Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 8 (2022) 120–137

http://www.keaipublishing.com/en/journals/jfds/

Forecasting earnings and returns: A review of recent advancements

Jeremiah Green*[,1], Wanjia Zhao [1]

Received 31 January 2022; revised 20 April 2022; accepted 26 April 2022
Available online 16 May 2022

Abstract

We selectively review recent advancements in research on predictive models of earnings and returns. We discuss why applying statistical, econometric, and machine learning advancements to forecasting earnings and returns presents difficult challenges. In the context of these challenges, we discuss recent papers that confront the challenges and present promising advancements and paths for future research.

Keywords: Earnings; Returns; Machine learning; Uncertainty; Measurement error; Forecasting

1. Introduction

Forecasting accounting earnings and forecasting stock returns have long histories in accounting and finance.[1–4] While earnings and returns forecasting have developed differently, the two are not independent; earnings and returns may measure the same underlying variable or forecasted earnings may be an important driver of returns.[1,5–9] Prior research has reviewed these literatures (e.g. Richardson et al[1]; Timmermann[10]) and other research has reviewed forecasting techniques more generally (e.g. Petropoulos et al[11]). However, a flurry of recent research applies new techniques, new variables, and new data to forecasting earnings or returns. As an example of these innovations at least for returns, Karolyi and Van Nieuwerburgh[12] introduces an entire issue of *The Review of Financial Studies* dedicated to methodological developments in explaining or forecasting returns.[a]

Despite the recent or continuing interest, forecasting earnings and returns present the forecaster with significant challenges. Prior research has discussed the challenges that come from forecasting financial data (e.g. Timmermann[10]). These include the unpredictable nature of financial data, a low signal-to-noise ratio in available variables, and model uncertainty. For support of the first order nature of these concerns, consider some relevant evidence:

* Corresponding author. James Benjamin Department of Accounting, Mays Business School, Texas A&M University, 449C Wehner Building, College Station, TX 77843, USA.

E-mail addresses: jgreen@mays.tamu.edu (J. Green), wzhao@mays.tamu.edu (W. Zhao).

Peer review under responsibility of KeAi.

[1] Both authors are at Texas A&M University.

[a] For a recent review of how machine learning has been applied to returns and asset pricing, see Weigand.[13]

https://doi.org/10.1016/j.jfds.2022.04.004

> Moreover, at a one year horizon, the random walk model performs as well as modern sophisticated methods that use larger predictor sets. This finding echoes an old result that, given recent applications of forecasts in the literature, may have been forgotten.[3]

> Outside of microcaps, the hedge returns to exploiting characteristics-based predictability also have been insignificantly different from zero since 2003.[14]

> … machine learning forecasts for return realizations beyond one year into the future are no better than a zero forecast in discriminating between high and low expected return firms.[15]

> … investments based on deep learning signals extract profitability from difficult-to-arbitrage stocks and during high limits-to-arbitrage market states. In particular, excluding microcaps, distressed stocks, or episodes of high market volatility considerably attenuates profitability.[16]

These quotations help to temper enthusiasm and inject reality into the surge towards new developments. For new developments to have a meaningful, lasting impact on our collective understanding of earnings and returns and for these developments to yield predictions that are conceptually and practically useful, they need to directly confront the primary obstacles that occur in these challenging data. As an example of what may be a successful approach, the last two quotations come from papers that argue that imposing restrictions on machine learning methods improves return forecasts. Perhaps this type of approach points to an important and fruitful path for future research.

Given the continuing developments in the literature and the apparent enthusiasm for new methods and data, we review recent advancements in this literature. To make the task manageable, we put some bounds on our review.[b] First, because our focus is on advancements for the purposes of prediction, we will not deal with papers considering important questions such as whether predictability is attributable to mispricing or risk or whether the results in the literature are subject to multiple testing problems except when these papers are relevant to our discussion. Second, one of the data constraints in the literature is a limited time-series, particularly in the case of low frequency accounting information. On the other hand, there are large cross-sections of companies, we therefore also focus only on the cross-section of earnings and equity returns for individual companies and stocks, not the time-series of returns for individual stocks or portfolios of stocks. Third, because many of these advancements are very recent and are currently being developed, we include in the review many working papers as well as published papers.[c]

We organize our review using the empirical challenges that forecasters encounter when forecasting financial data: unpredictability, a low signal-to-noise ratio, and model uncertainty. To begin the review, we provide a conceptual and empirical description of these empirical challenges.

Our empirical description and our review of the recent literature lead to some insights and questions. Beginning with the insights, first, even with recent advancements, returns and earnings appear to be largely unpredictable. Second, even with the large number of candidate predictors for returns,[4,17,18] recent research continues to find new predictors. On the other hand, the variables that predict earnings continue to be elusive[3] but there are some advancements that are encouraging. Third, nonlinearities and flexible modeling methods may have some benefits in forecasting, but out of the box methods may have limited ability to improve forecasting. Turning to the questions, fourth, what can the collective empirical developments teach us about the conceptual drivers and patterns in earnings and returns? Fifth, much like the current efforts to make sense of the "zoo" of return predictors (e.g. Hou et al[17]; Cochrane[19]), how do the advancements in different data and methods combine or interact?

[b] While we appreciate some of the technical differences developed in this research, we will also keep our review non-technical.

[c] Human forecasts, such as analyst forecasts and management forecasts, although taking considerable space in the literature, are not the focus of our review. However, we do discuss a few papers that are relevant to our points.

2. Empirical description of forecasting and the challenges with financial data

To help readers have a more concrete idea of the conceptual framework, we begin with the standard cross-sectional regression model for earnings or returns that posits a linear form.[d]

$$Y_{i,t+1} = \alpha_t + \beta_t X_{i,t} + \in_{i,t+1} \tag{1}$$

Here Y is the earnings or return variable to be predicted. It can be observed for company i in a time period $t + 1$ after time t, where t might be a fiscal year or quarter for earnings and a day, week, month, or some other period for returns. X is a vector of predictor variables observed at or before time t for each company i. When forecasting returns, forecasted earnings may be an important X variable in addition to being an important Y variable. α and β are to be estimated and ε is the error term. In the regression estimates α and β only have a time t and not a company i subscript. The reason is that these are cross-sectional regression models that are estimated in each cross-section following Fama and MacBeth.[20,e] When making a prediction, the forecasted Y variable becomes

$$\widehat{Y}_{i,t+1} = E_t[\alpha_{t+1}] + E_t[\beta_{t+1}]X_{i,t} \tag{2}$$

In this equation, E_t indicates the expectation of what α or β will be at t+1, with the information set that is available at time t.

As mentioned in the introduction, when forecasting financial variables with financial variables, the forecaster encounters important problems: unpredictability, a low signal-to-noise ratio, and model uncertainty. We first describe these problems conceptually and then outline their importance empirically.

2.1. *Forecasting challenges*

The first challenge in modeling earnings and returns is that despite the collective efforts of a large number of researchers over many years, the surviving models of earnings and returns only explain a small fraction of the variation in the Y variables. Put in another way, to a large degree, earnings and returns are unpredictable or the variance of ε is large.

The recognition of the noise in returns began early and one of the solutions was to model the returns to portfolios of stocks to hopefully "average away" the noise in returns. While the averaging of returns in portfolios makes sense for providing descriptive information or for forming tradeable portfolios, it is usually the case that the primary objective is the firm level prediction of returns. The unpredictable nature of earnings and returns may come from human actions that are idiosyncratic and unpredictable. On the other hand, the models of earnings and returns may fail to capture what could be predictable with otherwise omitted variables and more complex models.

The second challenge is the low signal-to-noise ratio of financial variables. This is related to the unpredictability of earnings and returns, but here we refer more specifically to the low signal-to-noise ratio of the predictor variables, X. To counter the noise in the X variables, research has presented two approaches.

The first of these counters to the low signal-to-noise ratio recognizes that there may exist important X variables that are not included in the model. Innovations to the basic model initially began with adding additional X variables. There is a large literature proposing return predictors, so the potential X variables number at least in the hundreds.[4,17,18] The large literature for return predictors follows earlier research that added to and aggregated from proposed predictors (e.g. Lewellen[21]; Jacobs and Levy[22]). For earnings, the number of proposed predictors is smaller (e.g. Li and Mohanram[2]) but earlier research did relate a larger number of variables to the prediction of earnings and to the prediction of returns (e.g. Abarbanell and Bushee[6]; Ou and Penman[7]; Lev and Thiagarajan[23]). While the paths to predict earnings and returns might have diverged, some research argues that the predictors for earnings and returns should be related.[1]

The second counter to the noise in X variables has been to assume that X variables represent a smaller number of observed or latent variables. Asset pricing research has attempted to find a smaller set of variables that capture the information from the large set of possible predictors. Recent research in asset pricing primarily focuses on this issue.[12]

[d] We recognize that the basic methodology we outline is a coarse description of the many variations to the modeling of earnings and returns. However, we believe this approach is general and helps to clearly illustrate the challenges and advancements in this literature.

[e] Note that α and β have subscripts t to mean that to predict Y, we must have an expectation for α and β to make a prediction. However, α and β are not estimated until Y is observed, so in a regression, they might have a subscript of $t + 1$ to recognize that they are estimates that are made using Y_{t+1}.

The third challenge that is also related to unpredictability and noise in X is model uncertainty. Model uncertainty can come from many sources but may be a first order concern.[24] We discuss model uncertainty in three forms. First, model uncertainty may arise because the form of the model is uncertain and most relevant for some of the recent advancements in the literature, the model of earnings and returns may not be linear. Second, even with a correctly specified model, model coefficients may vary over time or across companies, thus lowering the quality of the expectations for the coefficients when making forecasts.

2.2. *Empirical description of the challenges*

To provide some empirical context to the challenges of modeling earnings and returns, we use a common empirical design where we use a large cross-section of stocks and monthly returns from 2000 to 2020 with a large set of return predictors. We use a subset of predictors from Green et al[14] that can be calculated from CRSP and Compustat. The definitions for the firm characteristics are provided in the appendix. For the cross-section of monthly stock returns, we use predictors that we assume to be available as of the end of the month, and delisting adjusted returns are from the end of the month to the end of the subsequent month. To make the comparison straightforward, we use the same data set for earnings prediction and predict 12-month ahead accounting return on equity (*ROE*) and change in earnings (*SUE*). To deal with outliers, we winsorize the predictors each month at the 1st and the 99th percentiles and rank the predictors into deciles.[f]

We compare the pooled data in-sample R-squared value for alternative modeling approaches. Because these R-squared values are for in-sample fit, they likely represent an upper bound on the achievable fit out-of-sample.[g]

To highlight the challenges to forecasting returns and earnings, we compare the following models. We choose these models to test the relative magnitudes of the challenges presented in the previous section as discussed below.

In the first model Eq. (3), $\overline{Y}_{t+1}$ is the mean Y value for each month. i indexes company and t month. β is the single coefficient that is constant over time and across companies. α is an intercept and ε is an error term. The purpose of this model is to describe how well the realized market level variable explains the time-series panel of returns or earnings. This model is equivalent to a time fixed effect model. We refer to this model as the *market model*.

$$Y_{i,t+1} = \alpha + \beta \overline{Y}_{t+1} + \varepsilon_{i,t+1} \tag{3}$$

In the second model Eq. (4), $\overline{Y}_{t+1}$ is the two-digit SIC industry and month mean. β is the single coefficient that is constant over time and across companies. α is an intercept and ε is an error term. This model is equivalent to an industry-month fixed effect model. We refer to this model as the *industry model*. The first two equations follow prior research that finds that market and industry information explains stock returns and earnings (e.g. Ayers and Freeman[26]).

$$Y_{i,t+1} = \alpha + \beta \overline{Y}_{j,t+1} + \varepsilon_{i,t+1} \tag{4}$$

The first thing to note about these two equations is that the market and industry variables are only known ex-post. These regressions describe how well, if known, aggregate information explains firm-level stock returns and earnings. These equations provide a benchmark for understanding the extent to which earnings and returns are explained by aggregate variation. Or from the opposite view, how much variation in earnings and returns is idiosyncratic and thus cannot be explained by aggregate variation. We use this as a benchmark for how unpredictable earnings and returns are.

The next model evaluates the challenge of the noise in X.[h] In Eq. (5), X is a vector of predictors 1 to H known at the end of month t. The estimated coefficients are assumed constant for the pooled sample. We refer to this model as the *linear characteristics model*.

[f] The R code for our data and analysis are available in the appendix.

[g] Following Gu et al[25]; we calculate the R-squared for each model where the mean of the Y variable is assumed to be zero, i.e. total sum of squares is $\sum (y-0)^2$. As noted in that paper, the historical mean as the benchmark with which to compare the model against is unlikely to be informative about the value of the model because historical averages are poor predictors of returns. We note that the same adjustment may not be applicable for *ROE* because of the autocorrelation in the measure. However, to make R-squared comparable across the different regressions we use the same R-squared calculation.

[h] Another option would be to show models that use different subsets of the X variables we have available. Our purpose here is to use the full set of X variables as a benchmark. Recent research tries to find subsets of X variables that work as well as but not better than the full set of X variables.[27–30]

$$Y_{i,t+1} = \alpha + \sum_{1}^{H} \beta_h X_{h,i,t} + \varepsilon_{i,t+1} \tag{5}$$

We also use a variation to Eq. (5) where the model allows for non-linear relations between the *X* and *Y* variables using an extreme-gradient-boosting regression tree. We use this machine learning approach because it allows for highly non-linear associations between *X* and *Y*. We use the XGBOOST and the caret package in R and use the default training parameters. This equation is the first of our equations that addresses model uncertainty. We refer to this model as the *non-linear characteristics model*.

Our final two models allow coefficients to vary by month *t* and to vary by industry *j*. Eqs. (6) and (7) provide two additional models to evaluate model uncertainty. If allowing for timevarying and industry-varying coefficients increases the R-squared to a great extent, then it implies that forecasting with the assumption of static coefficients overtime or homogeneous coefficients across companies may struggle. We refer to these models as the *time-varying coefficients model* (Eq. (6)) and the *industry-varying coefficients model* (Eq. (7)).

$$Y_{i,t+1} = \alpha_t + \sum_{1}^{H} \beta_{h,t} X_{h,i,t} + \varepsilon_{i,t+1} \tag{6}$$

$$Y_{i,t+1} = \alpha_j + \sum_{1}^{H} \beta_{h,j} X_{h,i,t} + \varepsilon_{i,t+1} \tag{7}$$

Comparing the R-squared values from these models provides the basis for describing the empirical challenges to using and developing new models and approaches to forecasting returns and earnings.

Table 1 provides the R-squared values for these models. We first comment on the unpredictability of the *Y* variables. The highest R-squared value for returns, ROE, and SUE are 0.266, 0.620, and 0.085 respectively. Remember if these are in-sample upper limits, most of the variation in *Y*, and even more particularly for returns and changes in earnings, is left unexplained by the models.

Y may be in large part difficult or impossible to model. If we are to model returns and earnings successfully, because the error in *Y* is large, the primary drivers of earnings and returns may be idiosyncratic to firms and time periods. This first challenge introduces issues for which the benefits of data and methodological advancements may be small or difficult to achieve.[i]

Turning to the individual models, we next focus on the first three rows. The first two rows display the results from including respectively known ex-post aggregate market and industry returns, and aggregate *ROE* and *SUE*. The aggregate information explains a larger fraction of the variation in returns (0.167 for market returns and 0.232 for industry returns) than the variation in *ROE* (0.017 and 0.101) and *SUE* (0.002 and 0.064). The R-squared value for both earnings measures using the market mean of the corresponding earnings measures are very low. The differences between the returns and earnings models raise the possibility that an important driver of returns is the aggregate market but that differences across firms is more important for explaining earnings. For all three variables, there is a meaningful increase in R-squared when moving to the industry mean. However, in percentage terms, the increase in R-squared is larger for the earnings measures, suggesting that the variation across firms in earnings measures is at least partially

Table 1
In-sample R-squared.

	Returns	ROE	SUE
Market model	0.167	0.017	0.002
Industry model	0.232	0.101	0.064
Linear characteristics model	0.019	0.511	0.024
Non-linear characteristics model	0.044	0.620	0.063
Time-varying coefficients model	0.266	0.558	0.085
Industry-varying coefficients model	0.032	0.565	0.067

[i] Prior research on returns forecasting also recognizes the low model fit problem and focuses on differential improvements to fit.[31]

explained by industry variation. These patterns suggest that the challenges with unpredictability may be different for returns and earnings.

Moving to the third row, the inclusion of firm characteristics shows the other side of the same story. Firm characteristics explain a larger portion of earnings (0.511), but little variation in changes in earnings (0.024) or returns (0.019). The increase for *ROE* may reflect that lagged earnings explains a large portion of earnings, but very little else provides a meaningful improvement.[3] An important difference between rows one and two and row three is that the firm characteristics are known ex-ante and thus represent models that can be considered prediction models. If the *X* variables only provide a small prediction benefit, like for returns or changes in earnings, or provide little incremental benefit over the historical level of earnings for *ROE*, then the question becomes whether other unknown *X* variables could explain *Y* better. But, here we cannot provide any evidence about whether some *X* variables that are not included could otherwise better explain the *Y* variables.

Moving to model uncertainty, to address the assumption of linearity, we use XGBOOST. The descriptive information here is important because recent papers use non-linearity as a primary motivation for adopting machine learning methods. Row four shows the R-squared values. For all three *Y* variables, the increase from using a more complex model is meaningful. However, the increase for returns remains far below the ex-post aggregate models. This raises the possibility that improving the characteristics model for returns may only provide modest benefits, even if these benefits can provide improvements in tradeable strategies. However, relative to the linear case in row three, the improvements are large. The improvement from moving to non-linearity seems particularly useful for *SUE* relative to the prior rows.

The last two rows allow coefficients to vary by month or to vary across industries. Note that because the time-varying models include an intercept that is constant in the cross section, these models incorporate the aggregate ex-post mean as well as the individual *X*s. Accordingly, these models are not strictly prediction ready, but they provide a comparison with rows 1–3. The timevarying model for all three *Y* variables only increases the R-squared modestly relative to the best linear alternative models; for *Returns* 0.266 versus 0.232, *ROE* 0.558 versus 0.511, and *SUE* 0.085 versus 0.064. This means that the primary driver of the R-squared for returns is still the aggregate mean, and that the characteristics with the time-varying coefficients only slightly improves the model fit. In fact, the industry mean is almost as effective as this model. However, there does appear to be some benefit to using the characteristics when allowing the coefficients to vary by month. It is possible that the time-varying coefficients and characteristics largely capture the ex-post industry mean or perhaps they add new information to it.

For *ROE*, the improvement relative to including characteristics with constant coefficients is mediocre. This again raises the challenge of creating a model that meaningfully improves upon a model that only uses lagged *ROE*. For *SUE* this appears to be the best modeling approach, but again it is similar to the industry mean. Finally, allowing coefficients to vary across industries does little to improve the returns model (0.032), although it is slightly better than the constant coefficients model (0.019). The improvements to the earnings models are small as well and achieve approximately the same improvement as the non-linear model or the time-varying coefficients model.

With the context of the conceptual and empirical challenges, we now turn to a review of the recent advancements in the literature.

3. Literature review

In the following subsections, we review recent advancements in the literature. We primarily review advancements that have been made in the last few years. We follow the outline developed in the prior section where we organize studies by how they relate to the unpredictability of the *Y* variables, the noise in the *X* variables, and model uncertainty.

3.1. Unpredictability

The unpredictability of *Y* for returns has a long history in research. Prior research links fundamental information (dividends and earnings) to stock prices and finds that stock prices are too volatile.[32–37] The primary issue is that stock prices deviate from the fundamental drivers of value, i.e. earnings. The excess volatility in stock prices could occur because investors act based on sentiment or other behavioral based reasons[38,39] or because investors face uncertainty when learning about fundamentals.[40] Similar to the excess volatility in stock prices, analysts' forecasts for earnings appear to be too volatile and the excess volatility in analysts' forecasts may affect investors that rely on analysts' forecasts.[41–43] The excess volatility of returns reflects uncertainty or leads to uncertainty; if investor learning and behavioral trading are unpredictable, then an important portion of returns will also be unpredictable.

On the other hand, earnings, and more specifically expectations of earnings, are presumably one of the fundamental drivers of returns. If returns are too volatile because investors face uncertainty when forming expectations about earnings,[40] then a corollary is that earnings forecasting is an uncertain task.

Numerous papers have worked to deal with the unpredictability of returns. In one summary of this literature, Timmermann[10] discusses some potential solutions for confronting unpredictability, as well as model uncertainty which we come to later; for example, some solutions include using combinations of forecasts, using the history of model errors, shrinkage methods using Bayesian statistics, economic restrictions to forecasts, out-of-sample cross validation testing, and addressing multiple testing problems. We will refer to recent advancements related to some of these solutions.

As one way to explicitly recognize the unpredictable nature of financial variables, Timmermann[10] discusses forecasting the distribution of returns, density forecasts, from time series models.[j] Moving away from point forecasts of earnings or returns explicitly recognizes that forecasting is an uncertain task and the point forecast is likely to be far-off. Recent research applies cross-sectional models of returns and earnings to forecasting the distributions of Y. While the methods vary somewhat, the motivation and approaches are similar. Therefore, we only discuss in detail one specific approach.

Research in finance uses high frequency time-series data to forecast moments of a time series distribution.[10,45–47] However, when using cross-sections of data, forecasting a distribution for individual companies at a point in time requires different tools. New research has worked towards forecasting the distribution of returns and earnings from these cross-sections. While these methods, such as cross-sectional density forecasting models, are being developed,[48,49] perhaps the most widely used and intuitive method is quantile regression.

$$Y^{Q}_{i,t+1} = \alpha^{Q} + \sum_{1}^{H} \beta^{Q}_{h} X_{h,i,t} + \varepsilon_{i,t+1} \tag{8}$$

Here a quantile regression can be seen as following the same type of model as Eq. (5). The primary difference between the ordinary least squares model and the quantile regression model is that the coefficients, α and β, are estimated separately for each quantile of the Y distribution.[50,51,k] Estimating coefficients conditional on the Y variable requires a different estimation method because which quantile an observation belongs to is required for making a prediction. However, recent research uses the distribution of the αs and βs from different quantiles along with the characteristics for an observation to form a distribution of forecasted Y. More specifically, the point estimates for a company at different quantiles of a distribution are given by the estimates from quantile regressions. For example, for the 25th, 50th, and 75th percentiles, the forecasted point estimates would be given by the following equations.

$$Y^{25}_{i,t+1} = \alpha^{25} + \sum_{1}^{H} \beta^{25}_{h} X_{h,i,t}$$

$$Y^{50}_{i,t+1} = \alpha^{50} + \sum_{1}^{H} \beta^{50}_{h} X_{h,i,t}$$

$$Y^{75}_{i,t+1} = \alpha^{75} + \sum_{1}^{H} \beta^{75}_{h} X_{h,i,t}$$

Using these conditional forecasts, meaning that the forecast would be $Y^{25}_{i,t+1}$ conditional on a company ending up in the 25th percentile of the Y distribution, researchers have formed estimates of the distribution of forecasted returns and earnings. Because the estimates of each point of the distribution are separate, there is no imposed assumption about the distribution of Y. This may be important because other research finds that it is the tails of the distribution that may matter

[j] The noise in Y may be different for earnings and returns. Additionally, some aspects of returns (or earnings) may be more predictable than other aspects.[44]

[k] Many papers make further developments to the initial quantile regression proposed.[52]

for returns.[53,54] Because of this feature, some researchers examine particular quantiles or combine the conditional Y forecasts in different ways.[1]

Several papers apply cross-sectional quantile regressions to forecasting returns. Advancements in this effort are important because estimating the return probability density function is an ongoing, but elusive goal.[55–57] While we focus on the quantile regression, other papers apply functional data methods to estimate the return density functions with cross-sectional data. For example, Kokoszka et al[48] compares different functional data analysis methods for estimating the return probability density function. Their log quantile density transformation outperforms other approaches for forecasting the probability density functions of future returns.

While most of the quantile regression research has applied quantile regression to forecasting risk measures such as value-at-risk,[46,58,59] more recent research applies quantile regression as a means for creating a point forecast for returns. Gowlland et al[60] argue that one of the benefits of quantile regressions is that while ordinary linear regression models focus on forecasting at the mean of the distribution, investors may often primarily care about extremes of the distribution. This paper proposes that quantile regression could be useful for investors and show some possible applications using momentum and the bookto-market ratio.

Similarly, Ma and Pohlman[61] present potential methods by which investors might use quantile regression to make investing decisions. Using long only and long-short market neutral portfolios, Pohlman and Ma[62] test whether quantile regression can improve portfolio performance. The reason for applying quantile regression in this way is that the return distribution may not be well represented by the mean forecasting from ordinary least squares (OLS). However, when using quantile regression to create a point forecast for return prediction, a quantile needs to be specified. Pohlman and Ma[62] compare the forecasts from different quantiles. Using the lower part of the return distribution (10th quantile) outperforms other quantiles and OLS in terms of the portfolio return and Sharpe ratio.

Using quantile regressions in earnings prediction research is motivated in two ways: by an interest in future earnings uncertainty and/or by the inadequacy of the short time series of data available for forecasting higher order moments of earnings.[9] Konstantinidi and Pope[63] use forecasted earnings quantiles to create measures of variance, skewness, and kurtosis. For example, for the variance of the forecasted distribution, they use the forecasted earnings from the 75th quantile minus the forecasted earnings from the 25th quantile. This gives a measure of the spread of the distribution for every firm-year. They find that these measures of expected risk in earnings are associated with future return volatility, bond spreads, and analysts' measures of uncertainty and risk. Other recent papers that employ similar methods include Chang et al[64] and Tian et al[65] Chang et al[64] find that equity prices and credit spreads are associated with the higher moments of expected earnings.

Similar to returns, some recent earnings forecasting research has applied quantile regression to the forecasts of point estimates for earnings. Tian et al[65] use quantile regression to forecast earnings and find that quantile regression better forecasts earnings when the earnings distribution has heavier tails, i.e. it is different from a normal distribution. Hendriock[66] finds that quantile regression based methods forecast earnings better than OLS and these forecasts can be improved by using a machine learning quantile method (via artificial neural networks).

We highlight in this section some promising advancements that begin to address the unpredictability of returns and earnings, and more specifically the uncertainty about point forecasts. Additionally, the same tools may be useful particularly when earnings and return distributions have skewed or heavy tailed distributions. However, we are confronted with the problem that returns and earnings may be inherently largely unpredictable and these models may not capture this unpredictability. For example, using quantile regressions to forecast a distribution only works to the extent that the quantile regression model can forecast earnings and returns. Errors in the quantile regression model will still present the forecaster with the problem that much of the variation in returns and earnings may go unexplained.

3.2. Noisy X

The next challenge when working with financial data is that the X variables are noisy. We discuss two areas of recent research that continue to address this challenge. The first is finding new X variables, and the second is using new methods that exploit variation in existing X variables to find a smaller set of important or latent variables.

[1] An important side note is that these models can only address predictable uncertainty, meaning noise in Y that is not captured by the quantile regression model will still be unexplained by the forecasted distribution.

3.2.1. New X variables

Perhaps the most common approach to improving models that suffer from existing noisy *X* variables has been to find new variables. The growth in *X* variables for forecasting returns was highlighted in Green et al[4] and the list of variables has expanded.[17] Meanwhile, the growth in *X* variables for forecasting earnings has been less expansive and has perhaps gone the opposite direction beginning with a somewhat larger set of variables (e.g. Ou and Penman[6]; Abarbanell and Bushee[7]) and moving to models with fewer variables where the focus has been on using earnings forecasts to estimate a firm's cost of capital (e.g. Li and Mohanram[2]; Easton and Monahan[67]). Despite the different paths that returns and earnings prediction have taken, there should be some overlap in the predictors of earnings and returns.[1]

Some research continues to develop new *X* variables based on traditional data sets. For example, He and Narayanamoorthy He and Narayanamoorthy[68] find that changes in earnings growth forecasts returns. Avramov et al[69] find that shocks to moving averages of accounting items predict returns. Lyle and Yohn[70] incorporate accounting information into a portfolio optimization approach and find improved portfolio performance. For earnings, Azevedo et al[71] find that combining analyst forecasts with cross-sectional earnings forecasting models outperforms either method on its own in terms of forecasting earnings.

Some research also studies conditions for which return predictability is stronger. For example, Burger and Curtis[72] find that the use of margin debt is related to market overpricing. Han et al[73] find that trading volume amplifies mispricing.

Other research develops new *X* variables from business text, which were once alternative data sources but now constitute a large literature in accounting and finance.[74–76] Li[77] uses a naïve Bayesian method to classify the tone of forward-looking statements in the MD&A and finds that tone predicts future earnings. Similarly, Huang et al[78] use a naïve Bayesian approach to extract information signals from analyst reports. In more recent research, Karapandza[79] extracts future-oriented language; Meursault et al[80] measure unexpected earnings based on business text; Heston and Sinha[81] and Ke et al[82] use machine learning techniques to measure the sentiment of business news stories. These studies find that textual measures forecast stock returns. A related number of studies also use textual measures to forecast earnings (e.g. Bochkay and Levine[83]).

Moving beyond textual analysis, other research extracts return predictors from different data sources. For example, Obaid and Pukthuanthong[84] use pictures in news stories to measure aggregate sentiment and find that aggregate sentiment predicts market returns. deHaan et al[85] find that analysts experiencing bad weather are slow to process information, suggesting that weather can predict returns. Jame et al[86] show that crowd sourced earnings forecasts contain incremental signals about earnings and returns.

Finally, in contrast to searching for new *X* variables, some papers seek to improve the measurement of existing *X* variables. Some papers focus on the measurement of individual predictors. For example, Ball et al[87] and Ball et al[88] work to improve the measurement of profitability, Ball et al[89] the measurement of book values of equity, and Cooper et al[90] the measurement of asset growth. Mohanram and Gode[91] use a different approach by focusing on measurement errors in analysts' forecasts. They find that removing predictable analysts’ earnings forecast errors improves cost of capital estimates derived from analyst forecasts. Freyberger et al[92] takes another approach that focuses on a measurement issue that applies to multiple predictors. They find that imputing missing values with conditional mean imputations and using weighted least squares regressions improve out-of-sample prediction.

3.2.2. Finding a smaller set of predictor variables

The papers related to predicting returns summarized above offer a glimpse into the massive return forecasting literature, as over the time span of 1970–2010, a zoo of more than 300 return predictive signals have been reported.[4,19] At the same time, current research attempts to tame the enthusiasm of the factor zoo coming from different perspectives. Feng et al[93] propose estimating the marginal contribution of a new factor against the existing large number of factors. Harvey et al[18] posit that the large number of return predictors may result from overfitting commonly used data sources. They introduce a framework of multiple hypothesis testing and propose a t-statistic of 3.0 to account for the multiple testing problem. Relatedly, Chinco et al[94] employ an empirical Bayesian method to re-adjust the anomaly base rate over time, which then adjusts the posterior probability of discovering a true anomaly. These papers have a common theme, which is to use statistically motivated methods to raise the hurdle for identifying true return predictive signals.

Related research tries to synthesize the information in the large number of return predictors by identifying a smaller set of observed or latent variables that capture the variation in the large set of predictors. Some papers use or propose variable selection models and find that their methods select only a small number of the large set of proposed X variables.[27,93,95,96] Other papers (e.g. Kozak et al[97] use the large set of X variables to create a small set of latent variable factors that can explain portfolio time-series returns. The general conclusion from these papers is that the large number of proposed X variables may be smaller than the literature suggests and that a small set of variables can capture most of the information from the large set of variables. We note two complications with interpreting this research. First, the different methods do not arrive at the same variables or the same conclusions. Second, reducing the number of variables to a set of robust predictors makes the X variable set smaller, but doesn't necessarily lead to better return predictions.

3.3. *Model uncertainty*

The final challenge we discuss is that when forecasting earnings and returns, the forecaster has significant uncertainty about the chosen model. We focus on two aspects of uncertainty. The first is uncertainty about the form of the model with a specific focus on linear versus non-linear models. The second is uncertainty about whether the coefficients estimated in one sample will perform well when moving to a new sample. The broader forecasting literature proposes a number of ways for addressing model uncertainty, including techniques such as cross-validation and ensemble forecasts.[10,11,m] We only discuss here the most recent literature applied to earnings and returns forecasting.

3.3.1. *Non-linearity*

The first form of uncertainty we discuss is uncertainty about the form of the model and more specifically, the assumption of linearity. One of the common motivations for using machine learning methods is that these methods impose fewer assumptions and allow for highly complex, non-linear models. Because non-linearity is used as a motivation for using machine learning, we will also include in this section a more general discussion of machine learning methods.

Before jumping into papers using machine learning methods, other research shows that nonlinearities may be important for predicting returns. Relaxing the assumption of linearity is potentially important for at least two reasons. First, non-linear models can allow coefficients, such as the β_t in Eq. (1), to vary with the level of the independent variable(s), $X_{i,t}$. For example, extreme values of X can have weaker or stronger associations with Y. Second, if a non-linear function is a better fit for the data, a linear function may produce fitted or predicted $Y_{i,t+1}$ values that are highly inaccurate for extreme levels of X. Some research shows the importance of considering non-linear models. For example, Müller and Schmickler[100] find that double-sorting anomalies, i.e. anomaly interactions, generate higher returns than individual anomalies, and Caldeira et al[101] find that non-linearities improve portfolio performance.

Moving to machine learning, a growing number of papers compare machine learning methods (e.g., random forests, gradient boosting, neural networks, etc.) with linear models or random walk models of earnings.[102–107] As discussed in prior research, lagged earnings seems to be the most important predictor when forecasting future earnings[3]; however, recent studies provide some evidence that machine learning methods may improve on the random walk forecast for earnings.

While there has been a surge in the number of papers applying machine learning to earnings forecasting, we discuss two examples here. Easton et al[108] present a k-nearest-neighbors approach to forecasting earnings where nearest neighbors are determined by different lags of prior earnings. They find that this method provides more accurate forecasts compared to a random walk model of earnings. Importantly, models that include other predictors, such as profit margin, asset turnover and accruals, do not improve earnings prediction. This finding suggests again that lagged earnings is the most important predictor for future earnings, but that the path of earnings may also matter. Elamir[109] applies extreme gradient boosting to forecasting ROE and finds that in-sample modelling improves dramatically, but the out-of-sample improvement is small.

Machine learning techniques have also been applied to forecasting returns. Gu et al[25] compare machine learning methods for forecasting returns. They conclude that machine learning methods (regression trees and neural networks)

[m] However, we note that some research has found success applying different forecasting techniques from the broader forecasting literature. For example, Dong[98] finds that aggregating anomalies improves return predictability. Elliott et al[99] use a forecast combination approach and find that their approach improves return forecasts.

improve forecasting returns because they allow for complex non-linearities among return predictors. Other papers also find that machine learning may improve return prediction.[110–114,n]

Despite the recent interest in machine learning, other research finds that machine learning is not a panacea and presents its own problems, such as in-sample overfitting, out-of-sample lackof-fit, and model instability due to a lack of sufficient training data. For example, Baba Yara[15] finds that machine learning models do not generalize – a symptom of overfitting in-sample data; but imposing restrictions on the machine learning process improves return predictability. Avramov et al[16] use various versions of deep neural networks and find that these methods identify difficult-to-arbitrage stocks and time periods. However, they do find that there may be some profitability despite the difficult-to-arbitrage positions identified.

Other research is related to the machine learning issue of overfitting data and raises questions about whether improved return predictability is possible. For example, Hou et al[17] replicate a large number of anomalies and find that many do not replicate. They also find that focusing primarily on larger stocks further reduces the anomalies that seem to exist. Chordia et al[121] use randomly generated trading strategies and find that the likelihood of concluding that an anomaly is real when in reality it is not (i.e., false positive) is around 45%. Harvey et al[18] propose higher t-statistic cut-offs and conclude that many of the findings in return predictability are likely false. McLean and Pontiff[122] find that the returns to published anomalies decline after the anomaly is published. Kim[123] finds that returns to accounting anomalies declined after the introduction of SEC EDGAR system. Green et al[14] and Smith and Timmermann[124] among others find that anomaly returns have weakened or disappeared in recent periods. Chen and Velikov[125] find that after controlling for trading costs, the returns to anomalies are very small.

In contrast, some papers push against the pessimistic view of return prediction. Chen[126] shows that it is unlikely the case that all published return predictors are spuriously generated. DeMiguel et al[127] find that adjusting for transaction costs increases the number of significant anomalies because some anomalies can help diversify anomalies against higher turnover and higher transaction cost trading. To summarise, research applying non-linear methods including machine learning seems to present some promise for improving earnings and return prediction; however, the results are not conclusive.

3.3.2. Coefficient instability

The final challenge we discuss is the challenge that, when estimated in one sample, the coefficients may not be good expectations for what would occur in other samples; in other words, the model coefficients may not be stable over time or across companies. Because spurious relations between the *X* and *Y* variables and heterogeneity in the relations between the *X* and *Y* variables can both cause unstable coefficients, coefficients that vary over time or across companies can make prediction difficult, particularly when a researcher cannot identify the cause of unstable coefficients.

First, some research has documented over-time variation in coefficients. For example, Bianchi et al[128] provide evidence that the number of predictors that matter to returns varies over time.

One reason for the time-varying patterns in how *X* variables predict returns is that investors learn and adapt their trading behaviors over time.[129–132] Second, other research finds in various forms that model coefficients vary across companies. For example, with respect to earnings forecasts, Dichev and Tang[133] find that firms with more volatile earnings also have less predictable earnings. For returns, Bathke et al[134] find that for companies that have no first-order autocorrelation in their earnings surprise, the post-earnings-announcement-drift is a reversal rather than a continuation of returns. Evgeniou et al[135] classify firms into groups using machine learning and find that the number of predictors varies across groups of firms.

[n] In addition to improving firm-level or portfolio-level returns forecast, new methods have been applied to better measure or extract aggregate risk factors. Chen et al[103] apply deep neural networks to explain returns and extract factors from returns. Lettau and Pelger[115] use principal component analysis, a latent variable approach with firm characteristics, to extract macro factors in stock returns. In a similar vein, Gu et al[116] and Kelly et al[28] treat macro factors as latent variables that are conditional on firm characteristics and find that their method better fits returns. Differently, Gu et al[116] use autoencoder neural networks while Kelly et al[28] use instrumental principal component analysis. Kim et al[117]; using projected principal component analysis, create portfolios that use firm characteristics to estimate both factor loadings and abnormal returns. Bianchi and McAlinn[118] apply ensemble learning method to forecast equity premia at the market and industry level, using accounting information. Liao and Liu[119] find that using information from errors-in-variables improves risk premia estimates relative to previous two-pass regression methods. Bandi et al[120] use spectral factor models to decrease dimensionality of the systematic risk factors while allowing them to vary across frequencies, thereby revealing more risk information specific to varying business cycles.

Moving from documenting to addressing coefficient instability, we discuss various approaches used in recent research. We first introduce across–company coefficient variation because this literature is smaller when considering the purposes of forecasting.

The research forecasting earnings that explores the possibility of company-varying coefficients seems to be larger than the research forecasting returns exploring the same possibility. Among this stream of research forecasting earnings, company-varying coefficients seem to improve prediction. For example, Fairfield et al[136] find that industry models are more accurate when forecasting growth but not profitability. Vorst and Yohn[137] find that models conditioned on a firm's life cycle stage improves the accuracy of out-of-sample profitability and growth forecasts. However, some research forecasting returns also examines if company-varying coefficients improve forecasting performance. For example, Bryzgalova et al[138] use decision trees to group firms. From groups of similar firms, they create optimal portfolios and form risk factors, and this method delivers higher out-of-sample Sharpe ratios compared to the conventional sorting methods.

The larger literature dealing with coefficient instability is primarily concerned with the challenges presented by unstable coefficients over time rather than across companies. Additionally, a major focus of the time-varying coefficients studies is for returns rather than earnings. We first explain the general modeling approach.

The common approach to applying trained models to out-of-sample predictions is to first use data prior to time t to estimate coefficients such as in Eq. (5). Second, the forecaster combines the average coefficient over the prior period with observed characteristics, X variables, at time t to forecast Y. If coefficients change over time but do so slowly, using coefficients estimated in a short window prior to time t may perform better than using all prior periods in the coefficient estimation period. On the other hand, if variation in coefficients is spurious (or unpredictable), averaging coefficients over a longer time period may perform better.

One stream of research finds that averaging coefficients over a longer time period better predicts returns (e.g. Brandt et al[139]; Coqueret[140]). Some research uses other ways to reduce the instability of the coefficients. For example, Ohlson and Kim[141] use a robust estimation approach and find that the inter-temporal stability of coefficients increases relative to OLS estimation. Another stream of research attempts to exploit predictable or useful variation in the coefficients. As an example of how this might be possible, Wang et al[142] find momentum in anomaly returns such that variation in these returns may be predictable. Other research also finds that predictability may depend on macroeconomics variables,[143] day of the week,[144] or aggregate sentiment.[145–147] When applied the time varying coefficients idea, Henrique et al[148] find that machine learning models that are updated more frequently may perform better than non-updated machine learning models for predicting short window returns.

Using short or long window estimation periods is not the only ways to address the challenge that coefficients are not stable. For example, other methods include cross-validation and ensemble method forecasting. Cross-validation trains model parameters by resampling training data to arrive at a model that may perform better when applied to new data sets. Most machine learning software packages include some form of cross validation for tuning models.[o] van Binsbergen et al[149] and de Silva and Thesmar[150] use cross-validation to create earnings forecasts. Another forecasting practice when dealing with model uncertainty is to use a combination of forecasts, i.e. an ensemble method to forecasting. Recent research generally finds that different versions of these ensemble methods improve return forecasting.[99,151–153]

Finally, other research imposes some constraints on the coefficient estimation process when forecasting earnings or returns. The constraints can take the form of economic-based expectations[154] or statistical limits.[155–157]

Together, there is substantial evidence that model coefficients vary over time and across firms. A growing set of studies examine the predictability and instability of predictions when applied to out-of-sample prediction. In our view, the evidence is mixed. Some research finds that limiting the variation in coefficients improves prediction while other research finds that allowing for time-varying or across-firm variation in coefficients improves prediction.

4. Discussion and conclusion

The literature on forecasting earnings and returns, already large to begin with, continues to grow. We review recent advancements in the forecasting of earnings and returns. We follow prior research that argues that these tasks should be considered together.

[o] For example, in R the caret package includes cross-validation as a tool for model training.

Despite the continuing growth and excitement over new data and new methodologies, the challenges for a forecaster are substantial. Following prior research, we highlight three major challenges for a forecaster when working with financial data: unpredictability of earnings and returns, noisy *X* variables, and model uncertainty. Using these challenges as a way to organize the literature, we discuss recent research that advances our collective ability to understand and predict the cross-sections of earnings and returns.

Here we reiterate some important insights from the literature. First, even with recent advancements, finding new meaningful predictors remains an important effort. Second, new out-of-the-box methods may have limited usefulness, but the thoughtful use of estimation methods and constraints seems to present promising opportunities. Third, it continues to be the case that finding earnings predictors that provide better forecasts than lagged earnings is challenging. Fourth, sorting through, combining, and understanding different models and methods likely has a long way to go before we achieve anything close to recommended best practices.

Finally, we mention questions that to us seem to be important and unanswered. First, why are the drivers of earnings and returns different? For example, why do aggregate means matter so much for returns but firm level information matter so much for earnings? Second, what are other ways to modify machine learning methods to exploit some of the unique features of financial data to better forecast returns and earnings? Third, are there combinations of methods and models that together work better than others? Fourth, are some methods that have been applied more for predicting earnings also applicable for returns or vice versa? Fifth, what does our ability or lack thereof to forecast earnings and returns tell us about the underlying economic mechanisms?

These and other questions remain important going forward. While we might view recent advancements skeptically, we join in the enthusiasm that new methods and new data will increase our ability to understand and to forecast financial data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table A1Variable definitions. Variables used in section 2.2 and Table 1.

Variable	Definition
acc	Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at); if oancf is missing then set to change in act - change in che - change in lct + change in dlc + change in txp-dp
agr	Annual percent change in total assets (at)
bm	Book value of equity (ceq) divided by end of fiscal year-end market capitalization
capxint	Capital expenditures from the cash flow statement divided by average total assets
capxgr	Percent change in capital expenditures from the cash flow statement
cash	Cash and cash equivalents divided by average total assets
cashdebt	Earnings before depreciation and extraordinary items (ib+dp) divided by avg. total liabilities (lt)
cashpr	Fiscal year-end market capitalization plus long-term debt (dltt) minus total assets (at) divided by cash and equivalents (che)
chato	Annual change in asset turnover (sale/average total assets)
chcurr	Annual change in the current ratio (act-lct)/total assets
chobklg	Annual change in obklg
chpm	Annual change in profit margin (ib/sale)
cf	Annual operating cash flows scaled by average total assets (at)
cfp	Operating cash flows divided by fiscal-year-end market capitalization
chinv	Change in inventory (inv) scaled by average total assets (at)
chtx	Percent change in total taxes (txtq) from quartert-4 to t
depr	Depreciation divided by PP&E
dy	Total dividends (dvt) divided by market capitalization at fiscal year-end
egr	Annual percent change in book value of equity (ceq)
ep	Annual income before extraordinary items (ib) divided by end of fiscal year market cap
gma	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)
grCAPX	Percent change in capital expenditures from year t-2 to year t
grgw	change in goodwill divided by average total assets (at)
grltnoa	Growth in long-term net operating assets

(*continued*)

Variable	Definition
hire	Percent change in number of employees (emp)
invest	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)
lev	Total liabilities (lt) divided by fiscal year-end market capitalization
lgr	Annual percent change in total liabilities (lt)
mom12m	11-month cumulative returns ending one month before month end
mom1m	1-month cumulative return
mom6m	5-month cumulative returns ending one month before month end

Table A1
Variable definitions (continued)

Variable	Definition
mve	Natural log of market capitalization at end of the most recent fiscal year end
mve_m	Natural log of market capitalization at end of the most recent calendar month
obklg	order backlog divided by average total assets (at)
operprof	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity
pchdepr	Percent change in depr
pctacc	Same as acc except that the numerator is divided by the absolute value of ib; if ib = 0 then ib set to 0.01 for denominator
pps	month-end price per share
ps	Sum of 9 indicator variables to form fundamental health score
quick	current assets minus current liabilities divided by average total assets
rd mve	R&D expense divided by end-of-fiscal-year market capitalization
rd sale	R&D expense divided by sales (xrd/sale)
rsup	change in revenue (revtq) from quartert-4 to t divided by average total assets
roa	ib/average total assets
roaq	ibq/average total assets
roe	ib/average common equity
roevol	standard deviation of roe for 8 quarters
roic	Annual earnings before interest and taxes (ebit) minus nonoperating income (nopi) divided by non-cash enterprise value (ceq+lt-che)
salecash	Annual sales divided by cash and cash equivalents
saleinv	Annual sales divided by total inventory
salerec	Annual sales divided by accounts receivable
shgr	Annual percent change in common shares (csho)
sgr	Annual percent change in sales (sale)
SP	Annual revenue (sale) divided by fiscal year-end market capitalization
sue	Unexpected quarterly earnings divided by fiscal-quarter-end market cap. Unexpected earnings is the seasonally differenced quarterly earnings before extraordinary items from Compustat quarterly file
tang	Cash holdings + 0.715 × receivables + 0.547 × inventory + 0.535 × PPE/total assets
turnmo	monthly share turnover (vol/shrout)
xadint	advertising expense (xad) divided by average total assets (at)
xrdint	advertising expense (xrd) divided by average total assets (at)

References

1. Richardson S, Tuna I, Wysocki P. Accounting anomalies and fundamental analysis: a review of recent research advances. *J Account Econ*. 2010;50(2–3):410–454.
2. Li KK, Mohanram P. Evaluating cross-sectional forecasting models for implied cost of capital. *Rev Account Stud*. 2014;19(3):1152–1185.
3. Gerakos J, Gramacy R. *Regression-based Earnings Forecasts*. 2013. Chicago Booth Research Paper (12-26).
4. Green J, Hand JR, Zhang XF. The supraview of return predictive signals. *Rev Account Stud*. 2013;18(3):692–730.
5. Campbell JY, Shiller RJ. Stock prices, earnings, and expected dividends. *J Finance*. 1988;43(3):661–676.
6. Abarbanell JS, Bushee BJ. Abnormal returns to a fundamental analysis strategy. *Account Rev*. 1998:19–45.
7. Ou JA, Penman SH. Financial statement analysis and the prediction of stock returns. *J Account Econ*. 1989;11(4):295–329.
8. Nissim D, Penman SH. Ratio analysis and equity valuation: from research to practice. *Rev Account Stud*. 2001;6(1):109–154.
9. Monahan SJ. Financial statement analysis and earnings forecasting. *Foundations and Trends® in Accounting*. 2018;12(2):105–215.
10. Timmermann A. Forecasting methods in finance. *Ann Rev Financial Econom*. 2018;10:449–479.

11. Petropoulos F, Apiletti D, Assimakopoulos V, et al. Forecasting: theory and practice. *Int J Forecast*. 2022. In Press (available online).
12. Karolyi GA, Van Nieuwerburgh S. New methods for the cross-section of returns. *Rev Financ Stud*. 2020;33(5):1879–1890.
13. Weigand A. Machine learning in empirical asset pricing. *Financ Mark Portfolio Manag*. 2019;33(1):93–104.
14. Green J, Hand JR, Zhang XF. The characteristics that provide independent information about average us monthly stock returns. *Rev Financ Stud*. 2017;30(12):4389–4436.
15. Baba Yara F. *Machine Learning and Return Predictability across Firms, Time and Portfolios*. 2020. Available at SSRN 3696533.
16. Avramov D, Cheng S, Metzker L. *Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability*. 2021. Available at SSRN 3450322.
17. Hou K, Xue C, Zhang L. Replicating anomalies. *Rev Financ Stud*. 2020;33(5):2019–2133.
18. Harvey CR, Liu Y, Zhu H.... and the cross-section of expected returns. *Rev Financ Stud*. 2016;29(1):5–68.
19. Cochrane JH. Presidential address: discount rates. *J Finance*. 2011;66(4):1047–1108.
20. Fama EF, MacBeth JD. Risk, return, and equilibrium: empirical tests. *J Polit Econ*. 1973;81(3):607–636.
21. Lewellen J. The cross-section of expected stock returns. *Crit Finance Rev*. 2015;4(1):1–44.
22. Jacobs BI, Levy KN. Investing in a multidimensional market. *Financ Anal J*. 2014;70(6):6–12.
23. Lev B, Thiagarajan SR. Fundamental information analysis. *J Account Res*. 1993;31(2):190–215.
24. Avramov D, Kaplanski G, Subrahmanyam A. *Stock Return Predictability: New Evidence from Moving Averages of Prices and Firm Fundamentals*. 2018. Available at SSRN 3111334.
25. Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. *Rev Financ Stud*. 2020;33(5):2223–2273.
26. Ayers B, Freeman RN. Market assessment of industry and firm earnings information. *J Account Econ*. 1997;24(2):205–218.
27. Freyberger J, Neuhierl A, Weber M. Dissecting characteristics nonparametrically. *Rev Financ Stud*. 2020;33(5):2326–2377.
28. Kelly BT, Pruitt S, Su Y. Characteristics are covariances: a unified model of risk and return. *J Financ Econ*. 2019;134(3):501–524.
29. Light N, Maslov D, Rytchkov O. Aggregation of information about the cross section of stock returns: a latent variable approach. *Rev Financ Stud*. 2017;30(4):1339–1381.
30. Clarke C. The level, slope, and curve factor model for stocks. *J Financ Econ*. 2022;143(1):159–187.
31. Campbell JY, Thompson SB. Predicting excess stock returns out of sample: can anything beat the historical average? *Rev Financ Stud*. 2008;21(4):1509–1531.
32. Shiller RJ. Do stock prices move too much to be justified by subsequent changes in dividends?: Reply. *Am Econ Rev*. 1983;73(1):236–237.
33. Marsh TA, Merton RC. Dividend variability and variance bounds tests for the rationality of stock market prices. *Am Econ Rev*. 1986;76(3):483–498.
34. Shiller RJ. Comovements in stock prices and comovements in dividends. *J Finance*. 1989;44(3):719–729.
35. Change T, Change F. Do stock prices move too much to be justified by subsequent changes in dividends? Comment. *Am Econ Rev*. 1983;73(1):234–235.
36. Shiller RJ. The volatility of stock market prices. *Science*. 1987;235(4784):33–37.
37. Malkiel BG. Is the stock market efficient? *Science*. 1989;243(4896):1313–1318.
38. De Bondt WF, Thaler R. Does the stock market overreact? *J Finance*. 1985;40(3):793–805.
39. De Bondt WF, Thaler RH. Further evidence on investor overreaction and stock market seasonality. *J Finance*. 1987;42(3):557–581.
40. Timmermann AG. How learning in financial markets generates excess volatility and predictability in stock prices. *Q J Econ*. 1993;108(4):1135–1145.
41. De Bondt WF, Thaler RH. Do security analysts overreact? *Am Econ Rev*. 1990:52–57.
42. Lundholm RJ, Rogo R. Do analyst forecasts vary too much? *J Financial Rep*. 2016;1(1):101–123.
43. Lundholm R, Rogo R. Do excessively volatile forecasts impact investors? *Rev Account Stud*. 2020;25(2):636–671.
44. Haddad V, Kozak S, Santosh S. *Predicting Relative Returns*. 2017. Working Paper.
45. De Raaij G, Raunig B. Evaluating density forecasts from models of stock market returns. *Eur J Finance*. 2005;11(2):151–166.
46. Taylor JW. Estimating value at risk and expected shortfall using expectiles. *J Financ Econom*. 2008;6(2):231–252.
47. Timmerman A. Editorial: density forecasting in economics and finance. *J Forecast*. 2000;19(4):231–234.
48. Kokoszka P, Miao H, Petersen A, Shang HL. Forecasting of density functions with an application to cross-sectional and intraday returns. *Int J Forecast*. 2019;35(4):1304–1317.
49. Liu L. Density forecasts in panel data models: a semiparametric bayesian perspective. *J Bus Econ Stat*. 2021:1–42 (just-accepted).
50. Koenker R, Bassett Jr G. Regression quantiles. *Econometrica: J Econom Soc*. 1978:33–50.
51. Koenker R, Hallock KF. Quantile regression. *J Econ Perspect*. 2001;15(4):143–156.
52. Koenker R. Quantile regression: 40 years on. *Annu Rev Econom*. 2017;9:155–176.
53. Kelly B, Jiang H. Tail risk and asset prices. *Rev Financ Stud*. 2014;27(10):2841–2871.
54. Amaya D, Christoffersen P, Jacobs K, Vasquez A. Does realized skewness predict the cross-section of equity returns? *J Financ Econ*. 2015;118(1):135–167.
55. Ross S. The recovery theorem. *J Finance*. 2015;70(2):615–648.
56. Jackwerth JC, Menner M. Does the ross recovery theorem work empirically? *J Financ Econ*. 2020;137(3):723–739.
57. Audrino F, Huitema R, Ludwig M. An empirical implementation of the ross recovery theorem as a prediction device. *J Financ Econom*. 2021;19(2):291–312.
58. Chen M-Y, Chen J-E. Application of quantile regression to estimation of value at risk. *Rev Financial Risk Manag*. 2002;1(2):15.
59. Gaglianone WP, Lima LR, Linton O, Smith DR. Evaluating value-at-risk models via quantile regression. *J Bus Econ Stat*. 2011;29(1):150–160.

60. Gowlland C, Xiao Z, Zeng Q. Beyond the central tendency: quantile regression as a tool in quantitative investing. *J Portfolio Manag*. 2009;35(3):106–119.
61. Ma L, Pohlman L. Return forecasts and optimal portfolio construction: a quantile regression approach. *Eur J Finance*. 2008;14(5):409–425.
62. Pohlman L, Ma L. Return forecasting by quantile regression. *J Invest*. 2010;19(4):116–121.
63. Konstantinidi T, Pope PF. Forecasting risk in earnings. *Contemp Account Res*. 2016;33(2):487–525.
64. Chang W-J, Monahan SJ, Ouazad A, Vasvari FP. The higher moments of future earnings. *Account Rev*. 2021;96(1):91–116.
65. Tian H, Yim A, Newton DP. Tail-heaviness, asymmetry, and profitability forecasting by quantile regression. *Manag Sci*. 2021;67(8):5209–5233.
66. Hendriock M. *Forecasting Earnings with Predicted, Conditional Probability Density Functions*. 2021. Available at SSRN 3901386.
67. Easton PD, Monahan SJ. Review of recent research on improving earnings forecasts and evaluating accounting-based estimates of the expected rate of return on equity capital. *Abacus*. 2016;52(1):35–58.
68. He S, Narayanamoorthy GG. Earnings acceleration and stock returns. *J Account Econ*. 2020;69(1):101238.
69. Avramov D, Kaplanski G, Subrahmanyam A. *Post-fundamentals Drift in Stock Prices: A Regression Regularization Perspective*. 2020. Available at SSRN 3507512.
70. Lyle MR, Yohn TL. Fundamental analysis and mean-variance optimal portfolios. *Account Rev*. 2021;96(6):303–327.
71. Azevedo V, Bielstein P, Gerhart M. Earnings forecasts: the case for combining analysts' estimates with a cross-sectional model. *Rev Quant Finance Account*. 2021;56(2):545–579.
72. Burger M, Curtis A. Aggregate margin debt and the divergence of price from accounting fundamentals. *Contemp Account Res*. 2017;34(3):1418–1445.
73. Han Y, Huang D, Huang D, Zhou G. Expected return, volume, and mispricing. *J Financ Econ*. 2021;143(3):1295–1315.
74. Loughran T, McDonald B. Textual analysis in accounting and finance: a survey. *J Account Res*. 2016;54(4):1187–1230.
75. Guo L, Shi F, Tu J. Textual analysis and machine leaning: crack unstructured data in finance and accounting. *J Finance Data Sci*. 2016;2(3):153–170.
76. Das SR, Donini M, Zafar MB, He J, Kenthapadi K. Finlex: an effective use of word embeddings for financial lexicon generation. *J Finance Data Sci*. 2022;8:1–11.
77. Li F. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *J Account Res*. 2010;48(5):1049–1102.
78. Huang AH, Zang AY, Zheng R. Evidence on the information content of text in analyst reports. *Account Rev*. 2014;89(6):2151–2180.
79. Karapandza R. Stock returns and future tense language in 10-K reports. *J Bank Finance*. 2016;71:50–61.
80. Meursault V, Liang PJ, Routledge B, Scanlon M. *PEAD.txt: Post-earnings announcement Drift Using Text*. 2021. Available at SSRN 3778798.
81. Heston SL, Sinha NR. News vs. sentiment: predicting stock returns from news stories. *Financ Anal J*. 2017;73(3):67–83.
82. Ke ZT, Kelly BT, Xiu D. *Predicting Returns with Text Data*. 2019. Working Paper.
83. Bochkay K, Levine CB. Using MD&A to improve earnings forecasts. *J Account Audit Finance*. 2019;34(3):458–482.
84. Obaid K, Pukthuanthong K. A picture is worth a thousand words: measuring investor sentiment by combining machine learning and photos from news. *J Financ Econ*. 2021;144(1):273–297.
85. deHaan E, Madsen J, Piotroski JD. Do weather-induced moods affect the processing of earnings news? *J Account Res*. 2017;55(3):509–550.
86. Jame R, Johnston R, Markov S, Wolfe MC. The value of crowdsourced earnings forecasts. *J Account Res*. 2016;54(4):1077–1110.
87. Ball R, Gerakos J, Linnainmaa JT, Nikolaev VV. Deflating profitability. *J Financ Econ*. 2015;117(2):225–248.
88. Ball R, Gerakos J, Linnainmaa JT, Nikolaev V. Accruals, cash flows, and operating profitability in the cross section of stock returns. *J Financ Econ*. 2016;121(1):28–45.
89. Ball R, Gerakos J, Linnainmaa T, Nikolaev V. Book-to-market, retained earnings, and earnings in the cross section of stock returns. *J Finance Econ*. 2019;1(1):1–24.
90. Cooper MJ, Gulen H, Schill MJ. Asset growth and the cross-section of stock returns. *J Finance*. 2008;63(4):1609–1651.
91. Mohanram P, Gode D. Removing predictable analyst forecast errors to improve implied cost of equity estimates. *Rev Account Stud*. 2013;18(2):443–478.
92. Freyberger J, Höppner B, Neuhierl A, Weber M. *Missing Data in Asset Pricing Panels*. 2021. Available at SSRN 3932438.
93. Feng G, Giglio S, Xiu D. Taming the factor zoo: a test of new factors. *J Finance*. 2020;75(3):1327–1370.
94. Chinco A, Neuhierl A, Weber M. Estimating the anomaly base rate. *J Financ Econ*. 2021;140(1):101–126.
95. He A, Huang D, Li J, Zhou G. *Shrinking Factor Dimension: A Reduced-Rank Approach*. 2019. Available at SSRN 3205697.
96. Sun C. *Dissecting the Factor Zoo: A Correlation-Robust Machine Learning Approach*. 2020. Available at SSRN 3263420.
97. Kozak S, Nagel S, Santosh S. Shrinking the cross-section. *J Financ Econ*. 2020;135(2):271–292.
98. Dong MM. *Global Anomalies*. 2018. Working Paper.
99. Elliott G, Gargano A, Timmermann A. Complete subset regressions. *J Econom*. 2013;177(2):357–373.
100. Müller K, Schmickler S. *Interacting Anomalies*. 2020. Available at SSRN 3646417.
101. Caldeira J, AP Santos A, Torrent H. *Semiparametric Portfolio Policies*. 2021. Available at SSRN 3830435.
102. Anand V, Brunner R, Ikegwu K, Sougiannis T. *Predicting Profitability Using Machine Learning*. 2019. Available at SSRN 3466478.
103. Chen X, Cho YH, Dou Y, Lev B. *Fundamental Analysis of XBRL Data: A Machine Learning Approach*. 2020. Available at SSRN 3741015.
104. Chen L, Pelger M, Zhu J. *Deep Learning in Asset Pricing*. 2020. Available at SSRN 3350138.
105. Cao K, You H. *Fundamental Analysis via Machine Learning*. 2020. Working Paper.
106. Hansen JW, Thimsen C. *Forecasting Corporate Earnings with Machine Learning*. 2021. Working Paper.
107. Frank MZ, Yang K. *Predicting Firm Profits: From Fama-Macbeth to Gradient Boosting*. 2021. Available at SSRN 3919194.

108. Easton PD, Kapons M, Monahan SJ, Schütt HH, Weisbrod EH. *Forecasting Earnings Using K-Nearest Neighbor Matching*. 2020. Available at SSRN 3752238.
109. Elamir EA. Boosting algorithms to analyse firm's performance based on return on equity: an explanatory study. *Int J Comput Digit Syst*. 2020;10:1–17.
110. Azevedo V, Hoegner C. *Enhancing Stock Market Anomalies with Machine Learning*. 2020. Available at SSRN 3752741.
111. Luss R, d'Aspremont A. Predicting abnormal returns from news using text classification. *Quant Finance*. 2015;15(6):999–1012.
112. Tobek O, Hronec M. Does it pay to follow anomalies research? Machine learning approach with international evidence. *J Financ Mark*. 2021;56:100588.
113. Choi D, Jiang W, Zhang C. *Alpha Go Everywhere: Machine Learning and International Stock Returns*. 2021. Available at SSRN 3489679.
114. Cong LW, Tang K, Wang J, Zhang Y. *Alphaportfolio: Direct Construction through Deep Reinforcement Learning and Interpretable Ai*. 2021. Available at SSRN 3554486.
115. Lettau M, Pelger M. Factors that fit the time series and cross-section of stock returns. *Rev Financ Stud*. 2020;33(5):2274–2325.
116. Gu S, Kelly B, Xiu D. Autoencoder asset pricing models. *J Econom*. 2021;222(1):429–450.
117. Kim S, Korajczyk RA, Neuhierl A. Arbitrage portfolios. *Rev Financ Stud*. 2021;34(6):2813–2856.
118. Bianchi D, McAlinn K. *Divide and Conquer: Financial Ratios and Industry Returns Predictability*. 2020. Available at SSRN 3136368.
119. Liao Z, Liu Y. *Optimal Cross-Sectional Regression*. 2021. Available at SSRN 3719299.
120. Bandi FM, Chaudhuri SE, Lo AW, Tamoni A. Spectral factor models. *J Financ Econ*. 2021;142(1):214–238.
121. Chordia T, Goyal A, Saretto A. Anomalies and false rejections. *Rev Financ Stud*. 2020;33(5):2134–2179.
122. McLean RD, Pontiff J. Does academic research destroy stock return predictability? *J Finance*. 2016;71(1):5–32.
123. Kim YH. *Do Information Acquisition Costs Matter? the Effect of SEC EDGAR on Stock Anomalies*. 2021. Available at SSRN 3921785.
124. Smith S, Timmermann A. *Have Risk Premia Vanished?*. 2021. Available at SSRN 3846221.
125. Chen AY, Velikov M. *Zeroing in on the Expected Returns of Anomalies*. 2019. Available at SSRN 3073681.
126. Chen AY. The limits of p-hacking: some thought experiments. *J Finance*. 2021;76(5):2447–2480.
127. DeMiguel V, Martin-Utrera A, Nogales FJ, Uppal R. A transaction-cost perspective on the multitude of firm characteristics. *Rev Financ Stud*. 2020;33(5):2180–2222.
128. Bianchi D, Büchner M, Tamoni A. *What Matters when? Time-Varying Sparsity in Expected Returns*. 2019. WBS Finance Group Research Paper.
129. Milian JA. Unsophisticated arbitrageurs and market efficiency: overreacting to a history of underreaction? *J Account Res*. 2015;53(1):175–220.
130. Zaremba A, Umutlu M, Maydybura A. Where have the profits gone? Market efficiency and the disappearing equity anomalies in country and industry returns. *J Bank Finance*. 2020;121:105966.
131. Jacobs H, Müller S. Anomalies across the globe: once public, no longer existent? *J Financ Econ*. 2020;135(1):213–230.
132. Farmer L, Schmidt L, Timmermann A. *Pockets of Predictability*. 2019. Available at SSRN 3152386.
133. Dichev ID, Tang VW. Earnings volatility and earnings predictability. *J Account Econ*. 2009;47(1–2):160–181.
134. Bathke AW, Mason TW, Morton RM. Investor overreaction to earnings surprises and post-earnings-announcement reversals. *Contemp Account Res*. 2019;36(4):2069–2092.
135. Evgeniou T, Guecioueur A, Prieto R. *Uncovering Sparsity and Heterogeneity in Firm-Level Return Predictability Using Machine Learning*. 2021. Available at SSRN 3604921.
136. Fairfield PM, Ramnath S, Yohn TL. Do industry-level analyses improve forecasts of financial performance? *J Account Res*. 2009;47(1):147–178.
137. Vorst P, Yohn TL. Life cycle models and forecasting growth and profitability. *Account Rev*. 2018;93(6):357–381.
138. Bryzgalova S, Pelger M, Zhu J. *Forest through the Trees: Building Cross-Sections of Stock Returns*. 2020. Available at SSRN 3493458.
139. Brandt MW, Santa-Clara P, Valkanov R. Parametric portfolio policies: exploiting characteristics in the cross-section of equity returns. *Rev Financ Stud*. 2009;22(9):3411–3447.
140. Coqueret G. Persistence in factor-based supervised learning models. *J Finance Data Sci*. 2022;8:12–34.
141. Ohlson JA, Kim S. Linear valuation without OLS: the Theil-Sen estimation approach. *Rev Account Stud*. 2015;20(1):395–435.
142. Wang F, Yan X, Zheng L. Time-series and cross-sectional momentum in anomaly returns. *Eur Financ Manag*. 2021;27(4):736–771.
143. Favero CA, Melone A, Tamoni A. *Macro Trends and Factor Timing*. 2021. Available at SSRN 3940452.
144. Birru J. Day of the week and the cross-section of returns. *J Financ Econ*. 2018;130(1):182–214.
145. Stambaugh RF, Yu J, Yuan Y. The long of it: odds that investor sentiment spuriously predicts anomaly returns. *J Financ Econ*. 2014;114(3):613–619.
146. Stambaugh RF, Yu J, Yuan Y. The short of it: investor sentiment and anomalies. *J Financ Econ*. 2012;104(2):288–302.
147. Jacobs H. What explains the dynamics of 100 anomalies? *J Bank Finance*. 2015;57:65–85.
148. Henrique BM, Sobreiro VA, Kimura H. Stock price prediction using support vector regression on daily and up to the minute prices. *J Finance Data Sci*. 2018;4(3):183–201.
149. van Binsbergen JH, Han X, Lopez-Lira A. *Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases*. 2020. Working Paper.
150. de Silva T, Thesmar D. *Noise in Expectations: Evidence from Analyst Forecasts*. 2021. Working Paper.
151. De Nard G, Hediger S, Leippold M. *Subsampled Factor Models for Asset Pricing: The Rise of Vasa*. 2020. Available at SSRN 3557957.
152. Han Y, He A, Rapach D, Zhou G. *What Firm Characteristics Drive Us Stock Returns*. 2018. Working Paper.
153. Zhang H. *Empirical Asset Pricing and Ensemble Machine Learning*. 2021. Working Paper.
154. Pettenuzzo D, Timmermann A, Valkanov R. Forecasting stock returns under economic constraints. *J Financ Econ*. 2014;114(3):517–553.

155. Han Y, He A, Rapach D, Zhou G. *Firm Characteristics and Expected Stock Returns*. 2020. Available at SSRN 3185335.

156. DeMiguel V, Garlappi L, Nogales FJ, Uppal R. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Manag Sci*. 2009;55(5):798–812.

157. Evans ME, Njoroge K, Yong KO. An examination of the statistical significance and economic relevance of profitability and earnings forecasts from models and analysts. *Contemp Account Res*. 2017;34(3):1453–1488.