

Appendix A. Computation of topic-specific sentiment variables

To calculate topic-specific sentiment variables for a given transcript, we proceed in three steps. In the first step, we compute topic classification scores to assign a transcript's sentences to one of six topics based on topic-specific dictionaries.[25] Our selection of topics and topic-specific key words is based on the literature and on results of a topic clustering algorithm. We use the Latent Dirichlet allocation (LDA) algorithm introduced by Blei et al. (2003) to cluster textual content of earnings conference call transcripts based on our first training set, i.e., for earnings calls of the years 2007 up to and including 2012. Subsequently, we select three types of terms for our six core word lists from these content clusters. Terms that are related to accounting figures, others which express expectations of future developments and yet others which suggest changes or movements (expressed with directional terms) following Henry (2006), Tetlock et al. (2008) and Engelberg (2008). We use Gensim's word2vec module by Mikolov et al. (2013) to represents word tokens of the overall text corpus as word vectors in vector space. Thus, for each term of the core word lists, we add three further terms that can be found in local proximity (measured by the cosine-similarity of two word vectors) within the vector space model, to each word lists. In this way we want to counteract the loss of context, which is a bottleneck of dictionary based approaches for text analysis (Loughran and McDonald, 2016). We then compute an inverse-document-frequency factor IDF_w to weight terms in the word lists as the logarithm of the total number of sentences divided by the number of sentences containing term w. The terms and weights of our resulting word lists for the topic categories *Earnings*, *Revenues*, *Liquidity*, *Environment*, *Outlook* and *Changes* are consolidated in Table A.9. We calculate the topic classification score $TS_{d,k}$ for a given sentence (document d) and topic k by summing up IDF factors and normalizing with the number of tokens in the sentence, i.e.,

$$TS_{d,k} = \frac{1}{|W_d|} \sum_{w \in W_d} IDF_w, \tag{A.1}$$

where W_d is the list of tokens in the sentence, and $|W_d|$ denotes its length.

In the second step, we calculate sentiment polarity scores with a modified version of the Generic Parser program and the finance-specific sentiment dictionary of Loughran and McDonald (2011) in the updated version of 2018.[26] For every sentence d, we calculate positive and negative sentiment polarities $S_d^{p,n}$ by dividing the number of keyword matches from the respective word lists by the

[25] We apply a dictionary-based approach to extract the topic-specific sentiment of earnings conference call transcripts to facilitate the replication of results (Loughran and McDonald, 2016). We also believe that this will allow us to transparently present our process of thematically categorizing the information-content.

[26] The Generic Parser program and the L&M Master Dictionary can be found under https://sraf.nd.edu/textual-analysis/code/.

total number of tokens in the sentence.

In the third step, we compute topic-specific sentiment polarities as the average document sentiment polarity weighted by topic classification score, i.e.,

$$TSS_k^{p,n} = \frac{\sum_{d=1}^{D} TS_{d,k} \cdot S_d^{p,n}}{\sum_{d=1}^{D} TS_{d,k}}, \tag{A.2}$$

where D denotes the number of sentences in the transcript.

The following example illustrates how statements in earnings call transcripts are preprocessed, classified as specific topics and scored regarding their sentiment polarity. This statement was made in the prepared remarks session of Apples's fourth quarter earnings call 2018 by Timothy Donald Cook:

> "This year, we shipped our 2 billionth iOS device, celebrated the 10th anniversary of the App Store, and achieved the strongest revenue and earnings in Apple's history."

The following word tokens are created after the removal of stopwords and bi-grams construction:

> year, shipped billionth, io device, celebrated, anniversary, app store, achieved, strongest, revenue, earnings, apple, history

In this statement two non-zero topics scores are computed, *Earnings* 0.3733 and *Revenues* 0.4276. The computed sentiment polarity is *Positivity* 0.1538 and *Negativity* 0.000.

Earnings		**Revenues**		**Liquidity**		**Environment**		**Outlook**		**Changes**	
earnings	1.079	revenue	0.847	cash	1.284	market	0.862	expect	1.031	increase	0.961957
income	1.219	sale	0.915	cash_flow	1.568	industry	1.467	expected	1.336	increased	1.075717
margin	1.231	volume	1.411	free_cash	1.894	demand	1.472	future	1.352	higher	1.198822
loss	1.505	sequentially	1.899	working_capital	2.094	environment	1.660	guidance	1.375	lower	1.238356
gross_margin	1.652	sequential	2.013	distributable_cash	3.057	supply	1.977	long_term	1.514	significant	1.300992
ebitda	1.724	subscription	2.412	cashflow	4.693	economic	2.031	expectation	1.524	improvement	1.383697
profit	1.725	consumable	3.316	free_cashflow	5.082	economy	2.140	outlook	1.631	change	1.438271
adjusted_ebitda	1.738			outspend_cash	6.830	marketplace	2.175	anticipate	1.805	decline	1.509389
profitability	1.816			freecash_flow	6.830	economic_environment	2.773	estimate	1.828	improved	1.523814
eps	1.942			cashflow_operation	6.830	recession	2.796	going_forward	1.901	reduction	1.580934
diluted_share	2.035					climate	3.052	forecast	2.039	improve	1.628478
gross_profit	2.054					global_economy	3.099	estimated	2.113	decrease	1.642499
diluted	2.244					macro_environment	3.198	assumption	2.154	decreased	1.74284
continuing_operation	2.367					economic_recovery	3.280	likely	2.157	reduced	1.769513
adjusted_eps	2.494					macroeconomic_environment	3.307	projected	2.180	declined	1.809513
ebit	2.513					economic_climate	3.558	projection	2.228	improving	1.850733
fy	2.578					new_entrant	3.726	longer_term	2.370	favorable	1.873912
ffo	2.614					economic_situation	6.830	expecting	2.437	meaningful	2.110179
diluted_eps	2.764							assuming	2.495	substantial	2.127414
earning	2.781							forecasted	2.607	reducing	2.141645
roe	2.936							forecasting	2.687	rose	2.377796
affo	3.120							projecting	2.861	drop	2.392795
oibda	3.366							anticipating	2.934	changed	2.449554
noninterest_income	3.537							updated_guidance	3.089	declining	2.514022
ebitdar	3.624							outlook_remainder	3.139	movement	2.533472
ebita	3.721							estimating	3.238	lowered	2.793832
roa	3.734							revised_guidance	3.277	decreasing	2.908651
fad	3.976							planning_assumption	3.891	sizable	3.041518
cffo	4.089							ourexpectations	4.785	variation	3.064661
adjusted_ebitdar	4.108							previous_guidance	6.830	slight_decline	3.31683
ebidta	4.141							revised_outlook	6.830	slight_decrease	3.332008
adjusted_ebidta	4.260									higher_proportion	3.558572
decremental_margin	4.350									lesser	3.605648
minus_eur	4.368									greater_proportion	3.951659
ebtida	4.410									increasein	4.777102
pershare	4.480									greatly_improved	6.830181
earnings_pershare	4.632										

Table A.9: **Topic-term lists with term weighting factors.** This table reports the six predefined topic categories with respective IDF scores for term weighting.

Appendix B. Calculation of accumulated local effects

To visualize the effects of our features, we utilize accumulated local effects (ALE) plots (Apley and Zhu, 2019), a faster-to-compute alternative to partial dependence plots (Friedman, 2001) that is not negatively affected by correlated predictors. Suppose that we are interested in the effect of feature x_j on the prediction of the trained model, $f(x;\hat{\theta})$. We calculate the ALE curve of this feature, $ALE_j(x_j)$, by averaging the changes in the model's prediction and accumulating these differences:

$$ALE_j(x_j) = \int_{x_{min,j}}^{x_j} \mathbb{E}\left[f^j(X_j, X_{\setminus j}) \mid X_j = z_j\right] dz_j \; - \; ALE_0. \tag{B.1}$$

The vector $x_{\setminus j}$ denotes the set of features without the feature of interest, x_j, and $x_{min,j}$ is the lower boundary of the support of feature j's marginal probability distribution. In this representation of the ALE, we assume that f is differentiable in x_j and possesses a continuous partial derivative $f^j(x_j, x_{\setminus j}) = \frac{f(x_j, x_{\setminus j};\hat{\theta})}{\partial x_j}$ with respect to the feature of interest. Also, $\mathbb{E}\left[f^j(X_j, X_{\setminus j}) \mid X_j = z_j\right]$ must be continuous in z_j. The constant ALE_0 is chosen such that $ALE_j(x_j)$ has a mean of zero.

To obtain an estimate $\hat{ALE}_j(x_j)$ from actual data, we first partition the sample range of x_j into K intervals denoted as $\{N_j(k)(z_{k-1,j}, z_{k,j}] : k = 1, 2, \ldots, K\}$. Let $n_j(k)$ further denote the number of observations in interval j, and $k_j(x)$ the partition index that observation x falls into. In the actual application, we set $z_{k,j}$ to the k/K quantile of feature j. Given the partition, the ALE curve for feature x_j can be estimated by

$$\hat{ALE}_j(x_j) = \sum_{k=1}^{k_j(x_j)} \frac{1}{n_j(k)} \sum_{\{i:x_{i,j} \in N_j(k)\}} \left[f(z_{k,j}, x_{i,\setminus j};\hat{\theta}) - f(z_{k-1,j}, x_{i,\setminus j};\hat{\theta})\right] \; - \; \hat{ALE}_0. \tag{B.2}$$

The inner sum averages over the differences in prediction of all observations in the k-th interval, thereby exchanging observation i's value of feature j with the boundary values of the interval. The outer sum then accumulates all average differences for all intervals up to the interval of value x_j. As before, the constant $\hat{ALE}_0$ is chosen such that the mean ALE estimate is zero with respect to the marginal empirical distribution of x_j. Instead of considering just a single variable, the approach can also be extended to two features. Further details on the implementation in this case can be found in Apley and Zhu (2019). The notion of local effects is crucial for the correct interpretation of ALE plots: Each segment of the curve, computed from interval k, shows the local effect of x_j as the average prediction change from all observations in that interval.

Appendix C. Descriptive variable statistics

	Mean	Std. Dev.	Q05	Q25	Q50	Q75	Q95
$MV_{s,q}$	9.6119	0.6779	8.6006	9.1200	9.5532	10.0656	10.8176
$BM_{s,q}$	0.5430	1.8267	0.0647	0.2481	0.4274	0.6854	1.2657
$EP_{s,q}$	0.0110	0.1843	-0.0076	0.0084	0.0135	0.0190	0.0342
$SP_{s,q}$	0.2550	0.5416	0.0304	0.0742	0.1412	0.2726	0.7957
$CP_{s,q}$	0.0274	0.1046	-0.0069	0.0119	0.0202	0.0342	0.0842
$DY_{s,q}$	0.0045	0.0107	0.0000	0.0000	0.0030	0.0067	0.0138
$DR_{s,q}$	0.3411	2.0895	0.0000	0.0000	0.1615	0.4194	1.4518
$BMFE_{s,q}$	0.0056	1.7123	-0.0962	-0.0039	0.0000	0.0039	0.0602
$EPFE_{s,q}$	-0.0007	0.1723	-0.0052	-0.0002	0.0006	0.0019	0.0081
$SPFE_{s,q}$	-0.0008	0.1925	-0.0214	-0.0020	0.0005	0.0037	0.0230
$DYFE_{s,q}$	0.0001	0.0098	-0.0002	0.0000	0.0000	0.0000	0.0004
$CPFE_{s,q}$	0.0002	0.0575	-0.0253	-0.0005	0.0000	0.0021	0.0266
$V\text{-}EPS_{s,q}$	0.0012	0.0135	0.0000	0.0001	0.0003	0.0008	0.0039
$D\text{-}EPS_{s,q}$	0.0016	0.0086	0.0001	0.0003	0.0006	0.0013	0.0048
$C\text{-}EPS_{s,q}$	9.7776	7.2650	2.0000	5.0000	8.0000	13.0000	24.0000
$N\text{-}EPS_{s,q}$	11.7004	7.9057	2.0000	5.0000	10.0000	17.0000	27.0000
$D\text{-}REV_{s,q}$	0.0070	0.0412	0.0001	0.0008	0.0020	0.0055	0.0250
$N\text{-}REV_{s,q}$	9.4420	6.7057	2.0000	4.0000	8.0000	13.0000	23.0000
$D\text{-}DIV_{s,q}$	0.0002	0.0046	0.0000	0.0000	0.0000	0.0000	0.0006
$Q\text{-}LEN_{s,q}$	4.3868	0.3493	3.9960	4.3004	4.4467	4.5481	4.6699
$I\text{-}LEN_{s,q}$	4.2446	0.2197	3.9434	4.1544	4.2668	4.3663	4.4971
$N\text{-}ANA_{s,q}$	7.8884	3.8662	2.0000	5.0000	8.0000	10.0000	15.0000
$I\text{-}P_{s,q}$	0.0318	0.0151	0.0111	0.0214	0.0301	0.0403	0.0581
$I\text{-}N_{s,q}$	0.0243	0.0158	0.0052	0.0133	0.0212	0.0318	0.0532
$Q\text{-}P_{s,q}$	0.0280	0.0198	0.0000	0.0136	0.0250	0.0390	0.0642
$Q\text{-}N_{s,q}$	0.0264	0.0192	0.0000	0.0136	0.0230	0.0351	0.0606
$FA\text{-}P_{s,q}$	0.0300	0.0177	0.0000	0.0183	0.0283	0.0400	0.0609
$FA\text{-}N_{s,q}$	0.0323	0.0180	0.0053	0.0206	0.0306	0.0419	0.0628
$SA\text{-}P_{s,q}$	0.0290	0.0172	0.0000	0.0175	0.0273	0.0385	0.0586
$SA\text{-}N_{s,q}$	0.0326	0.0173	0.0062	0.0215	0.0312	0.0422	0.0625
$I\text{-}RE\text{-}P_{s,q}$	0.0242	0.0187	0.0024	0.0114	0.0203	0.0324	0.0580
$I\text{-}RE\text{-}N_{s,q}$	0.0248	0.0320	0.0000	0.0058	0.0150	0.0324	0.0809
$I\text{-}EA\text{-}P_{s,q}$	0.0257	0.0231	0.0012	0.0101	0.0202	0.0347	0.0678
$I\text{-}EA\text{-}N_{s,q}$	0.0214	0.0316	0.0000	0.0032	0.0108	0.0268	0.0789
$I\text{-}OU\text{-}P_{s,q}$	0.0313	0.0238	0.0025	0.0154	0.0268	0.0416	0.0744
$I\text{-}OU\text{-}N_{s,q}$	0.0139	0.0159	0.0000	0.0034	0.0094	0.0189	0.0431
$I\text{-}EN\text{-}P_{s,q}$	0.0538	0.0369	0.0000	0.0288	0.0489	0.0722	0.1199
$I\text{-}EN\text{-}N_{s,q}$	0.0230	0.0265	0.0000	0.0031	0.0150	0.0336	0.0744
$I\text{-}LI\text{-}P_{s,q}$	0.0255	0.0354	0.0000	0.0000	0.0123	0.0384	0.0946
$I\text{-}LI\text{-}N_{s,q}$	0.0070	0.0178	0.0000	0.0000	0.0000	0.0057	0.0380
$I\text{-}CH\text{-}P_{s,q}$	0.0502	0.0380	0.0083	0.0255	0.0426	0.0653	0.1144
$I\text{-}CH\text{-}N_{s,q}$	0.0258	0.0306	0.0000	0.0070	0.0171	0.0341	0.0776
$Y_{s,q}^{-5}$	0.0031	0.0767	-0.1157	-0.0347	0.0030	0.0411	0.1214
$Y_{s,q}^{5}$	0.0021	0.0552	-0.0738	-0.0236	0.0002	0.0250	0.0832
$Y_{s,q}^{20}$	0.0019	0.0912	-0.1201	-0.0409	-0.0012	0.0393	0.1294
$Y_{s,q}^{60}$	0.0045	0.1574	-0.2065	-0.0706	-0.0010	0.0700	0.2187

Table C.10: **Descriptive variable statistics.** This table provides descriptive statistics for the cleaned sample of earnings announcement events ($n = 45560$). To economize on space, we do not report statistics for topic-sentiment variables of the Q&A section of the earnings call.

Appendix D. Robustness checks

	Forecast horizon									
	5 days					60 days				
	Min.	Max.	Mean	Std. dev.	Baseline	Min.	Max.	Mean	Std. dev.	Baseline
Panel A: All out-of-sample earnings events										
Mean	0.312	0.331	0.321	0.0072	0.319	0.508	0.579	0.545	0.0222	0.567
t-statistic	11.102	11.801	11.443	0.2586	11.357	6.659	7.585	7.147	0.2909	7.430
Median	0.190	0.205	0.199	0.0049	0.1973	0.3010	0.3682	0.3334	0.0182	0.3445
Dir. Acc.	52.239	52.415	52.321	0.0571	52.376	50.607	50.836	50.729	0.0665	50.678
Panel B: Earnings events from the flop and top deciles										
Count	6504	6586	6540.5	21.7983	6529	6464	6542	6497.6	25.2947	6464
Mean	0.910	0.978	0.943	0.0211	0.954	1.723	1.936	1.849	0.0626	1.936
t-statistic	10.281	11.389	10.822	0.3576	11.140	7.277	8.132	7.771	0.2442	8.078
Median	0.573	0.629	0.597	0.018	0.5838	0.988	1.158	1.073	0.0546	1.135
Dir. acc.	55.228	55.751	55.499	0.1666	55.416	53.119	53.641	53.376	0.1730	53.573

Table D.11: **Random forest seed robustness.** The table presents distributional characteristics of main performance metrics from the application of the random forest model with 10 different seed values of the random number generator. The column *Baseline* restates the results given in the main text.

	Forecast horizon									
	5 days					60 days				
	All events		Top-flop events			All events		Top-flop events		
Parameter setting	Mean	Acc.	Count	Mean	Acc.	Mean	Acc.	Count	Mean	Acc.
Baseline	0.319	52.38	6529	0.954	55.42	0.567	50.68	6464	1.936	53.57
Max. tree depth $J = 10$	0.268	52.10	6560	0.932	55.44	0.505	50.25	6482	1.842	52.98
Max. tree depth $J = 30$	0.322	52.60	6577	0.909	55.20	0.458	50.69	6516	1.760	53.16
Number of trees $B = 2500$	0.307	52.40	6523	0.939	55.62	0.541	50.67	6456	1.892	53.64
Number of trees $B = 10000$	0.330	52.31	6541	0.966	55.57	0.528	50.77	6489	1.906	53.66
Ranking window $W = 500$	0.311	52.38	6523	0.916	55.25	0.537	50.68	6575	1.731	52.92
Ranking window $W = 1000$	0.312	52.38	6539	0.948	55.64	0.558	50.68	6539	1.867	53.39

Table D.12: **Results from alternative model configurations.** The table presents results from alternative parameter settings for the random forest. Each alternative configuration changes the stated parameter and keeps all remaining parameters unchanged.

Appendix E. Further breakdown of mean abnormal returns

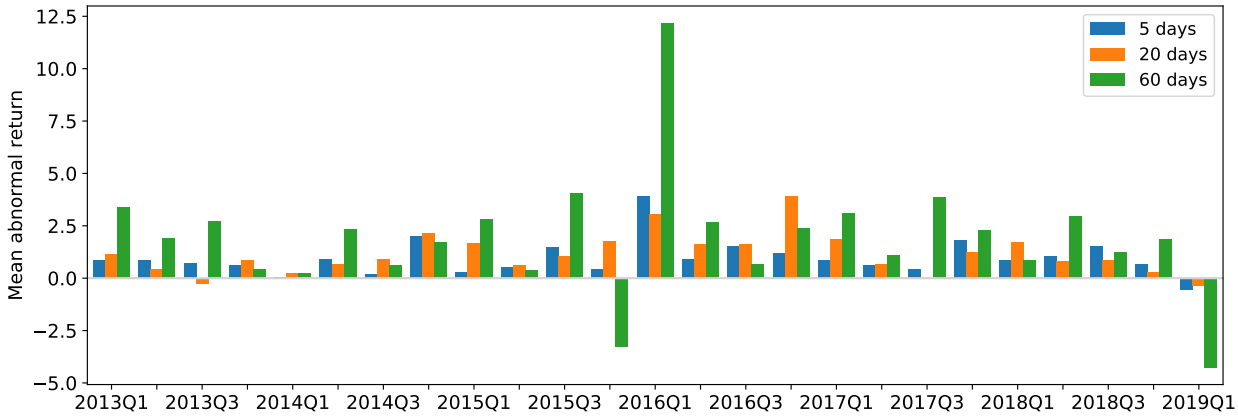

Figure E.12: **Subperiod breakdown of mean top-flop abnormal returns.** This figure depicts mean abnormal top-flop returns from the random forest model for forecast horizons of 5 (blue bars), 20 (orange bars) and 60 days (green bars) conditional on the shown year-quarter bins, which are based on the date of the earnings call.

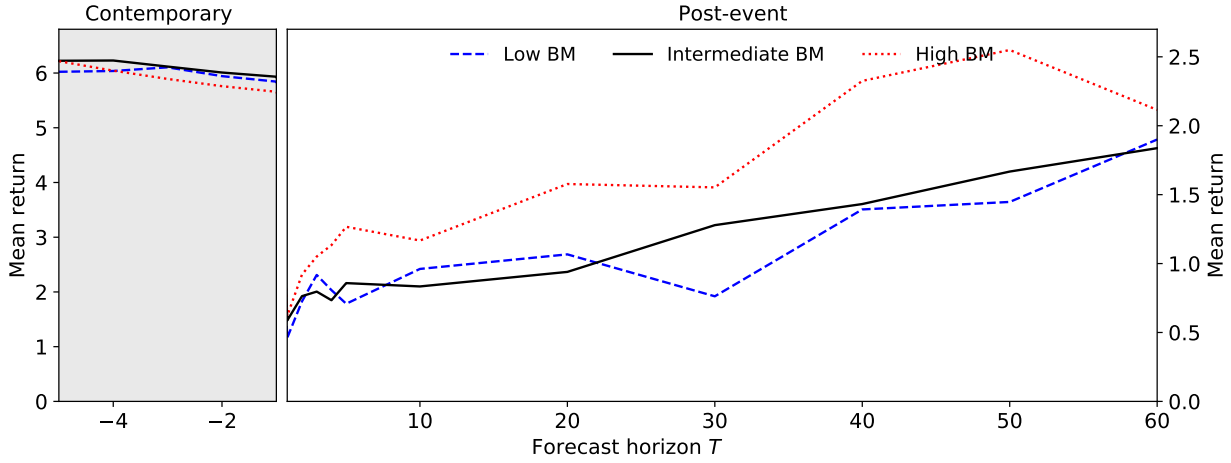

Figure E.13: **Evolution of mean top-flop abnormal returns by book-to-market ratio.** This figure compares the evolution of mean buy-and-hold abnormal returns for top-flop earnings events, conditional on the book-to-market ratio (BM). The dashed blue (dotted red) lines depict the evolution of mean returns for events with a book-to-market ratio in the lowest (highest) quintile. The solid black line shows mean returns for the intermediate three quintiles.