THE ACCOUNTING REVIEW
Vol. 99, No. 3
May 2024
pp. 115–140

American Accounting Association
DOI: 10.2308/TAR-2021-0478

Forecasting Earnings Using *k*-Nearest Neighbors

Peter D. Easton
University of Notre Dame

Martin M. Kapons
University of Amsterdam

Steven J. Monahan
The University of Utah

Harm H. Schütt
Tilburg University

Eric H. Weisbrod
The University of Kansas

ABSTRACT: We use a simple *k*-nearest neighbors algorithm (hereafter, k-NN*) to forecast earnings. k-NN* forecasts of one-, two-, and three-year-ahead earnings are more accurate than those generated by popular extant forecasting approaches. k-NN* forecasts of two- and three-year (one-year)-ahead *EPS* and aggregate three-year *EPS* are more (less) accurate than those generated by analysts. The association between the unexpected earnings implied by k-NN* and the contemporaneous market-adjusted return (i.e., the earnings association coefficient (EAC)) is positive and exceeds the EAC on unexpected earnings implied by alternate approaches. A trading strategy that is long (short) firms for which k-NN* predicts positive (negative) earnings growth earns positive risk-adjusted returns that exceed those earned by similar trading strategies that are based on alternate forecasts. The k-NN* algorithm generates an empirically reliable *ex ante* indicator of forecast accuracy that identifies situations when the k-NN* EAC is larger and the k-NN* trading strategy is more profitable.

Data Availability: Data are available from the public sources described in the text.

JEL Classifications: C21; C53; G17; M41.

Keywords: earnings; forecasting; machine learning; nearest neighbors.

We thank Daniel J. Taylor (editor), two anonymous reviewers, Jeff Coulton, Michael Ettredge, Elia Ferracuti (discussant), Stephan Hollander, Andrew Jackson, Philip Joos, Stephannie Larocque, Richard Morris, Erik Peek, Annelies Renders, Arthur Stenzel, Stephen Taylor, Andrew Yim, David Veenman, and Xiao-Jun Zhang for helpful comments and suggestions. We also thank workshop participants at the Accounting Design Project, the 2020 Financial Accounting and Reporting Section Midyear Meeting, and the 2021 European Accounting Association Annual Congress and City, University of London, Bayes Business School; Egyptian Online Seminars in Business, Accounting, and Economics; UNSW Sydney; University of St. Gallen; HEC Montreal; Tilburg University; and The University of Kansas for helpful comments on earlier drafts.

Peter D. Easton, University of Notre Dame, Mendoza College of Business, Department of Accountancy, Notre Dame, IN, USA; Martin M. Kapons, University of Amsterdam, Amsterdam Business School, Accounting Section, Amsterdam, Netherlands; Steven J. Monahan, The University of Utah, David Eccles School of Business, School of Accounting, Salt Lake City, UT, USA; Harm H. Schütt, Tilburg University, Tilburg School of Economics and Management, Department of Accountancy, Tilburg, Netherlands; Eric H. Weisbrod, The University of Kansas, School of Business, Accounting Academic Area, Lawrence, KS, USA.

Editor's note: Accepted by Daniel J. Taylor, under the Senior Editorship of W. Robert Knechel.

Submitted: July 2021
Accepted: October 2023
Early Access: November 2023

I. INTRODUCTION

Earnings forecasting is important. It is at the heart of equity valuation. Lenders evaluate earnings forecasts when assessing creditworthiness and negotiating covenants. Managers form earnings expectations when making investment decisions, developing budgets, writing contracts, etc. Consequently, earnings forecasts are a key variable of interest in many accounting and finance studies, and a number of studies propose and evaluate different forecasting approaches. Nonetheless, several fundamental issues are not well understood. In this study, we consider two of these issues: the use and usefulness of comparable-firm-based forecasts and the linearity assumption that underpins most extant forecasting approaches.

Most extant studies of earnings forecasting sweep aside issues relating to the choice and use of comparable firms. In fact, we are aware of only three studies that center the analysis around these issues: Fairfield, Ramnath, and Yohn (2009; hereafter, FRY), Vorst and Yohn (2018; hereafter, VY), and Blouin, Core, and Guay (2010; hereafter, BCG). VY (FRY) show that regression-based forecasts are more (less) accurate when the training data are restricted to firms in the same lifecycle (industry) group. BCG, who build on the study by Barber and Lyon (1996), show that forecasts implied by groups of firms that are matched to the subject firm on the basis of profitability and size are more accurate than forecasts generated by the random walk. We address two key unanswered questions within this genre: how do forecasts based on comparable firms perform relative to one another and to alternate approaches? Are forecasts based on comparable firms useful in the context of security analysis?

We use *k*-nearest neighbors (i.e., k-NN) to study these questions. k-NN is a simple, effective, longstanding, nonparametric forecasting approach.[1] It involves matching the subject firm-year to a set of nearest neighbors (on the basis of one or more predictors) and then basing the forecast of the subject firm-year's earnings on the nearest neighbors' earnings.[2] Hence, k-NN is a natural way of (1) integrating comparable firms into the earnings forecasting process and (2) avoiding the linearity assumption that underpins most extant forecasting approaches.

We base our approach for identifying a subject firm-years' nearest neighbors on an intuitive argument: earnings reflect economic performance, and firms with similar past performance are more likely to perform similarly in the future. That is, earnings is likely to be a key predictor. Consequently, the primary k-NN algorithm that we study (hereafter, k-NN*) matches each subject firm-year to firm-years with similar earnings histories.[3] Then, we set the forecast equal to the median of the lead earnings for the matched firm-years. To assure that our forecasts are out of sample, we require that matched firm-years precede the subject firm-year by at least h years (h is the length of the forecast horizon).[4]

We begin our empirical analyses by establishing three basic properties of the k-NN* algorithm. First, we "tune" two key parameters: (1) the length of the earnings history, M, that we use to find matches (we allow M to vary between one and five years) and (2) the number of nearest neighbors, k, that we match to each subject firm-year. To avoid look-ahead bias, we use the "tuning" sample, which contains subject firm-years drawn from 1978 to 1994. (Our holdout, or "testing" sample, contains firm-years drawn from 1998 to 2018.) Using minimum mean absolute forecast error (i.e., *MAFE*) as our criterion, we find that the optimal value of M is two and the optimal value of k is 120. Hence, learning more about each firm is less important than learning about more firms.

Second, we compare the k-NN* algorithm with four "naive" approaches, in which we set the forecast equal to (1) the subject firm-year's current earnings (i.e., the random walk forecast), (2) the median of the nearest neighbors' matched-year (instead of lead) earnings, (3) the median of lead earnings for all firm-years ending during the ten-year period that ends h years before the subject firm-year, and (4) the median of lead earnings for a sample of k randomly selected firm-years. We find that, regardless of the value of k, the k-NN* algorithm is superior to all of the naive approaches. This implies that selecting *nearest* neighbors (instead of *all* or *random* neighbors) and using their *lead* earnings instead of their *matched-year* earnings are both important determinants of the performance of k-NN*.

Third, we compare the subject firm-years in our sample with the nearest neighbors chosen by the k-NN* algorithm. Per the Fama-French 12 industries classification scheme, the average percentage of nearest neighbors in the same

[1] The practice of finding similar histories, or "*k*-nearest neighbors," on which to base predictions appears in texts dating back as early as the 11th century (Chen and Shah 2018). Modern applications include forecasting a baseball player's future performance by comparison to similar players (Silver 2003) and forecasting a state's election results by incorporating polling trends from similar states (Silver 2008).

[2] The nearest neighbors consist of firm-years that precede the subject firm-year by a sufficient number of years to assure that the forecasts are out of sample.

[3] We find matches on the basis of earnings deflated by equity market value. We evaluate alternate deflators, such as equity book value, total assets, and revenues. In a set of untabulated results, we find that our results regarding relative (to other approaches) forecast accuracy do not depend on the choice of deflator.

[4] For example, when forecasting one- (two-) year-ahead earnings for a subject firm in 2010, the most recent year from which a matched firm-year can be drawn is 2009 (2008). This ensures that the "lead" earnings of the matched firm-year (i.e., the matched firm-year's 2010 earnings) are observable at the time the forecast is made. Specifically, we search for matches from the set of all firm-years ending within the ten-year period that ends h years before the subject firm-year, and we include the subject's prior firm-years in this set.

industry as the subject firm-year is 17 percent, and when industry is defined on the basis of two-digit SIC code (Global Industry Classification Standard (GICS) code), this amount falls to 7.7 (4.4) percent. The average percentage of nearest neighbors in the same lifecycle group, as defined in Dickinson (2011), (size decile) as the subject firm is 37 (13.6) percent. And the average (median) difference between the subject firm-year and the nearest neighbor firm-years is 5.5 (6) years. Hence, the matching approach underlying the k-NN* algorithm is noticeably different from conventional approaches in which the subject firm-year is matched to its contemporaries in the same industry, same lifecycle group, or same size stratum. Whether this difference leads to more accurate forecasts is the empirical question that we evaluate next.

We compare the k-NN* forecasts with a broad set of forecasts, including forecasts generated by (1) k-NN algorithms that find matches using alternate (e.g., profit margin and asset turnover) or additional (e.g., accruals) predictors; (2) the comparable-firm-based models developed by FRY, VY, and BCG; and (3) the popular regression models described in Hou, van Dijk, and Zhang (2012; hereafter, HVZ) and Li and Mohanram (2014; hereafter, LM). We use the testing sample and show that the k-NN* forecasts are more accurate than all of the alternate forecasts. This result is robust. It holds regardless of the forecast horizon—i.e., one, two, or three years—and regardless of the criterion that we use to evaluate accuracy—i.e., *MAFE*, median absolute forecast error (i.e., *MDAFE*), or mean square error (i.e., *MSE*). We also compare the k-NN* forecasts of earnings per share, *EPS*, with those generated by analysts. We find that analysts' year-ahead *EPS* forecasts are more accurate than those generated by the k-NN* algorithm. However, when horizons of either two or three years are considered, the k-NN* forecasts are superior. Moreover, the k-NN* forecasts of aggregate three-year *EPS* are superior to those generated by analysts. Hence, from an overall perspective, the k-NN* algorithm outperforms analysts.

Next, we evaluate the cross-sectional variation in the accuracy of the k-NN* forecasts. A key feature of the k-NN* algorithm is that it generates a measure of the dispersion of the nearest neighbors' lead earnings around the point forecast. We examine the usefulness of this feature by computing the variable $MAD_{i,t}$, which equals the median absolute deviation of the nearest neighbors' lead earnings.[5] We find that $MAD_{i,t}$ has a strong, negative relation with both the absolute and relative accuracy of k-NN* forecasts. We also find that, in terms of explanatory power, $MAD_{i,t}$ dominates other popular indicators of forecast accuracy, such as analyst following, the book-to-market ratio, equity market value, etc. Hence, we conclude that $MAD_{i,t}$ is a novel and reliable *ex ante* indicator of forecast accuracy.

In our final set of tests, we assess the usefulness of k-NN* forecasts within the context of security analysis. We do two sets of tests. In the first set, we evaluate the slope coefficient from a regression of market-adjusted returns for year t+1 on the contemporaneous unexpected earnings implied by the forecasts. We refer to this coefficient as the earnings association coefficient (i.e., the EAC). We show that the EAC on the unexpected earnings implied by the k-NN* algorithm is positive and larger than the EACs on the unexpected earnings implied by the alternate approaches. We also find that the EAC on the unexpected earnings implied by the k-NN* algorithm decreases as $MAD_{i,t}$ increases. In the second set of tests, we implement a trading strategy in which we take long (short) positions in stocks for which the k-NN* forecasts predict positive (negative) year-ahead earnings growth. We find that the risk-adjusted returns earned by this strategy are positive, decreasing in $MAD_{i,t}$, and larger than those earned by similar strategies that are based on forecasts from alternate approaches. Taken together, these results imply that (1) k-NN* forecasts are useful within the context of security analysis, (2) they have incremental usefulness *vis-à-vis* forecasts from alternate approaches, and (3) the variable $MAD_{i,t}$ provides investors with useful information about when they should put more (less) weight on the k-NN* forecasts.

We make three contributions. First, we develop a k-NN-based forecasting approach (i.e., k-NN*) that we argue is the new benchmark because it (1) is easy to understand, easy to use, easy to explain, and easy to modify; (2) outperforms competing approaches; and (3) naturally self-assesses via the variable $MAD_{i,t}$, which is a useful *ex ante* indicator of forecast accuracy.

Second, we provide initial evidence about the usefulness of k-NN within the context of earnings forecasting. A key advantage of k-NN is that it combines the subject firm-year's historical performance with the historical performance of the neighbor firm-years in a nonparametric manner. Consequently, compared with regression models, k-NN accommodates nonlinearities and is less sensitive to extreme values. Moreover, k-NN is a natural and objective way of integrating comparable firms into the forecasting process.

Finally, we make a methodological contribution regarding the best way of identifying matched firms/samples, which is an important research design issue. Many studies (e.g., FRY and VY) assume that (1) it is desirable to match on structural factors that are either unobservable or difficult to measure and (2) these factors vary systematically across

[5] $MAD_{i,t}$ is the median of the absolute values of the differences between each nearest neighbor's lead earnings and the median of the nearest neighbors' lead earnings. As discussed in Arachchige, Prendergast, and Staudte (2022), $MAD_{i,t}$ is a superior measure of dispersion for skewed and/or leptokurtotic distributions, which describes our data.

industries, lifecycle groups, or size strata. Hence, matching on industry, lifecycle, or size is common. However, we find that our simple, performance-based matching algorithm in which we identify firms with similar two-year earnings histories is better. We also find that the curse of dimensionality is real: matching on longer histories and/or additional predictors typically leads to *worse* forecasts.

II. BACKGROUND AND RESEARCH QUESTIONS

Longstanding literature focuses on the use of historical accounting data to forecast earnings. Early studies (e.g., Ball and Watts 1972; Albrecht, Lookabill, and McKeown 1977; Watts and Leftwich 1977) evaluate autoregressive integrated moving average (i.e., ARIMA) models. Although ARIMA models are useful for forecasting many time-series processes, they have a number of limitations within the context of forecasting annual earnings. We focus on three limitations that are especially relevant to our study: (1) ARIMA models only consider the information contained in the subject firm's historical earnings, (2) they ignore information contained in comparable firms' financial statements, and (3) they embed the assumption that there is a linear relation between future earnings and historical earnings.

Linear panel regressions are not subject to the first two limitations: they accommodate multiple predictors, and they can be trained on a sample of comparable firms. Consequently, linear panel regression models have displaced ARIMA models. For example, the HVZ model is widely adopted and is often referred to as the benchmark approach (e.g., Evans, Njoroge, and Yong 2017; So 2013). In addition, the earnings persistence (EP) model described in LM is also popular and it forms the basis for the industry- and lifecycle-based forecasts described in FRY and VY, respectively.

Although the panel regression models used in extant studies have clear advantages over ARIMA models, they still embed the linearity assumption. With this in mind, we consider an alternate forecasting approach that is essentially the polar opposite of linear regression models: k-nearest neighbors (i.e., k-NN). The basic logic of k-NN is straightforward. First, identify the subject firm-year's k-nearest neighbors, which are the k firm-years that (1) have already reported earnings and (2) have predictors that are nearest to those of the subject firm-year. Second, set the forecast of the subject firm-year's earnings as the median of the k-nearest neighbors' lead earnings.

Unlike regression models, k-NN algorithms accommodate nonlinear relations well. In fact, there is no need for the researcher to assume a functional form. Rather, k-NN algorithms naturally adapt to the data. Moreover, as discussed in Hastie, Tibshirani, and Friedman (2015), if the training data consist of an arbitrarily large number of neighborhoods, each of which contains an arbitrarily large number of neighbors that have predictors close to those of a unique subject firm-year, k-NN generates the optimal forecast. That is, if the training data consist of many densely populated neighborhoods, k-NN works well.

There are two disadvantages to using k-NN. First, if the training data consist of a few overlapping, sparsely populated neighborhoods, k-NN is inaccurate. Second, k-NN does not accommodate multiple predictors as well as linear panel regression models. This is a manifestation of the curse of dimensionality (Bellman 1961). As discussed in Hastie et al. (2015), this phenomenon implies that, holding k constant, as the number of predictors (i.e., dimensions) increases, (1) the neighborhoods in the training data begin to overlap and become sparsely populated and (2) the fraction of a subject firm-year's nearest neighbors that are located near the extremes of the *training data* increases.

The above discussion motivates four empirical questions. First, how is forecast accuracy affected by the choice of the two defining attributes of every k-NN algorithm: (1) the number of nearest neighbors (i.e., k) and (2) the predictor set? Regarding the choice of k, if the training dataset is arbitrarily large, increasing k causes the sample median to become a more precise estimate of the population median. However, when there are finite training data, increasing k causes the nearest neighbors of a particular subject firm-year to become less similar to this firm-year. Consequently, the choice of k is an empirical question.

Regarding the choice of predictors, we begin with a predictor set that consists of the subject firm-year's M-year earnings history. Hence, a subject firm-year's nearest neighbors are those with earnings trends that are the most similar to the recent trend in the subject firm's earnings. This raises two questions: how long of a trend (i.e., how many lags of earnings) should we include in our predictor set? How do k-NN algorithms that use simple "earnings only" predictor sets perform relative to k-NN algorithms that use different and/or additional variables? To answer these questions, we evaluate alternate predictor sets by varying dimensionality in two ways: (1) the choice of variables and (2) the number of lags of the variables. Increasing dimensionality has costs and benefits. On the one hand, adding predictors adds information. On the other hand, per the curse of dimensionality, adding predictors causes the nearest neighbors of each subject firm-year to become less similar to this firm-year.

After establishing that k-NN*—i.e., the k-NN algorithm that uses 120 nearest neighbors (i.e., $k = 120$) and a predictor set consisting of current and lagged earnings (i.e., $M = 2$)—is superior, we turn to our second research question: are the forecasts generated by k-NN* more accurate than those generated by (1) the random walk, (2) the approach used by

BCG, (3) the linear regression models used by HVZ, LM, FRY, and VY, and (4) analysts? We view the random walk as the baseline because it is simple, well known, and has been shown to perform well relative to other extant models. We consider BCG's approach because it is a comparable-firm-based approach. The comparison of k-NN* with linear panel regression models is important because, as discussed above, k-NN algorithms and linear regression models are polar opposites: k-NN algorithms accommodate nonlinearities, whereas regression models do not. However, linear panel regression models accommodate large predictor sets better than k-NN algorithms. We consider analysts' forecasts because they are the default choice for many researchers and practitioners.

We show that k-NN* is more accurate than the alternate approaches. However, no approach, including k-NN*, will generate the best forecast for every observation. This leads to our third research question: are the absolute and relative (to other approaches) accuracy of the forecasts generated by k-NN* predictable *ex ante*? We evaluate the variable $MAD_{i,t}$, which reflects the extent to which the nearest neighbors "disagree" about what the subject firm-year's future earnings will be, and thus, we predict that k-NN* forecasts will perform better (worse) when $MAD_{i,t}$ is low (high).

Our final research question is: are k-NN* forecasts useful to decision makers? We focus on equity valuation, and we evaluate (1) the association between market-adjusted returns for year t+1 and contemporaneous unexpected earnings implied by the k-NN* forecasts (i.e., the EAC) and (2) the risk-adjusted return earned on a hedge portfolio that is long (short) firms for which the k-NN* forecasts imply positive (negative) earnings growth. We show that (1) the EAC (hedge portfolio return) that is based on the k-NN* forecasts is larger than the EAC (hedge portfolio return) that is based on the forecasts generated by any of the alternate approaches and (2) this EAC (hedge portfolio return) is larger when $MAD_{i,t}$ is low—i.e., when $MAD_{i,t}$ indicates high accuracy.

III. FORECASTING APPROACHES

For each subject firm-year *i,t*, we calculate (1) an *h*-year-ahead ($h \in \{1, 2, 3\}$) forecast of earnings before special items and (2) a forecast of aggregate earnings before special items for years t+1 through t+3. We refer to earnings before special items as *EBSI*.

k-NN

In general, k-NN is implemented in three steps. First, choose a set of predictors. Second, identify the *k*-nearest neighbors. Third, use the known outcomes of the *k*-nearest neighbors to forecast the unknown outcome for the subject firm-year. In the space below, we discuss how we implement this three-step process. We summarize the discussion in Figure 1.

We focus on a simple "earnings-only" predictor set that consists of the most recent *M*-year history of scaled earnings before special items, *SEBSI*. We use equity market value as our scaler because we do not want size to be the dominant determinant of our matches. Specifically, for subject (neighbor) firm-year *i,t* (*j,s*), we deflate all the dollar amounts in its most recent *M*-year history of *EBSI* by its equity market value at the end of year *t* (*s*).

Next, we choose the subject firm-year's nearest neighbors, which are the *k* neighbors with the smallest values of the variable $DIST^{M}_{i,t,j,s}$:

$$DIST^{M}_{i,t,j,s} = \sqrt{\Sigma^{M}_{m=1}\left(SEBSI_{i,t-m+1} - SEBSI_{j,s-m+1}\right)^2} \tag{1}$$

In Equation (1), $DIST^{M}_{i,t,j,s}$ is the Euclidean distance between subject firm *i*'s *M*-year earnings history ending in year *t* and neighbor firm *j*'s *M*-year earnings history ending in year *s*.[6]

Finally, we compute the forecast of *EBSI* by multiplying the median of the nearest neighbors' *SEBSI* in year $s + h$ by the subject firm's equity market value at the end of t. And we set the variable $MAD_{i,t}$ equal to the median absolute deviation of the nearest neighbors' realized *SEBSI* in year $s + 1$.

We refer to the best "earnings only" k-NN algorithm as k-NN*. We compare k-NN* with eight alternate k-NN algorithms. Each of the eight alternate algorithms is similar to the k-NN* algorithm, except it uses different sets of predictors to identify nearest neighbors. To implement each algorithm, we follow the same process described above, but we allow the distance measure to include additional sets of (normalized) predictor variables.

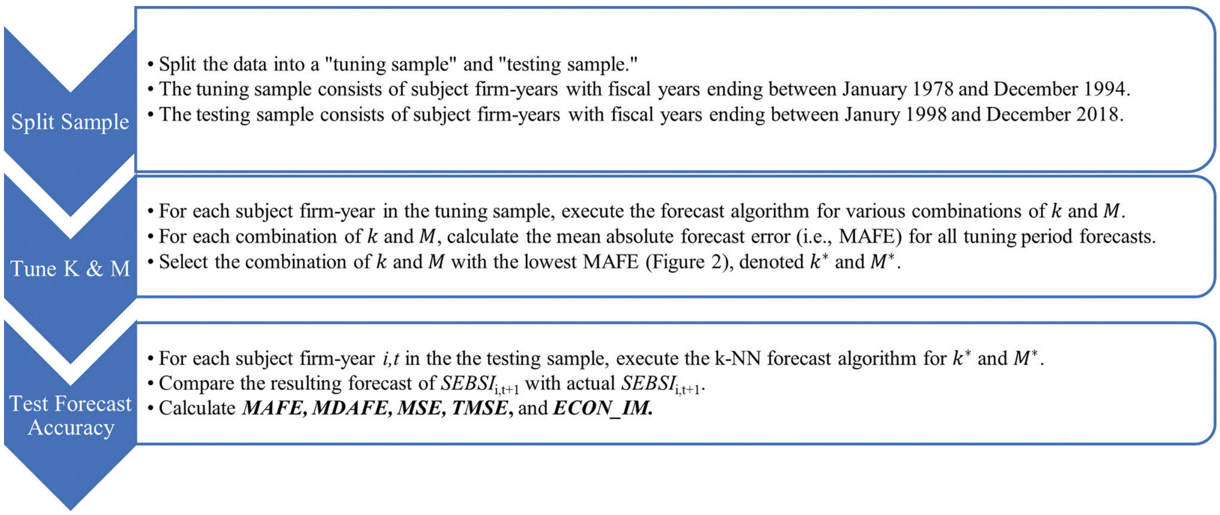
[6] When calculating $DIST^{M}_{i,t,j,s}$, we use normalized values of *SEBSI*. We calculate normalized values by subtracting the contemporaneous cross-sectional average of *SEBSI* from its raw value and then dividing this amount by the contemporaneous cross-sectional standard deviation of *SEBSI*.

FIGURE 1
k-NN Forecasting Process

Panel A: k-NN Forecast Algorithm

1. Determine the most recent M-year earnings history for subject (neighbor) firm-year i, t (j, s), and then deflate each element of this history by $MVE_{i,t}$ $(MVE_{i,s})$.
2. Calculate the Euclidean distance between the subject firm-year's scaled earnings history and the scaled earnings history of neighbor firm j, s, and then select the k neighbors with the smallest Euclidean distance. These are the subject firm-year's nearest neighbors.
3. Set $MAD_{i,t}$ equal to the median absolute deviation of the nearest neighbors' scaled lead earnings and use the median of the nearest neighbors' scaled lead earnings to impute the forecast of the subject firm-year's earnings.

Panel B: Tuning and Testing Process



Panel C: Timeline

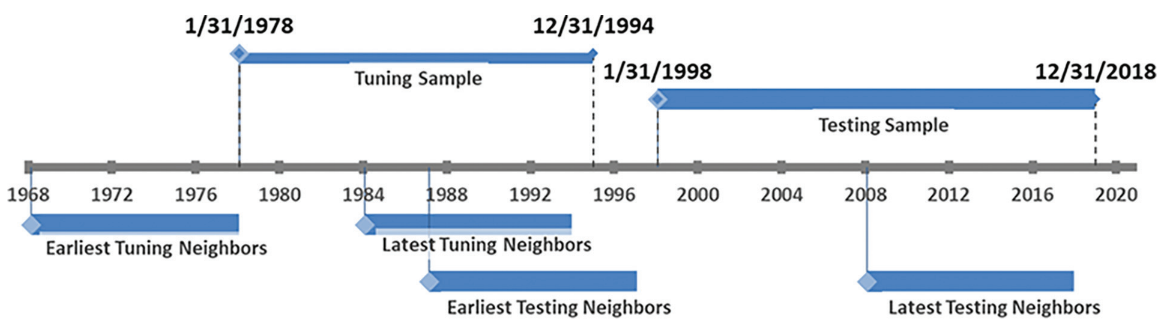


Figure 1 describes the k-NN forecasting process. Panel A describes the forecasting algorithm. Panel B provides an overview of the parameter tuning and testing procedures. Panel C provides a timeline showing the beginning and ending dates of the tuning and testing samples.
(The full-color version is available online.)

The eight alternate algorithms fall into two categories: (1) the DuPont (i.e., DUP) category and (2) the HVZ category. When implementing the k-NN$_{\text{DUP1}}$ algorithm, we identify nearest neighbors by using a root predictor set that consists of profit margin, *PM*, and asset turnover, *ATO*. The k-NN$_{\text{DUP2}}$ (k-NN$_{\text{DUP3}}$) algorithm adds sales growth, *SGrow*, (leverage, *LEV*) to the root predictor set; and the k-NN$_{\text{DUP4}}$ algorithm uses all four predictors. The k-NN$_{\text{HVZ1}}$ algorithm uses *SEBSI* and *ACC* to identify nearest neighbors; and the algorithms k-NN$_{\text{HVZ2}}$ through k-NN$_{\text{HVZ4}}$ are formed by progressively adding the predictors *AT*, *DIV*, and *LOSS*, which are defined below.

Finally, in order to implement each k-NN algorithm, we need to choose the values of k and M. To avoid look-ahead bias, we use the tuning sample, which contains subject firm-years drawn from 1978 to 1994, whereas our holdout (i.e., testing) sample contains firm-years drawn from 1998 to 2018 (see Figure 1, Panels A and B). Then, we evaluate the accuracy of each k-NN algorithm for various combinations of M and k. We elaborate on our tuning procedure in Section V.

BCG's Approach

BCG's approach is similar to a k-NN algorithm in the sense that they base their forecast of the subject firm-year's earnings on the earnings of a set of matched firm-years. To implement the BCG approach, we follow the four-step process described in BCG. First, we rank observations on their *SEBSI* for year $t-2$ and then we form two negative *SEBSI* groups and four positive *SEBSI* groups. Second, within each of these six groups, we rank observations into quintiles based on their average assets for year $t-2$. This yields 30 *SEBSI*-size bins: ten negative *SEBSI*-size bins and 20 positive *SEBSI*-size bins. Third, we randomly select a sample of 50 observations from the *SEBSI*-size bin to which subject firm-year i,t belongs, and we calculate the median earnings growth from $t-2$ to $t-1$. Finally, we determine our BCG forecast of the subject firm's year $t+1$ earnings by multiplying its earnings for year t by the median earnings growth rate.[7]

Regression Models

We evaluate forecasts based on the regression model proposed by HVZ and the EP regression model described in LM. The HVZ (EP) forecasts are obtained using the estimated coefficients from the ordinary least squares (i.e., OLS) regression shown in Equation (2) (Equation (3)) below.

$$SEBSI_{i,t+h} = \alpha_0 + \alpha_1 \times TA_{i,t} + \alpha_2 \times DD_{i,t} + \alpha_3 \times DIV_{i,t} + \alpha_4 \times SEBSI_{i,t} + \alpha_5 \times LOSS_{i,t} + \alpha_6 \times ACC_{i,t} + \varepsilon_{i,t+h} \tag{2}$$

$$SEBSI_{i,t+h} = \beta_0 + \beta_1 \times SEBSI_{i,t} + \beta_2 \times LOSS_{i,t} + \beta_3 \times (SEBSI_{i,t} \times LOSS_{i,t}) + \in_{i,t+h} \tag{3}$$

In the above equations, $SEBSI_{i,t+h}$ denotes firm i's scaled earnings before special items for year $t+h$, $TA_{i,t}$ denotes firm i's scaled total assets at the end of year t, $DD_{i,t}$ is an indicator variable that equals 1 (0) if firm i paid (did not pay) a dividend in year t, $DIV_{i,t}$ denotes firm i's scaled dividends for year t, $SEBSI_{i,t}$ denotes firm i's scaled earnings before special items for year t, $LOSS_{i,t}$ is an indicator variable that equals 1 (0) if $SEBSI_{i,t}$ is (is not) negative, $ACC_{i,t}$ denotes firm i's scaled accruals for year t, and $\varepsilon_{i,t+h}$ ($\in_{i,t+h}$) is the error term. With the exception of the indicator variables $DD_{i,t}$ and $LOSS_{i,t}$, all of the variables are deflated by firm i's equity market value at the end of year t.

Following FRY and VY, we estimate three versions of the EP model. In the first version, we use all firm-years in the estimation sample. In the other two versions, we use the subset of the estimation sample that consists of those firm-years that either have the same GICS code as the subject firm-year or are in the same lifecycle group (as defined by Dickinson 2011) as the subject firm-year. We refer to these two versions of the EP model as the EP-GICS model and the EP-LIFE model, respectively.

Rolling-Window Forecasting Procedure

We apply a rolling-window forecasting procedure. This assures that all of our forecasts are out of sample and that our returns tests are not affected by lookahead bias. Specifically, for each month of each calendar year, we identify nearest neighbors and estimate regressions coefficients using the sample of training data that consists of firm-years with fiscal years ending within the ten-year period that started $\left((10+h)\times 12\right)$ months earlier (h is the length of the forecast horizon in years). For example, suppose that we want to forecast one-year-ahead earnings for firms with fiscal years that ended in January 2013. Then, the training data would consist of firm-years with fiscal years that ended between February 2002 and January 2012 inclusive.

[7] BCG rank observations on the basis of their return on assets, *ROA*, and they use the average growth rate instead of the median growth rate. We rank on *SEBSI* and use the median growth rate so that our implementation is comparable to the other approaches that we evaluate. We also evaluate forecasts based on *ROA* sorts and/or the average growth rate, and in a set of untabulated results, we find that these forecasts are less accurate than the k-NN* forecasts.

IV. SAMPLE CONSTRUCTION AND DESCRIPTIVE STATISTICS

Sample Construction

We obtain our data from the Compustat Fundamentals Annual file. To avoid lookahead bias, we use the tuning sample, which contains subject firm-years drawn from 1978 to 1994, to tune the parameters of the k-NN algorithms. Because we need data from the cash flow statement to implement the lifecycle model, the main testing sample that we use to compare the k-NN* forecasts with those from alternate approaches spans the years 1998 through 2018.

In Table 1, we summarize how we construct the various forecast comparison samples contained in the testing sample. First, we identify the firm-years with (1) nonmissing *EBSI* for years t and t+1 and (2) nonmissing equity market value. We then remove financial firms and regulated firms, and to minimize the effect of database errors and small deflators, we also remove firm-years with equity market value that is less than ten million U.S. dollars or for which the absolute value of $SEBSI_{i,t}$ exceeds 1. After making these deletions, we arrive at the random walk forecast sample, which consists of 72,306 observations.

Second, we create a number of "comparison" samples that contain testing sample observations for which a forecast can be generated and evaluated for both (1) the k-NN* algorithm and (2) one of the alternate approaches. Finally, we construct the tuning sample that we use to determine the optimal values of k and M for each of the k-NN algorithms. We construct this sample in a manner that is similar to the way that we construct the samples described above. However, to avoid lookahead bias, we limit the tuning sample to subject firm-years between 1978 and 1994 inclusive. In addition, in order to be able to evaluate values of M between one and five years, we only include a firm-year if its *EBSI* for the current and previous four years are nonmissing.

In Appendix A, we provide a complete list of all of the variables that we use and describe how we calculate each variable.

TABLE 1

Sample Composition

Data Filter	Firm-Years
Total Compustat Observations 1998–2018	**179,662**
Less missing *EBSI*	−34,091
Less missing MVE	−17,452
Less MVE < \$10M	−18,678
Less absolute scaled earnings greater than 1	−2,158
Less missing future *EBSI*	−7,762
Less financial firms and regulated firms	−27,215
Random Walk Forecast Sample	**72,306**
Less missing lagged *EBSI*	−1,952
k-NN* Forecast Sample (Random Walk Comparison Sample)	**70,354**
k-NN* Forecast Sample	**70,354**
Less missing BCG data	−108
BCG Comparison Sample	**70,246**
k-NN* Forecast Sample	**70,354**
Less missing HVZ data	−1,489
HVZ Comparison Sample	**68,865**
k-NN* Forecast Sample	**70,354**
Less missing lifecycle data	−31
EP-LIFE Comparison Sample	**70,323**
k-NN* Forecast Sample	**70,354**
Less missing industry-level data	−1,549
EP-GIC Comparison Sample	**68,805**

Table 1 shows the effect of our data requirements on the final sample composition.

TABLE 2

Descriptive Statistics

Panel A: Summary Statistics for the Regression Estimation Sample

Variables	n	Mean	Std. Dev.	P05	P25	Med	P75	P95
ACC	118,709	−0.05	14.08	−0.41	−0.09	−0.02	0.00	0.13
DD	118,709	0.26	0.44	0.00	0.00	0.00	1.00	1.00
DIV	118,709	0.01	0.08	0.00	0.00	0.00	0.00	0.04
SEBSI	118,709	−0.03	0.19	−0.41	−0.06	0.03	0.07	0.14
LOSS	118,709	0.37	0.48	0.00	0.00	0.00	1.00	1.00
TA	118,709	1.53	31.31	0.11	0.44	0.89	1.68	4.45
PM	107,722	−3.98	162.78	−1.86	−0.03	0.03	0.08	0.19
ATO	107,722	1.18	0.93	0.14	0.59	1.02	1.54	2.76
LEV	107,722	5.21	289.48	1.13	1.41	1.91	2.81	7.20
SGrow	107,722	1.29	71.39	−0.31	−0.01	0.09	0.27	1.21

Panel B: Coefficients of the In-Sample Estimation of the HVZ Regression Model

Average n	Intercept	*ACC*	*DD*	*DIV*	*SEBSI*	*LOSS*	*TA*	Adjusted R^2
38,190	0.01	−0.07	0.02	−0.06	0.56	−0.05	0.00	0.48
	[17.37]	[−25.90]	[146.74]	[−9.17]	[93.92]	[−65.14]	[−7.87]	

Panel C: Coefficients of the In-Sample Estimation of the EP Regression Model

Lifecycle	Average n	Intercept	*SEBSI*	*LOSS*	*LOSS* ∗ *SEBSI*	Adjusted R^2
Unrestricted	41,680	0.02	0.50	−0.07	0.03	0.45
		[17.00]	[28.13]	[−58.44]	[1.05]	
Introduction	8,663	−0.01	0.55	−0.05	0.04	0.42
		[−6.57]	[18.31]	[−29.75]	[1.04]	
Growth	11,230	0.02	0.44	−0.04	0.03	0.24
		[34.59]	[30.25]	[−60.82]	[1.66]	
Mature	13,584	0.03	0.54	−0.04	−0.21	0.21
		[41.04]	[42.79]	[−43.62]	[−11.57]	
Shake-Out	4,649	0.01	0.52	−0.06	−0.03	0.36
		[10.81]	[29.19]	[−35.57]	[−1.01]	
Decline	3,337	−0.01	0.32	−0.04	0.20	0.35
		[−8.42]	[8.33]	[−35.44]	[4.29]	

Table 2, Panel A provides pooled summary statistics for the variables included in the regression and k-NN algorithms. Panel B (Panel C) shows the average coefficients of the ten-year rolling-window regressions for the HVZ (EP) model. t-statistics are derived from Fama-MacBeth standard errors.
See Appendix A for remaining variable definitions.

Descriptive Statistics

In Table 2, Panel A, we provide descriptive statistics for the predictors in (1) the HVZ model and (2) the DuPont category of k-NN algorithms.[8] Three comments are warranted. First, when constructing the training data for the approaches that rely on the HVZ variables, we do not remove observations with missing values of the DuPont variables and *vice versa*. Hence, the number of observations with available HVZ variables differs from the number of observations with available DuPont variables. Second, the descriptive statistics for the HVZ variables are similar to the descriptive statistics shown in other studies. Finally, the medians of the DuPont variables are similar to the medians shown in other

[8] The statistics relate to a pooled sample that we create by identifying every firm-year that is a member of any of the ten-year rolling samples of training data.

American Accounting Association

studies. However, because we neither delete observations with extreme values of the Dupont variables nor Winsorize extreme values of the DuPont variables, the means, standard deviations, and tails of the distributions are different.[9]

In Table 2, Panels B and C, we summarize the estimates of the regression coefficients for the HVZ model and the EP model, respectively. We show the average number of observations used in each rolling window of training data. Before estimating each regression, we "clean" the training data by removing observations for which (1) either $|SEBSI_{i,t}| > 1$ or $|SEBSI_{i,t+1}| > 1$ or (2) any of the regression variables are greater (less) than the first (99th) percentile. The coefficients (R^2s) are the time-series averages from the rolling-window regressions. Each t-statistic equals the average coefficient divided by its time-series standard error. Regarding the HVZ model, all of the coefficients are statistically different from 0 and the R^2 is 0.48.[10] Similarly, when we estimate the EP model on the unrestricted sample, we find that all of the coefficients are statistically different from 0 and the R^2 is 0.45. However, when we estimate a separate regression for each life-cycle group, we find that there is considerable variation in the magnitudes of the coefficients and the R^2s.

V. BASIC PROPERTIES OF THE k-NN* ALGORITHM

Parameter Tuning

To avoid lookahead bias, we use the tuning sample to choose the optimal values of M and k. We consider values of M between one and five years and values of k between 10 and 500 neighbors. We vary M by increments of one and k by increments of ten. We define the optimal combination of M and k (i.e., M^* and k^*) as the combination that minimizes the *MAFE*. We use the *MAFE* rather than the *MSE* because the *MAFE* is less sensitive to extreme forecast errors.[11] We express forecast errors as percentages of the subject firm-year's equity market value at the end of year t.

We show our parameter tuning results in Figure 2. For each of the five different values of M, we plot the *MAFE*s corresponding to the different values of k. Then, we identify the optimal value of k, which we refer to as k^*. The graph reveals four facts. First and foremost, the k-NN algorithm in which $M = 2$ and $k = 120$ is best. Second, regardless of the value of M, there are diminishing marginal returns to increasing k. Third, there is a negative monotonic relation between M and k^*. This is a manifestation of the curse of dimensionality.[12] That is, as the number of dimensions used to find matches increases, there are fewer good matches. Finally, matching on longer histories—i.e., higher values of M—is inferior with the important exception that using a history of two years—i.e., current and lagged *SEBSI*—is better than using only current *SEBSI*.

Benefits of Matching and Using the Implied Earnings Growth Rate

In this subsection, we evaluate the benefits of matching on *SEBSI* and basing the forecast on the growth rate implied by the nearest neighbors' lead earnings. To evaluate the importance of matching on *SEBSI*, we compare the k-NN* algorithm with two naive algorithms that are identical to the k-NN* algorithm except that their forecasts are based on the earnings of different sets of neighbors. Specifically, the random (economy-wide) k-NN algorithm matches the subject firm-year to k randomly selected firm-years (all firm-years) drawn from the ten-year period that ends in year $t-1$. To evaluate the importance of using the implied growth rate, we compare the k-NN* algorithm with a k-NN algorithm that is identical except that it bases the forecast on the median of the k-nearest neighbors' matched-year (i.e., year s) earnings instead of their lead (i.e., year $s+1$) earnings. We refer to this as the no-growth algorithm. Finally, we also compare the k-NN* algorithm with the random walk, which is a naive k-NN algorithm that matches the subject firm-year to itself and assumes zero growth. Consequently, this comparison is informative about the benefits of both matching and using the implied growth rate.

We document the results of the comparisons described above in Figure 3, which has the same format as Figure 2. The k-NN* algorithm is noticeably better than all of the naive algorithms.[13] This has two implications. First, the way matches are formed matters. Although the *MAFE* of the random k-NN algorithm decreases as k increases, it converges

[9] Our results are not attributable to extreme values. In untabulated tests, we conduct a battery of robustness tests in which we define extreme values differently (e.g., removing observations with $|PM_{i,t}| > 1$ or $|SGrow_{i,t}| > 1$, total assets or sales less than ten million U.S. dollars, etc.), and we find that, regardless of the way that we define extreme values, the k-NN* forecasts are the most accurate.

[10] All of our inferences about relative forecast accuracy remain unchanged if we use forecasts generated by HVZ regressions that are based on undeflated variables.

[11] We acknowledge that the *MAFE* loss function that we use to tune the k-NN algorithms is different from the *MSE* loss function that we use to train the regression models. We use OLS regressions, which embed an *MSE* loss function, because they are the most popular way of training regression-based models in the extant literature. In untabulated tests, we retrain all of the regression-based models using median regressions, which embed a *MAFE* loss function, and we find that the k-NN* forecasts are superior to the forecasts generated by the median regressions.

[12] Dimensionality is a function of both the number of variables that are used to find matches and the length of the history in the variables (i.e., M).

[13] We obtain similar results (untabulated) when we evaluate either the median absolute forecast error or the mean square error.

FIGURE 2
Mean Absolute Forecast Error by *k* and *M*

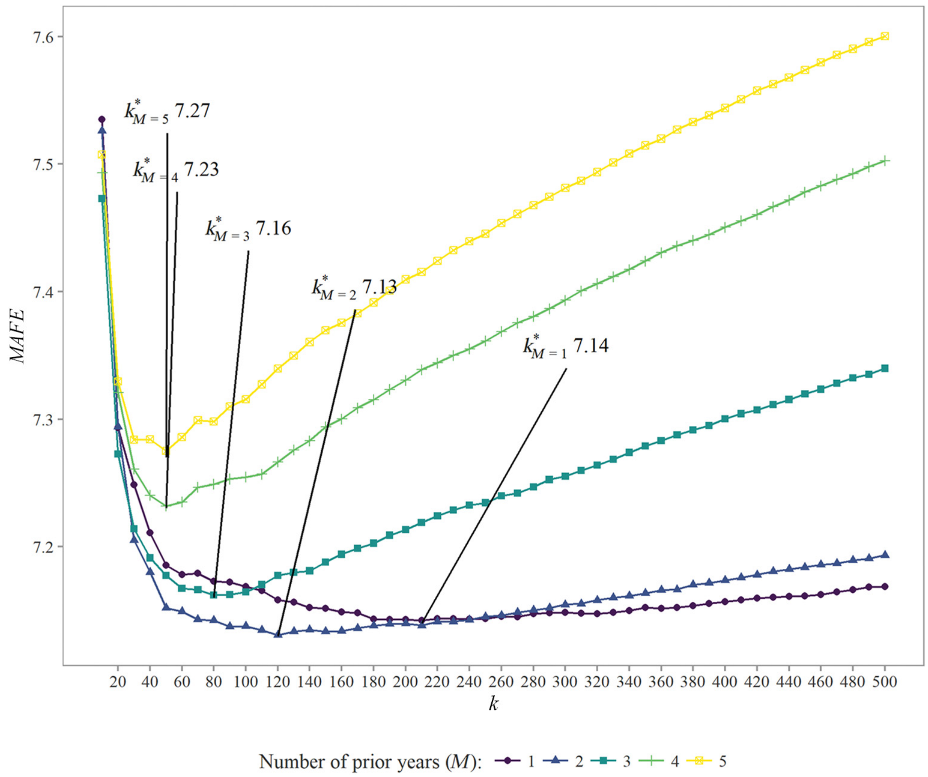


Figure 2 shows the *MAFE* for each combination of *k* and *M* from a common sample with data for all combinations. Each line plots the *MAFE* by *k* for a specific value of *M*. The labels for each line point to the number of neighbors (k*) that generate the minimum *MAFE* for a given value of *M*. (The full-color version is available online.)

to the *MAFE* of the economy-wide k-NN algorithm, which is greater than 11.5 and much larger than the *MAFE* of the k-NN* algorithm. Second, growth also matters. Regardless of the value of k, forecasts based on the median of the nearest neighbors' matched-year earnings are less accurate than forecasts generated by the k-NN* algorithm, which are based on the median of the nearest neighbors' lead earnings.

Comparison of Subject Firm-Years to Their Nearest Neighbors

In Table 3, we compare the subject firm-years in the testing sample with the nearest neighbors chosen by the k-NN* algorithm. The statistics shown in Panel A are calculated via a two-step process. First, for each subject firm-year i,t and nearest neighbor j,s, we calculate the variable shown in the first column of the panel. Second, we calculate and tabulate the mean, standard deviation, etc. of each variable. To calculate the statistics shown in Panel B, for each subject firm-year and nearest neighbor pair, we create an indicator variable that equals 100 if the pair share a common attribute and 0 if they do not. Then, we calculate the mean of this indicator variable, which yields the percentage of nearest neighbors that have this attribute in common with the subject firm-year to which they are matched.

As shown in Panel A, the average, median, and interquartile range of differences in *SEBSI* are each equal to 0.00. Thus, subject firm-years have very similar scaled earnings as their nearest neighbors, which is not surprising given that *SEBSI* is the variable that we use to find matches. That said, there are some clear dis-similarities between the subject firm-years and their nearest neighbors. For example, the nearest neighbor firm-years typically precede the subject firm-year by more than five years.

As shown in Panel B, the percentage of nearest neighbors with the same Fama-French industry (two-digit SIC) code as their subject firm-year is 16.97 (7.74), and 4.38 (36.51) share the same GICS code (are in the same lifecycle group) as

FIGURE 3
Decomposition of *MAFE* by Components of the Matching Algorithm

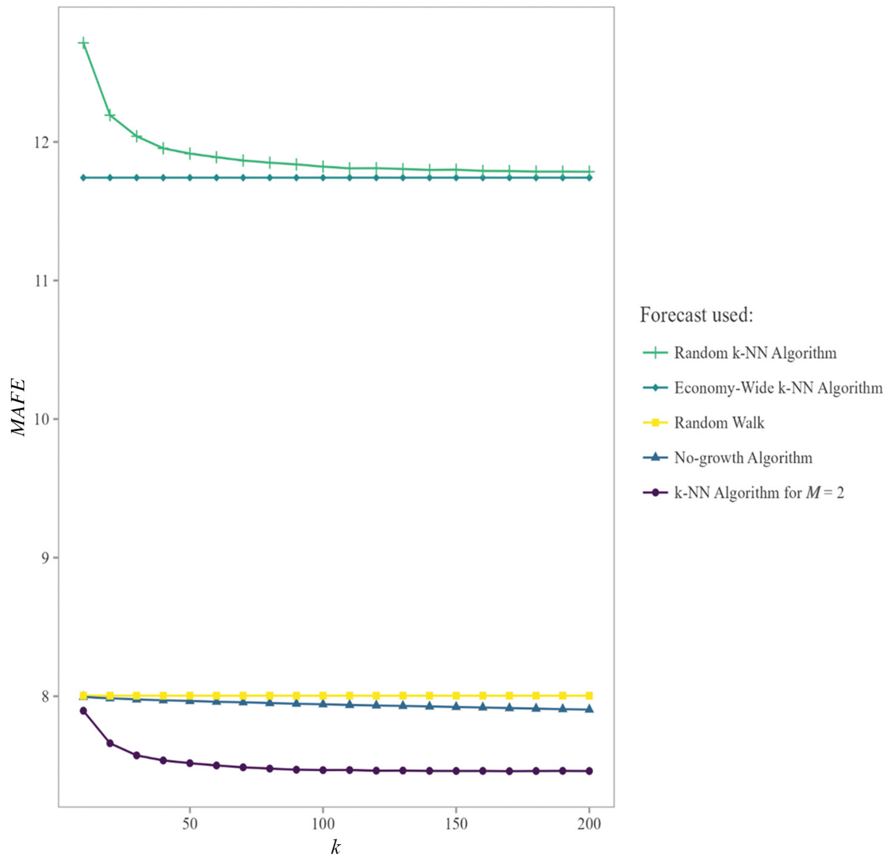


Figure 3 compares the mean absolute forecast error (*MAFE*) of the k-NN* algorithm with the *MAFE* of various naive k-NN algorithms for firm-year observations in the testing sample. We vary the number of selected firms, *k*. In the random (economy-wide) k-NN algorithm, we select a random set of *k* (all possible) neighbors. In the no-growth algorithm, we set the forecast equal to the median of the nearest neighbors (per the k-NN* algorithm) matched-year (i.e., year *s*) earnings.
(The full-color version is available online.)

their subject firm-year. Moreover, only 13.64 percent of the nearest neighbors are in the same size decile as the subject firm-year to which they are matched. Finally, subject firms are rarely matched to a younger version of themselves—i.e., less than 1 percent of the nearest neighbors have the same GVKEY as their subject firm-year.

These results lead to two conclusions. First, the nearest neighbors are very similar in terms of scaled earnings, but otherwise quite heterogeneous. Second, our matching algorithm is different from conventional algorithms that match the subject firm-year to its contemporaries in the same industry, same size stratum, or same lifecycle group.

VI. COMPARISON TO ALTERNATE FORECASTS

In this section, we compare the forecasts generated by the k-NN* algorithm with a number of alternate approaches. We use the forecast comparison sample, and we evaluate five measures of forecast accuracy. The first four measures are (1) the mean of the absolute scaled forecast errors, *MAFE*; (2) the median of the absolute scaled forecast errors, *MDAFE*; (3) the mean of the squared scaled forecast errors, *MSE*; and (4) the mean of the squared trimmed scaled forecast errors, *TMSE*.[14] Higher values of these four measures imply lower accuracy.

[14] Each scaled forecast error equals 100 multiplied by the ratio of the difference between realized and forecasted $EBSI_{i,t+1}$ to $MVE_{i,t}$. To calculate *TMSE*, we first delete the top and bottom 0.1 percent of the scaled forecast errors and then we calculate the mean square error.

TABLE 3

Comparison of Subject Firms with Their Nearest Neighbors

Panel A: Descriptive Statistics for Comparable and Subject Firms (n = 9,688,080)

Variables	Mean	Std. Dev.	P05	P25	Med	P75	P95
$SEBSI_{i,t} - SEBSI_{j,s}$	0.00	0.03	−0.02	0.00	0.00	0.00	0.02
$ACC_{i,t} - ACC_{j,s}$	0.00	0.27	−0.37	−0.07	0.00	0.06	0.38
$FEG_{i,t} - FEG_{j,s}$	0.00	0.06	−0.05	0.00	0.00	0.00	0.05
$Year_{i,t} - Year_{j,s}$	5.57	2.89	1.00	3.00	6.00	8.00	10.00
$Age_{i,t} - Age_{j,s}$	2.05	17.17	−28.00	−7.00	1.00	11.00	35.00

Panel B: Industry and Lifecycle Membership (n = 9,688,080)

Variable: *Percent Same*	FF12	Sic2	GIC	GVKEY	Lifecycle	MVE
Mean	16.97	7.74	4.38	0.00	36.51	13.64

Table 3, Panel A presents descriptive statistics that we use to compare the nearest neighbors chosen by the k-NN* algorithm with the subject firm-year. The suffix *i,t* denotes the subject firm-year, and the suffix *j,s* denotes the relevant nearest neighbor firm-year. Panel B reports the percentage of matched nearest neighbors (*Percent Same*) that are in the same industry (lifecycle) as the subject firm-year or in the same size decile (MVE). FF12 is the Fama-French 12 industries classification, and Sic2 is the two-digit SIC industry code. *t* is the first year of the two-year earnings sequence. See Appendix A for remaining variable definitions.

Our fifth measure of forecast accuracy, $IMPROVE$, equals the percentage of observations for which a particular approach generates a "meaningful" improvement in forecast accuracy *vis-à-vis* the alternate approach. $IMPROVE$ is a nonparametric measure of the percentage of observations for which a given forecasting approach is considerably more useful than the alternate approach. It is inspired by and similar to a measure used by Fairfield, Sweeney, and Yohn (1996). Fairfield et al. (1996) define an "economically significant" improvement in forecast accuracy as a forecast that is 0.5 percent (of equity *book* value) more accurate than the alternate forecast. Because we deflate by equity market value, we define a meaningful forecast improvement as a forecast that is 0.5 percent (of equity *market* value) more accurate than the alternate forecast. We note that, unlike $MAFE$, $MDAFE$, MSE, and $TMSE$, higher values of $IMPROVE$ imply higher accuracy.

Comparison to Alternate k-NN Algorithms

We compare the k-NN* algorithm with the eight alternate k-NN algorithms described in Section III. For each algorithm, we use the tuning sample and the approach described in Section V to choose the optimal values of M and k. We present the results of these analyses in Table 4. First and foremost, the k-NN* algorithm is never less accurate and usually more accurate than all of the alternate k-NN algorithms. Second, the k-NN* algorithm dominates the algorithms in the DuPont category. For example, regardless of the error metric used, k-NN_{DUP1} is the best DuPont algorithm. Nonetheless, its $MAFE$ ($MDAFE$) is 21.8 (42.0) percent larger than the $MAFE$ ($MDAFE$) of the k-NN* algorithm and its MSE ($TMSE$) is 19.6 (33.2) percent larger than the MSE ($TMSE$) of the k-NN* algorithm. Moreover, when compared with k-NN_{DUP1}, the k-NN* algorithm generates meaningful improvements in accuracy for 54.68 percent of the observations. Finally, the algorithms in the HVZ category perform relatively well. For example, the k-NN_{HVZ1} and k-NN_{HVZ2} algorithms are roughly as accurate as the k-NN* algorithm. That said, the k-NN_{HVZ3} and k-NN_{HVZ4} algorithms are less accurate than the k-NN* algorithm.

Comparison with Extant Approaches

We compare the forecasts generated by the k-NN* algorithm with forecasts generated by (1) the random walk (i.e., RW), (2) the BCG approach (i.e., BCG), (3) the HVZ model (i.e., HVZ), (4) the EP model, (5) the EP-GICS model, and (6) the EP-LIFE model. We present the results of these analyses in Table 5, which has the same format as Table 4. These results lead to one overarching conclusion: the k-NN* algorithm is superior to all of the extant approaches. Specifically, regardless of the forecast-accuracy measure used, the k-NN* algorithm is always the best and all the differences are statistically significant.

TABLE 4

Comparison of the k-NN* Algorithm with Alternate k-NN Algorithms (t+1)

Algorithm	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
k-NN*	67,672	7.40	2.69	4.04	2.14	54.68
k-NN$_{DUP1}$	67,672	9.01	3.82	4.83	2.85	29.00
k-NN$_{DUP1}$ – k-NN*		1.61***	1.13***	0.79***	0.71***	−25.68***
k-NN*	63,394	6.85	2.55	3.47	1.78	57.52
k-NN$_{DUP2}$	63,394	8.99	3.89	4.50	2.72	27.91
k-NN$_{DUP2}$ – k-NN*		2.15***	1.34***	1.03***	0.94***	−29.61***
k-NN*	66,853	7.36	2.67	4.02	2.13	55.97
k-NN$_{DUP3}$	66,853	9.19	3.87	4.93	2.96	28.54
k-NN$_{DUP3}$ – k-NN*		1.83***	1.20***	0.91***	0.83***	−27.43***
k-NN*	62,763	6.82	2.53	3.48	1.78	57.36
k-NN$_{DUP4}$	62,763	8.99	3.87	4.52	2.72	27.57
k-NN$_{DUP4}$ – k-NN*		2.17***	1.34***	1.04***	0.94***	−29.79***
k-NN*	68,865	7.42	2.72	3.98	2.11	25.88
k-NN$_{HVZ1}$	68,865	7.46	2.73	4.00	2.14	25.23
k-NN$_{HVZ1}$ – k-NN*		0.04**	0.01	0.02	0.03**	−0.65**
k-NN*	68,865	7.42	2.72	3.98	2.11	28.23
k-NN$_{HVZ2}$	68,865	7.45	2.73	3.98	2.12	27.54
k-NN$_{HVZ2}$ – k-NN*		0.03	0.01	0.00	0.01	−0.69
k-NN*	68,865	7.42	2.72	3.98	2.11	34.78
k-NN$_{HVZ3}$	68,865	7.58	2.88	4.04	2.17	29.56
k-NN$_{HVZ3}$ – k-NN*		0.16***	0.16***	0.06*	0.06***	−5.22***
k-NN*	68,865	7.42	2.72	3.98	2.11	34.75
k-NN$_{HVZ4}$	68,865	7.55	2.87	3.98	2.15	29.71
k-NN$_{HVZ4}$ – k-NN*		0.12***	0.16***	0.00	0.03**	−5.04***

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 4 tabulates the forecast error metrics that we use to compare the k-NN* algorithm with alternate k-NN algorithms. The available forecasts of the k-NN* algorithm are matched to the available forecasts of the alternate algorithm. Below each forecast error metric, the table provides the difference between the error metric of the k-NN* algorithm and the error metric for the alternate algorithm. Statistical significance of the differences in mean error metrics is determined based on t-statistics clustered by firm and calendar year. The statistical significance of differences in *MDAFE* is determined using quantile regression tests for differences in the median of the absolute forecast error distribution between approaches. See Appendix A for definitions of the forecast evaluation metrics tabulated in each column.

Analyst Following and Forecast Horizon

Given the necessity of having an accurate statistical forecasting approach for firms that are not covered by analysts, in this subsection, we separately evaluate firm-years that are covered by analysts and those that are not. We also evaluate forecasts of two- and three-year-ahead earnings as well as forecasts of aggregate (i.e., cumulative) earnings for years t+1 through t+3. For the sake of brevity, we compare the k-NN* algorithm with the random walk, the HVZ model, and the EP-LIFE model.

We present the results of these analyses in Table 6, which has the same format as Tables 4 and 5. In Panel A, we show how the absolute and relative accuracy of k-NN* forecasts vary with analyst coverage. Two conclusions warrant discussion. First, each of the approaches we evaluate is more accurate when it is used to forecast the earnings of firms that are covered by analysts. Second, the relative performance of the k-NN* forecasts do not depend on analyst coverage.

In Panel B, we show the results relating to forecasts of earnings for years t+2 and t+3 as well as aggregate earnings for years t+1 through t+3. Each forecasting approach performs worse as the forecast horizon lengthens, which is not surprising. Nonetheless, and perhaps more important, regardless of the forecast horizon, the k-NN* forecasts are more accurate than the forecasts that are generated by the other approaches.

Comparison with Analysts' Forecasts

Given the popularity of analysts' forecasts, we compare them with the k-NN* forecasts. To generate k-NN* forecasts that are comparable to analysts' *EPS* forecasts, we match nearest neighbors to the subject firm-year on the basis of

TABLE 5

Comparison of the k-NN* Algorithm with Extant Approaches (t+1)

Approach	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
k-NN*	70,354	7.46	2.73	3.99	2.14	36.94
RW	70,354	8.00	2.84	4.27	2.43	29.34
RW – k-NN*		0.54***	0.11***	0.28***	0.29***	−7.59***
k-NN*	70,246	7.46	2.73	3.98	2.14	37.36
BCG	70,246	8.11	3.06	4.28	2.44	26.79
BCG – k-NN*		0.65***	0.33***	0.30***	0.30***	−10.57***
k-NN*	68,865	7.42	2.72	3.98	2.11	45.07
HVZ	68,865	7.80	3.23	4.08	2.21	31.37
HVZ – k-NN*		0.38***	0.52***	0.10***	0.09***	−13.69***
k-NN*	70,354	7.46	2.73	3.99	2.14	48.56
EP-ALL	70,354	7.89	3.27	4.07	2.23	32.35
EP-ALL – k-NN*		0.42***	0.54***	0.08***	0.08***	−16.2***
k-NN*	68,805	7.49	2.75	4.04	2.16	46.54
EP-GIC	68,805	7.87	3.28	4.13	2.24	32.57
EP-GIC – k-NN*		0.38***	0.53***	0.09***	0.08***	−13.97***
k-NN*	70,323	7.46	2.73	3.99	2.14	42.26
EP-LIFE	70,323	7.72	3.19	4.05	2.19	32.14
EP-LIFE – k-NN*		0.26***	0.46***	0.05*	0.05***	−10.12***

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 5 tabulates the forecast error metrics that we use to compare the k-NN* algorithm with extant earnings forecasting approaches. The available forecasts of the k-NN* algorithm are matched to the available forecasts of the extant approach. Below each forecast error metric, the table provides the difference between the error metric for the k-NN* algorithm and the error metric for the alternate approach. Statistical significance of the differences in mean error metrics is determined based on t-statistics clustered by firm and calendar year. The statistical significance of differences in *MDAFE* is determined using quantile regression tests for differences in the median of the absolute forecast error distribution between approaches.
See Appendix A for definitions of the forecast evaluation metrics tabulated in each column.

TABLE 6

Performance after Partitioning by Analyst Coverage and for Longer Forecast Horizons

Panel A: Split by Analyst Coverage

Approach	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
t+1 Forecast Error with Analyst Coverage						
k-NN*	47,315	5.77	2.12	2.54	1.34	36.05
RW	47,315	6.24	2.22	2.80	1.55	27.42
RW – k-NN*		0.46***	0.10***	0.26***	0.22***	−8.63***
k-NN*	46,305	5.72	2.10	2.49	1.29	44.46
HVZ	46,305	6.11	2.61	2.58	1.36	28.88
HVZ – k-NN*		0.40***	0.51***	0.09***	0.07***	−15.59***
k-NN*	47,307	5.77	2.11	2.54	1.34	41.12
EP-LIFE	47,307	6.04	2.51	2.60	1.38	29.74
EP-LIFE – k-NN*		0.27***	0.39***	0.07***	0.04***	−11.38***
t+1 Forecast Error without Analyst Coverage						
k-NN*	23,039	10.94	4.70	6.98	3.97	38.75
RW	23,039	11.63	4.83	7.29	4.35	33.30
RW – k-NN*		0.70***	0.14*	0.31**	0.38***	−5.45***

(continued on next page)

American Accounting Association

TABLE 6 (continued)

Approach	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
k-NN*	22,560	10.92	4.68	7.02	4.00	46.30
HVZ	22,560	11.27	4.95	7.15	4.14	36.50
HVZ – k-NN*		0.34***	0.27***	0.13*	0.14***	−9.80***
k-NN*	23,016	10.94	4.70	6.99	3.97	44.61
EP-LIFE	23,016	11.18	5.10	7.01	4.03	37.07
EP-LIFE – k-NN*		0.24***	0.41***	0.02	0.06	−7.54***

Panel B: Forecast Errors for Different Forecast Horizons

Approach	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
***t*+2 Forecast Error**						
k-NN*	62,293	9.39	3.86	4.92	3.07	42.95
RW	62,293	10.25	4.06	5.39	3.60	34.96
RW – k-NN*		0.86***	0.20***	0.47***	0.52***	−7.99***
k-NN*	60,945	9.34	3.84	4.91	3.06	44.06
HVZ	60,945	9.65	4.19	5.05	3.20	35.09
HVZ – k-NN*		0.31***	0.35***	0.13***	0.15***	−8.98***
k-NN*	62,272	9.39	3.86	4.92	3.07	40.89
EP-LIFE	62,272	9.55	4.15	4.94	3.11	34.94
EP-LIFE – k-NN*		0.16***	0.29***	0.02	0.04***	−5.96***
***t*+3 Forecast Error**						
k-NN*	55,133	10.69	4.54	7.40	3.96	43.99
RW	55,133	11.72	4.82	7.97	4.57	36.65
RW – k-NN*		1.04***	0.28***	0.56***	0.61***	−7.34***
k-NN*	53,915	10.62	4.51	7.42	3.94	43.61
HVZ	53,915	10.81	4.73	7.46	4.05	37.32
HVZ – k-NN*		0.20***	0.22***	0.04	0.11***	−6.28***
k-NN*	55,118	10.68	4.54	7.40	3.96	41.02
EP-LIFE	55,118	10.82	4.75	7.38	3.97	35.36
EP-LIFE – k-NN*		0.14***	0.21***	−0.02	0.01	−5.66***
Aggregate Forecast Error (*t*+1 + *t*+2 + *t*+3)						
k-NN*	55,058	22.95	9.94	26.48	17.17	50.23
RW	55,058	25.46	10.58	30.43	21.26	41.49
RW – k-NN*		2.50***	0.64***	3.94***	4.09***	−8.75***
k-NN*	53,843	22.83	9.88	26.48	17.11	50.81
HVZ	53,843	23.70	10.96	27.53	18.14	41.27
HVZ – k-NN*		0.88***	1.08***	1.05**	1.03***	−9.54***
k-NN*	55,043	22.95	9.94	26.48	17.17	48.81
EP-LIFE	55,043	23.46	10.84	26.69	17.53	41.44
EP-LIFE – k-NN*		0.51***	0.89***	0.21	0.36***	−7.37***

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 6 tabulates the forecast error metrics that we use to compare the k-NN* forecasts within subsamples. Below each forecast error metric, the table provides the difference between the error metric for the k-NN* algorithm and the error metric for the alternate approach. Panel A examines forecast errors for firm-years with and firm-years without analyst coverage. Panel B tabulates forecast error metrics from the k-NN* algorithm for various forecast horizons. Statistical significance of the differences in mean error metrics is determined based on t-statistics clustered by firm and calendar year. The statistical significance of differences in *MDAFE* is determined using quantile regression tests for differences in the median of the absolute forecast error distribution between approaches.
See Appendix A for definitions of the forecast evaluation metrics tabulated in each column.

TABLE 7

Comparison of the k-NN* Algorithm with Analysts

Approach	n	*MAFE*	*MDAFE*	*MSE*	*TMSE*	*IMPROVE*
***t*+1 Forecast Error**						
k-NN*	50,284	3.89	1.30	0.90	0.71	23.06
ANALYST	50,284	3.53	1.08	0.88	0.65	33.52
ANALYST – k-NN*		−0.37***	−0.23***	−0.03	−0.06*	10.46***
***t*+2 Forecast Error**						
k-NN*	40,486	4.46	1.97	0.87	0.69	35.97
ANALYST	40,486	4.65	2.02	0.98	0.72	33.47
ANALYST – k-NN*		0.19*	0.05**	0.11***	0.03	−2.49
***t*+3 Forecast Error**						
k-NN*	27,556	4.50	2.34	0.78	0.59	44.75
ANALYST	27,556	5.18	2.74	0.95	0.73	33.14
ANALYST – k-NN*		0.68***	0.40***	0.17***	0.14***	−11.61***
Aggregate Forecast Error (*t*+1 + *t*+2 + *t*+3)						
k-NN*	27,526	9.36	4.68	3.29	2.53	46.27
ANALYST	27,526	10.14	5.07	3.78	2.84	40.62
ANALYST – k-NN*		0.78**	0.40***	0.49**	0.32*	−5.65

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 7 tabulates the forecast error metrics that we use to compare the k-NN* algorithm with sell-side analysts (ANALYST). The k-NN* algorithm uses I/B/E/S actuals earnings data to obtain forecasts. Below each forecast error metric, the table provides the difference between the error metric for the k-NN* algorithm and the error metric for sell-side analysts. Statistical significance of the differences in mean error metrics is determined based on t-statistics clustered by firm and calendar year. The statistical significance of differences in *MDAFE* is determined using quantile regression tests for differences in the median of the absolute forecast error distribution between approaches.
See Appendix A for definitions of the forecast evaluation metrics tabulated in each column.

its history of I/B/E/S *EPS* actuals (i.e., I/B/E/S street earnings).[15] We compare k-NN* *EPS* forecasts with the mean consensus analyst *EPS* forecast for year $t+h$ per the first summary file that became available after the year t announcement date.[16] When evaluating forecasts of *EPS* for years t+1 and t+2, we use analysts' explicit forecasts. Following Bradshaw, Drake, J. Myers, and L. Myers (2012), we impute analysts' forecasts of *EPS* for year t+3 from their forecasts of *EPS* for year t+2 and their forecasts of long-term growth. The availability of I/B/E/S actual *EPS* and consensus analysts' forecasts leads to comparison samples of 50,284, 40,486, and 27,556 firm-years for horizons t+1, t+2, and t+3, respectively.

The results of comparing k-NN* forecasts with analysts' forecasts are shown in Table 7. Three results are noteworthy. First, when we consider a one-year horizon, analysts' forecasts are more accurate than the k-NN* forecasts. This is not surprising. Analysts have access to a large amount of quantitative and qualitative information that is not embedded in the k-NN* forecasts.

Second, when we consider longer forecast horizons, the k-NN* forecasts are more accurate than analysts' forecasts. Moreover, the differences in accuracy increase as the forecast horizon lengthens. Finally, k-NN* forecasts of aggregate *EPS* for years t+1 through t+3 are more accurate than analysts' forecasts. Therefore, although k-NN* forecasts of one-year-ahead *EPS* are inferior, when these forecasts are aggregated with the superior longer horizon k-NN* forecasts, the k-NN* forecast of aggregate *EPS* is more accurate than analysts' forecast of aggregate *EPS*.[17]

VII. CROSS-SECTIONAL VARIATION IN FORECAST ACCURACY

In this section, we evaluate the cross-sectional variation in the absolute and relative (to other approaches) accuracy of the k-NN* forecasts. To do this, we estimate regressions of either the absolute or relative k-NN* forecast error on a

[15] We continue to use $M = 2$ and $k = 120$. We deflate *EPS* by the end-of-year share price per CRSP, and we deflate the forecast errors by the subject firm's share price at the end of year t. We adjust share prices for stock splits and dividends using the CRSP price adjustment factor.
[16] We require the first available summary file date (STATPERS) to occur no later than 45 days after the earnings announcement date (ANNDATS).
[17] We also compare the k-NN* *EPS* forecasts with those generated by the random walk and the EP-ALL models estimated using I/B/E/S actuals. We find that (results untabulated), regardless of the forecast horizon and the measure of accuracy, the forecasts generated by the k-NN* algorithm are more accurate than those generated by either the random walk or EP-ALL model.

number of independent variables that are defined in Appendix A. In the first four regressions, we evaluate forecasts of year-ahead *EBSI*. In the first of these four regressions, the dependent variable is the absolute value of the scaled k-NN* forecast error for subject firm-year i,t, $|kNNFE_{i,t}|$. The dependent variables in regressions two through four are the difference between $|kNNFE_{i,t}|$ and the absolute value of the scaled forecast error for subject firm-year i,t generated by the random walk, the EP-LIFE model, and the HVZ model, respectively. In regression five, the dependent variable is the absolute value of the scaled k-NN* *EPS* forecast error, and the dependent variable in regression six is the difference between the absolute k-NN* *EPS* forecast error and that of the consensus analyst *EPS* forecast.

We estimate OLS panel regressions, and we report standard errors that are clustered by firm and year. For ease of interpretation, we standardize each continuous variable so that it has a mean of 0 and a standard deviation of 1.[18] Consequently, the absolute values of the coefficients can be used to rank the independent variables with respect to their relative importance.

In Table 8, we show the coefficient estimates from each of the regressions described above. Two results stand out. First, the self-assessment feature that is built into the k-NN* algorithm (i.e., $MAD_{i,t}$) is reliable. Specifically, as expected, when $MAD_{i,t}$ is low (high), the absolute and relative accuracy of the k-NN* forecasts are high (low). That is, the more disagreement there is among a subject firm-year's nearest neighbors, the lower the accuracy of their forecast. Moreover, the coefficient on $MAD_{i,t}$ is always one of the three largest coefficients (in terms of absolute value) and it is the largest (second largest) in two (three) of the regressions.

Second, relative to the other *statistical* forecasting approaches, the k-NN* algorithm is at its best when it makes extreme forecasts. Specifically, the coefficient (t-statistic) on $AbsFEG_{i,t}$ is −0.40 (−10.66) in the regression that compares the k-NN* forecast errors with the random walk forecast errors. Moreover, the coefficients (t-statistics) on $AbsFEG_{i,t}$ in the regressions that compare the k-NN* algorithm with the EP-LIFE model and the HVZ model are −0.12 and −0.15 (−4.90 and −4.71), respectively. These results demonstrate the advantages of using the growth rate implied by the nearest neighbors' earnings trends instead of assuming either no growth in earnings (i.e., the random walk) or an average rate of mean reversion (i.e., linear panel regression models).

VIII. USEFULNESS OF k-NN* FORECASTS FOR SECURITY ANALYSIS

In this section, we evaluate the usefulness of the k-NN* algorithm within the context of security analysis. We conduct two sets of tests. In the first set, we compare the earnings association coefficient (i.e., EAC) on the unexpected earnings implied by the k-NN* forecast with the EAC on the unexpected earnings implied by an alternate forecast. These analyses are inspired by the analyses described in Brown, Hagerman, Griffin, and Zmijewski (1987). In our second set of tests, we evaluate a trading strategy that is based on the forecasts generated by the k-NN* algorithm and then we compare the performance of this strategy with the performance of similar strategies that are based on forecasts generated by the HVZ model and the EP-LIFE model.

EAC-Based Tests

We estimate regression models in which each firm-year contributes *two* observations to the regression sample: (1) an observation for which unexpected earnings are based on the k-NN* forecast and (2) a second observation for which unexpected earnings are based on an alternate forecast (e.g., the forecast implied by the HVZ model). Hence, the number of observations in the regression sample is double the number of underlying firm-years. Specifically, we estimate variations of the following regression model:

$$BHAR_{i,t+1} = \gamma_1 \times UE_{f,i,t+1} + \gamma_2 \times UE_{f,i,t+1} \times Alt_{f,i,t} + \zeta_{f,i,t+1} \tag{4}$$

In Equation (4), the variable $BHAR_{i,t+1}$ is firm i's market-adjusted buy-and-hold stock return for the 12-month period beginning three months after the end of fiscal year t. The variable $UE_{f,i,t+1}$ denotes unexpected earnings, and it equals the (deflated by $MVE_{i,t}$) difference between firm i's realized *EBSI* for year t+1 and either the k-NN* forecast (when f = k-NN*) or the alternate forecast (when f = alternate) of *EBSI* for year t+1. The variable $Alt_{f,i,t}$ is equal to 0 (1) when the observation pertains to the k-NN* (alternate) forecast that was made at the end of year t. Equation (4) includes firm and year fixed effects, and we interact each fixed effect with the variable $Alt_{f,i,t}$. Hence, to avoid perfect collinearity, we do not include the variable $Alt_{f,i,t}$ as a separate regressor. We cluster the standard errors by firm and by year.

[18] Specifically, we standardize each variable by subtracting the panel average of the variable from its raw value and then dividing this amount by the panel standard deviation of the variable.

TABLE 8

Determinants of Absolute and Relative Forecast Accuracy

Dep. Variable:	|*k-NN*FE*|	|*k-NN*FE*| − |*RWFE*|	|*k-NN*FE*| − |*LIFEFE*|	|*k-NN*FE*| − |*HVZFE*|	|*k-NN*FE*| (*I/B/E/S*)	|*k-NN*FE*| − |*ANALYSTFE*|
Intercept	−0.02	0.00	0.00	0.00	−0.02	0.01
	[−1.13]	[0.03]	[0.47]	[0.22]	[−0.95]	[0.38]
MAD	0.12***	0.07*	0.11**	0.14***	0.46***	0.19***
	[5.00]	[1.94]	[2.45]	[2.94]	[20.57]	[6.92]
AbsFEG	0.06***	−0.40***	−0.12***	−0.15***	−0.02*	−0.04*
	[2.90]	[−10.66]	[−4.90]	[−4.71]	[−1.85]	[−1.87]
FOLLOW	−0.02***	−0.01	−0.00	−0.01**	−0.05***	0.04***
	[−2.75]	[−1.00]	[−0.60]	[−2.30]	[−5.29]	[3.40]
BP	0.06***	0.01	0.00	0.01	−0.02**	0.01
	[2.90]	[0.88]	[0.24]	[0.79]	[−1.99]	[0.51]
LnMVE	−0.02	−0.01	−0.00	−0.00	−0.14***	0.00
	[−1.43]	[−0.88]	[−0.39]	[−0.13]	[−6.24]	[0.24]
SPI	0.01	−0.04***	0.04***	0.05***	0.04**	0.02
	[0.94]	[−3.50]	[4.11]	[5.61]	[2.47]	[1.58]
R&D	0.00	0.01	−0.02**	−0.02	−0.08***	0.01
	[0.19]	[1.21]	[−2.39]	[−1.55]	[−7.62]	[1.42]
LOSS	0.04***	0.08***	−0.09***	−0.11***	−0.01	0.00
	[2.69]	[2.68]	[−4.14]	[−4.16]	[−0.25]	[0.12]
AbsACC	0.20***	0.02	−0.01	−0.01	0.07***	0.02
	[11.97]	[1.51]	[−0.41]	[−0.44]	[5.30]	[0.79]
RetVol	0.04***	−0.01	0.04***	0.04***	0.02*	0.02**
	[4.07]	[−0.76]	[4.00]	[4.09]	[1.77]	[2.00]
AbsLEG	0.12***	−0.04	−0.00	−0.04	0.03*	0.02
	[4.78]	[−1.43]	[−0.04]	[−1.06]	[1.96]	[1.38]
DISP						−0.02
						[−1.36]
Adjusted R^2	0.22	0.12	0.01	0.01	0.34	0.03
n	58,490	58,490	58,490	58,490	43,612	34,931

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 8 reports coefficients from regressions of the absolute value of the scaled forecast errors of the k-NN* algorithm and of the absolute value of the scaled forecast error differences between the k-NN* algorithm and the alternate approach on firm characteristics. All continuous explanatory variables are Winsorized at the 1st and 99th percentile. All the continuous dependent and independent variables are standardized (subtracting the panel mean of the variable and dividing the difference by the panel standard deviation of the variable). The number of observations differs from that reported in Tables 4 and 5 due to requiring return volatility data. t-statistics are reported in brackets and are based on standard errors clustered by firm and by year.

Equation (4) is similar to a seemingly unrelated regression (i.e., SUR) model as described in Zellner (1962). However, Equation (4) does not generate a separate estimate of the EAC for the unexpected earnings implied by the k-NN* forecast and the unexpected earnings implied by the alternate forecast. Rather, it estimates the EAC for the unexpected earnings implied by the k-NN* forecast (i.e., γ_1) and the difference between this EAC and the EAC on the unexpected earnings implied by the alternate forecast (i.e., γ_2). Hence, Equation (4) allows us to directly test the difference between the two EACs. Moreover, it allows us to easily embed fixed effects and interaction terms and to easily calculate clustered standard errors.[19]

In Table 9, we show the estimated coefficients and R^2s from various versions of Equation (4). In Panel A, we compare the EAC on the unexpected earnings implied by the k-NN* forecast with the EAC on the unexpected earnings

[19] In a set of untabulated results, we find that using a SUR model leads to the same inferences as Equation (4). However, the two-way clustered standard errors from Equation (4) are more conservative (i.e., larger) than the standard errors from the SUR model.

TABLE 9

Contemporaneous Earnings-Return Relation Tests

Panel A: Regressions of Abnormal Returns on Unexpected Earnings

	(1) k-NN* Only	(2) k-NN* versus RW	(3) k-NN* versus HVZ	(4) k-NN* versus EP-LIFE
UE	1.080***	1.080***	1.080***	1.080***
	(8.459)	(8.459)	(8.459)	(8.459)
UE × *ALT*		−0.108***	−0.116***	−0.145***
		(−3.010)	(−4.905)	(−5.243)
n	60,001	120,002	118,700	120,002
Adjusted R^2	0.048	0.048	0.046	0.045
Firm × *ALT* F.E.	Yes	Yes	Yes	Yes
Year × *ALT* F.E.	Yes	Yes	Yes	Yes

Panel B: Regressions of Abnormal Returns on Unexpected Earnings Interacted with *MAD4*

	k-NN* Only	k-NN* versus RW	k-NN* versus HVZ	k-NN* versus EP-LIFE
UE	2.576***	2.576***	2.576***	2.576***
	(6.022)	(6.022)	(6.022)	(6.022)
MAD4	0.317***	0.317***	0.317***	0.317***
	(5.098)	(5.098)	(5.098)	(5.098)
UE × *MAD4*	−1.816***	−1.816***	−1.816***	−1.816***
	(−4.006)	(−4.006)	(−4.006)	(−4.006)
UE × *ALT*		−0.028	−0.050	−0.150***
		(−0.590)	(−1.425)	(−3.115)
MAD4 × *ALT*		0.006*	0.033***	0.029***
		(1.937)	(6.593)	(6.649)
UE × *MAD4* × *ALT*		−0.051	−0.049	0.042
		(−0.505)	(−1.256)	(0.687)
UE	2.576***	2.576***	2.576***	2.576***
	(6.022)	(6.022)	(6.022)	(6.022)
n	60,001	120,002	118,700	120,002
Adjusted R^2	0.064	0.118	0.116	0.116
Firm × *ALT* F.E.	Yes	Yes	Yes	Yes
Year × *ALT* F.E.	Yes	Yes	Yes	Yes

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

Table 9 tabulates the estimated coefficients from regressions of returns on unexpected earnings. The dependent variable in all models is *BHAR*, the market-adjusted buy-and-hold abnormal return from three months after fiscal year end t to three months after fiscal year t+1, calculated as the firm's 12-month cumulative return less the 12-month cumulative return on the CRSP value-weighted market index. Unexpected earnings (*UE*) are realized $SEBSI_{i,t+1}$ less $FSEBSI_{i,t+1}$ based on the relevant forecasting approach. *MAD4* is defined as the *MAD* quartile, scaled to range evenly from 0 (low *MAD*) to 1 (high *MAD*). Firms are sorted into *MAD* quartiles by year. Columns labeled "k-NN* Only" present results from panel regressions that include one observation per firm-year and that base *UE* on the forecasts from the k-NN* algorithm. Columns labeled "k-NN* versus" present results from panel regressions that include two observations per firm-year: (1) one observation where *UE* is calculated based on the forecast from the k-NN* algorithm and (2) one observation where *UE* is calculated based on the forecast from the alternate approach. *ALT* is an indicator variable equal to 1 if that observation's *UE* is calculated using the alternate approach and 0 if that observation's *UE* is calculated using the k-NN* algorithm. All regressions are estimated with pooled OLS and include firm and year fixed effects interacted with the *ALT* dummy variable. Standard errors are clustered by firm and by year.

implied by the alternate forecasts. In the first column, we estimate the baseline EAC for the k-NN* forecasts. To do this, we estimate a baseline version of Equation (4), in which there is only *one* observation per firm-year and unexpected earnings is based on the k-NN* forecast. The EAC on the unexpected earnings implied by the k-NN* forecast is 1.08 (t-statistic of 8.46).

In the remaining columns of Panel A, we show the results from estimating full specifications of Equation (4), in which we compare the EAC implied by the k-NN* forecasts with the EAC implied by the random walk, HVZ, and

EP-LIFE forecasts. We note that γ_2, which is the coefficient on $UE_{f,i,t+1} \times Alt_{f,i,t}$, is always negative and statistically significant. This indicates that the EAC implied by each of the alternate forecasts is significantly smaller than EAC implied by the k-NN* forecast, which implies that the k-NN* forecasts are the most accurate.

In Panel B, we evaluate how the EACs vary with our *ex ante* indicator of forecast accuracy—i.e., $MAD_{i,t}$. To do this, we calculate the variable $MAD4_{i,t}$, which equals the annual quartile rank of $MAD_{i,t}$ scaled to range between 0 and 1. Then, we include $MAD4_{i,t}$ as a separate regressor in Equation (4), and we interact it with the other regressors. The results demonstrate that the *ex ante* indicator of forecast accuracy matters. For example, consider the results in the first column, which are generated by a baseline regression in which we only include one observation per firm-year and we base unexpected earnings on the k-NN* forecasts. The results shown in this column indicate that the EAC implied by the k-NN* forecasts is 2.58 (t-statistic of 6.02) for firms in the lowest MAD quartile but only 0.76 (i.e., 2.58 plus −1.82) for firms in the highest MAD quartile. Hence, when the *ex ante* expected accuracy of a k-NN* forecast is high (low), the EAC on the unexpected earnings implied by this forecast is also high (low). In other words, MAD is a useful *ex ante* predictor of forecast accuracy.

In the remaining columns of Panel B, we show the results from estimating regressions in which we compare the effect of $MAD4_{i,t}$ on the EAC implied by the k-NN* forecasts and the EACs implied by the alternate forecasts. The negative and statistically significant coefficients on $UE_{f,i,t+1} \times MAD4_{i,t}$ imply that the results shown in the first column are robust—i.e., the EAC implied by the k-NN* forecasts is higher (lower) when $MAD_{i,t}$ indicates high (low) forecast accuracy. However, the insignificant coefficients on $UE_{f,i,t+1} \times MAD4_{i,t} \times ALT_{f,i,t}$ imply that the differences between the EAC implied by the k-NN* forecasts and the EACs implied by the alternate forecasts do not vary with $MAD_{i,t}$.

Trading Strategy

We evaluate trading strategies that are based on the forecasts generated by the k-NN* algorithm, the HVZ model, and the EP-LIFE model. We use a calendar-time, portfolio-formation approach. As discussed in Fama (1998), Kausar, Taffler, and Tan (2009), and Mitchell and Stafford (2000), this approach controls for cross-sectional dependence in event-firm returns, and thus, it is more conservative than conventional approaches that evaluate annual buy-and-hold abnormal returns.

When implementing the trading strategy, we assume that a firm's earnings and financial statements for a particular fiscal year become publicly available three months after the end of that year and thus forecasts of EBSI for fiscal year t+1 become available three months after the end of fiscal year t. Then, for each calendar month between 1998 and 2018 and for each firm, we calculate the difference between the firm's *latest available* forecast of *EBSI* and its *latest available* realization of *EBSI*. For example, suppose a firm's fiscal year ends on June 30, 2010. Then, for each month occurring between October 2010 and September 2011, we calculate the difference between the forecast of *EBSI* that became available on September 30, 2010 and reported *EBSI* for the fiscal year that ended on June 30, 2010.

Next, for each calendar month between 1998 and 2018, we use the difference described above to assign each firm to either the UP or DOWN portfolio. Firms in the UP (DOWN) portfolio have positive (negative) differences—i.e., the latest available forecast of *EBSI* implies that their earnings will go up (down). We note that, after being initially assigned to either the UP or DOWN portfolio, a firm remains in that portfolio for the next 12 months, at which time it is reassigned. Then, after 12 more months, the firm is reassigned again, etc. We also note that, at any point in time, the UP (and DOWN) portfolio will consist of firms for which the latest forecast became available recently (e.g., became available one month earlier) and firms for which the latest available forecast is fairly stale (e.g., became available 11 months earlier).

After creating the UP and DOWN portfolios, we tabulate the monthly excess returns and alphas for monthly hedge portfolios that take equal-weighted long (short) positions in firms that are in the UP (DOWN) portfolios (i.e., UP − DOWN).[20] Additionally, we show UP-DOWN hedge portfolio returns and alphas for different quartiles of MAD. Specifically, for each calendar month, we determine the value of MAD that pertains to the latest available k-NN* forecast of *EBSI* and then we independently assign firms to MAD quartiles.

In Table 10, Panel A, we show the average monthly excess returns and alphas earned by the k-NN* hedge portfolio (i.e., the hedge portfolio that is based on the k-NN* forecasts). Then, in Panel B, we compare these returns and alphas with those earned on hedge portfolios that are formed on the basis of the forecasts generated by either the HVZ model or the EP-LIFE model. Three results stand out. First, focusing on Panel A and the column with the heading "All" (i.e.,

[20] Excess returns equal the difference between the raw portfolio return and the contemporaneous risk-free rate. We calculate monthly alphas relative to the Fama-French three-factor model. Results are qualitatively similar (yet somewhat weaker) when we use the Fama-French five-factor model. This is expected. Specifically, the Fama-French five-factor model includes a profitability factor that is likely associated with the earnings growth expectations captured by our portfolio assignments.

TABLE 10

Return Predictability Tests

Panel A: Excess Hedge Returns and Fama-French Adjusted Hedge Returns for k-NN* Algorithm

		MAD Quartile				
	All	**Low**	**Q2**	**Q3**	**High**	**Low − High**
UP – DOWN	0.26***	0.91***	0.24	0.31***	−0.05	0.96*
k-NN*	[2.90]	[2.82]	[1.49]	[2.73]	[−0.19]	[1.94]
UP – DOWN	0.34***	1.10***	0.36***	0.31***	−0.15	1.25***
k-NN* (FF3)	[4.33]	[4.06]	[2.63]	[2.66]	[−0.66]	[3.21]

Panel B: k-NN* Algorithm Hedge Returns versus Alternate Model Hedge Returns

		MAD Quartile				
	All	**Low**	**Q2**	**Q3**	**High**	**Low − High**
UP – DOWN	0.13	0.99***	0.28**	0.04	−0.10	1.09***
k-NN* – HVZ	[1.00]	[2.87]	[2.15]	[0.27]	[−0.81]	[3.02]
UP – DOWN	0.02	0.81**	0.17	−0.01	−0.14	0.95**
k-NN* – EP-LIFE	[0.18]	[2.20]	[1.34]	[−0.06]	[−1.24]	[2.43]
UP – DOWN	0.26**	1.12***	0.33***	0.13	−0.10	1.23***
k-NN* – HVZ (FF3)	[2.48]	[3.85]	[2.79]	[0.95]	[−0.81]	[3.73]
UP – DOWN	0.13	1.00***	0.25**	0.02	−0.17	1.17***
k-NN* – EP-LIFE (FF3)	[1.50]	[3.47]	[2.32]	[0.17]	[−1.52]	[3.71]

***, **, * Denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.
Table 10 reports average future monthly excess returns (Panels A and B) and Fama-French adjusted hedge returns ("alphas") for firms sorted into portfolios on forecasted growth and (independently) on the median absolute deviation (*MAD*) of the k-NN* algorithm. Excess returns are returns in excess of the risk-free rate. Portfolios are formed each month according to the observed forecasted positive (UP) or negative (DOWN) growth of the respective forecasting approach (k-NN*, HVZ, or LIFE). Portfolios are rebalanced monthly. Quintiles of *MAD* are formed monthly and independently of the forecasted growth portfolios. Fama-French adjusted hedge returns are the returns that cannot be explained by the factor returns of the Fama-French three-factor model (commonly referred to as alphas: the intercept of the regression of the respective monthly portfolio return on the monthly returns of the Fama and French three factors). The Fama and French three-factor model factors are Rm-Rf, SMB, and HML. Fama-MacBeth t-values are reported in parentheses.

ignoring variation in $MAD_{i,t}$), we find that, regardless of whether or not we adjust for risk, the hedge portfolio (i.e., UP – DOWN) that is formed on the basis of the k-NN* forecasts always earns positive and statistically significant returns.

Second, the variable $MAD_{i,t}$ is important. When it indicates high (low) forecast accuracy, the k-NN* hedge portfolio earns relatively high (low) returns. For example, when this portfolio consists of firm-years in the lowest *MAD* quartile (i.e., *highest ex ante* accuracy), it earns an excess monthly return of 0.91 and the Fama-French three-factor alpha is 1.10. These amounts are statistically significant. Moreover, as shown in the column "Low – High," the differences between these excess returns and alphas and the excess returns and alphas earned by the hedge portfolios that consist of firm-years in the highest *MAD* quartile (i.e., *lowest ex ante* accuracy) are (1) all positive and (2) all statistically significant.

Finally, as shown in Panel B, the returns earned on the k-NN* hedge portfolio are larger than those earned on either the HVZ hedge portfolio or the EP-LIFE hedge portfolio. When we focus on all firm-years, the differences between the returns on the k-NN* hedge portfolio and those on the HVZ and EP-LIFE hedge portfolios are relatively small and not always statistically significant. However, when we consider variation in $MAD_{i,t}$, we find that the differences are large and statistically significant (small and statistically insignificant) for the low (high) *MAD* quartiles. This further buttresses our conclusion that $MAD_{i,t}$ is a useful *ex ante* indicator of forecast accuracy.

IX. SUMMARY AND CONCLUSIONS

Expected earnings play a central role in many business decisions, and they are a key variable of interest in many academic studies. Moreover, practitioners often use comparable firms to gain insights about a subject firm. Nonetheless, evidence about the use and usefulness of comparable-firm-based forecasts is sparse.

In this study, we examine the efficacy of comparable-firm-based forecasts. We eschew complicated approaches and use a simple k-NN algorithm that matches each subject firm-year to firm-years with comparable current and lagged earnings. We do this for three reasons. First, there is substantial and compelling empirical evidence that earnings are a useful performance indicator, and *a priori*, it is reasonable to assume that firms with similar past performance will have similar future performance. Second, simplicity is a virtue. Simple approaches are easy to understand, use, explain, and modify, and they are less subject to overfitting. Finally, we use k-NN because it is a natural and objective way of integrating comparable firms into the forecasting process.

Despite (or perhaps because of) its simplicity, the k-NN* algorithm performs well. Its forecasts are significantly more accurate than forecasts generated by the random walk, more complicated k-NN algorithms, the approach developed by BCG, and extant regression models. Moreover, its forecasts of longer horizon *EPS* and aggregate *EPS* are more accurate than those developed by analysts. The k-NN* algorithm is also unique in the sense that it self-assesses via the variable $MAD_{i,t}$, which is a reliable *ex ante* indicator of forecast accuracy. This ability to self-assess is useful because it provides equity investors with useful information about when they should put more (less) weight on the forecasts generated by k-NN*.

Our results offer new insights into the usefulness of reported earnings. The simple k-NN* algorithm is best. Adding more predictors or using a longer earnings history does not lead to better forecasts. Hence, a firm's recent earnings history is informative about what its future earnings will be. The trick to uncovering this information is to put this history into the correct context by identifying firms with similar histories.

REFERENCES

Albrecht, W. S., L. L. Lookabill, and J. C. McKeown. 1977. The time-series properties of annual earnings. *Journal of Accounting Research* 15 (2): 226–244. https://doi.org/10.2307/2490350

Arachchige, C. N. P. G., L. A. Prendergast, and R. G. Staudte. 2022. Robust analogs to the coefficient of variation. *Journal of Applied Statistics* 49 (2): 268–290. https://doi.org/10.1080/02664763.2020.1808599

Ball, R., and R. Watts. 1972. Some time series properties of accounting income. *The Journal of Finance* 27 (3): 663–681. https://doi.org/10.1111/j.1540-6261.1972.tb00991.x

Barber, B. M., and J. D. Lyon. 1996. Detecting abnormal operating performance: The empirical power and specification of test statistics. *Journal of Financial Economics* 41 (3): 359–399. https://doi.org/10.1016/0304-405X(96)84701-5

Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Blouin, J., J. E. Core, and W. Guay (BCG). 2010. Have the tax benefits of debt been overestimated? *Journal of Financial Economics* 98 (2): 195–213. https://doi.org/10.1016/j.jfineco.2010.04.005

Bradshaw, M. T., M. S. Drake, J. N. Myers, and L. A. Myers. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies* 17 (4): 944–968. https://doi.org/10.1007/s11142-012-9185-8

Brown, L. D., R. L. Hagerman, P. A. Griffin, and M. E. Zmijewski. 1987. An evaluation of alternative proxies for the market's assessment of unexpected earnings. *Journal of Accounting and Economics* 9 (2): 159–193. https://doi.org/10.1016/0165-4101(87)90004-8

Chen, G. H., and D. Shah. 2018. *Explaining the Success of Nearest Neighbor Methods in Prediction*. Hanover, MA: Now Publishers.

Dickinson, V. 2011. Cash flow patterns as a proxy for firm life cycle. *The Accounting Review* 86 (6): 1969–1994. https://doi.org/10.2308/accr-10130

Evans, M. E., K. Njoroge, and K. O. Yong. 2017. An examination of the statistical significance and economic relevance of profitability and earnings forecasts from models and analysts. *Contemporary Accounting Research* 34 (3): 1453–1488. https://doi.org/10.1111/1911-3846.12307

Fairfield, P. M., S. Ramnath, and T. L. Yohn (FRY). 2009. Do industry-level analyses improve forecasts of financial performance? *Journal of Accounting Research* 47 (1): 147–178. https://doi.org/10.1111/j.1475-679X.2008.00313.x

Fairfield, P. M., R. J. Sweeney, and T. Yohn. 1996. Accounting classification and the predictive content of earnings. *The Accounting Review* 71 (3): 337–355.

Fama, E. F. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49 (3): 283–306. https://doi.org/10.1016/S0304-405X(98)00026-9

Hastie, T., R. Tibshirani, and J. Friedman. 2015. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hou, K., M. A. van Dijk, and Y. Zhang (HVZ). 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics* 53 (3): 504–526. https://doi.org/10.1016/j.jacceco.2011.12.001

Kausar, A., R. J. Taffler, and C. Tan. 2009. The going-concern market anomaly. *Journal of Accounting Research* 47 (1): 213–239. https://doi.org/10.1111/j.1475-679X.2008.00317.x

Li, K. K., and P. Mohanram (LM). 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies* 19 (3): 1152–1185. https://doi.org/10.1007/s11142-014-9282-y

Mitchell, M. L., and E. Stafford. 2000. Managerial decisions and long-term stock price performance. *The Journal of Business* 73 (3): 287–329. https://doi.org/10.1086/209645

Silver, N. 2003. Introducing PECOTA. In *Baseball Prospectus 2003*, edited by G. Huckabay, C. Kahrl, and D. Pease, 507–514. Dulles, VA: Brassey's Publishers.

Silver, N. 2008. Frequently asked questions. FiveThirtyEight.com. https://fivethirtyeight.com/predictors/frequently-asked-questions-last-revised/ (last accessed November 27, 2020).

So, E. C. 2013. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics* 108 (3): 615–640. https://doi.org/10.1016/j.jfineco.2013.02.002

Vorst, P., and T. L. Yohn (VY). 2018. Life cycle models and forecasting growth and profitability. *The Accounting Review* 93 (6): 357–381. https://doi.org/10.2308/accr-52091

Watts, R. L., and R. W. Leftwich. 1977. The time series of annual accounting earnings. *Journal of Accounting Research* 15 (2): 253–271. https://doi.org/10.2307/2490352

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association* 57 (298): 348–368. https://doi.org/10.1080/01621459.1962.10480664

APPENDIX A

Variable Definitions

Panel A: Variables and Subscripts to Describe Matching Algorithms

Variables	Definition
i,t	Subscript for subject firm (i) and year (t).
M	Years of earnings history.
j,s	Subscript for neighbor firm (j) and year (s).
$DIST^{M}_{i,t,j,s}$	Euclidean distance between the subject firm-year's most recent M-year earnings history and neighbor firm j's M-year earnings history ending in year s.
$MAD_{i,t}$	Median absolute deviation of the nearest neighbors realized $SEBSI$ at time s+1.
k	Number of nearest neighbors with the smallest value of $DIST^{M}_{i,t,j,s}$.
h	Forecast horizon in years.
k^*	The value of k for which an increase in k does not lead to a statistically significant decrease in $MAFE$ (at the 5 percent level) for a given M.
M^*	The value of M for which k* is the lowest.

Panel B: Financial Variables

Variables	Definition	Construction
$EBSI_{i,t}$	Earnings before special items for firm i at time t.	$ib_{i,t} - spi_{i,t}$
$MVE_{i,t}$	Equity market value for firm i at the end of fiscal year t.	$prcc\_f_{i,t} * csho_{i,t}$
$SEBSI_{i,t}$	$EBSI_{i,t}$ scaled by $MVE_{i,t}$.	$(ib_{i,t} - spi_{i,t})/MVE_{i,t}$
$FSEBSI_{i,t+h}$	Forecast of $EBSI_{i,t+h}$ scaled by $MVE_{i,t}$.	
$FEBSI_{i,t+h}$	Forecast of $EBSI_{i,t+h}$.	$FSEBSI_{i,t+h} * MVE_{i,t}$
$Age_{i,t}$	Number of annual Compustat appearances for the firm i from first year in Compustat to the end of year t.	
$ACC_{i,t}$	Accruals for firm i at time t scaled by $MVE_{i,t}$.	$[\Delta(act_{i,t} - che_{i,t}) - \Delta(lct_{i,t} - dlc_{i,t} - txp_{i,t}) - dp_{i,t}]/MVE_{i,t}$
$TA_{i,t}$	Total assets for firm i at time t scaled by $MVE_{i,t}$.	$at_{i,t}/MVE_{i,t}$
$DIV_{i,t}$	Dividends for firm i at time t scaled by $MVE_{i,t}$.	$dvc_{i,t}/MVE_{i,t}$
$DD_{i,t}$	Indicator variable equal to 1 for dividend payers and 0 otherwise at time t.	$\mathbb{1}(DIV_{i,t} > 0)$
$LOSS_{i,t}$	Indicator variable equal to 1 for firms with negative $SEBSI_{i,t}$ and 0 otherwise.	$\mathbb{1}(SEBSI_{i,t} < 0)$
$PM_{i,t}$	Profit margin for firm i at time t.	$EBSI_{i,t}/sale_{i,t}$

(continued on next page)

APPENDIX A (continued)

Variables	Definition	Construction
$ATO_{i,t}$	Asset turnover for firm *i* at time *t*.	$sale_{i,t}/at_{i,t}$
$LEV_{i,t}$	Leverage for firm *i* at time *t*.	$at_{i,t}/ceq_{i,t}$
$SGrow_{i,t}$	Sales growth for firm *i* at time *t*.	$sale_{i,t}/sale_{i,t-1} - 1$
$FEG_{i,t}$	Forecasted dollar earnings growth for firm *i* at time *t* scaled by $MVE_{i,t}$.	$FSEBSI_{i,t+1} - SEBSI_{i,t}$
$AbsFEG_{i,t}$	Absolute value of $FEG_{i,t}$.	
$AbsLEG_{i,t}$	Absolute value of realized dollar earnings growth for firm *i* at time *t* scaled by $MVE_{i,t}$.	$(EBSI_{i,t} - EBSI_{i,t-1})/MVE_{i,t}$
$BP_{i,t}$	Book-to-market ratio of firm *i* at time *t*.	$ceq_{i,t}/(prcc\_f_{i,t} * csho_{i,t})$
$FOLLOW_{i,t}$	Indicator variable equal to 1 if firm *i* at time *t* is followed by at least one analyst and 0 otherwise.	
$LnMVE_{i,t}$	Natural logarithm of $MVE_{i,t}$.	
$SPI_{i,t}$	Indicator variable equal to 1 for nonzero special items and 0 otherwise at time *t*.	$\mathtt{1}(\text{abs}(spi_{i,t}) > 0)$
$R\&D_{i,t}$	R&D expenditures scaled by total assets for firm *i* at time *t*.	$xrd_{i,t}/at_{i,t}$
$AbsACC_{i,t}$	Absolute values of $ACC_{i,t}$ for firm *i* at time *t*.	
$RetVol_{i,t}$	Standard deviation of returns during the fiscal year for firm *i* at time *t*.	
$AbsLEG_{i,t}$	Absolute value of realized dollar earnings growth for firm *i* at time *t* scaled by $MVE_{i,t}$.	
$DISP_{i,t}$	Standard deviation of individual analyst forecasts for firm *i* at time *t* divided by the maximum of the absolute of the mean consensus forecast or 0.1.	
$BHAR_{i,t+1}$	The market-adjusted buy-and-hold abnormal return for firm *i* from three months after fiscal year end *t* to three months after fiscal year *t*+1, calculated as the firm's 12-month cumulative return less the 12-month cumulative return on the CRSP value-weighted market index.	
$Alt_{i,t}$	Indicator variable equal to 0 (1) when the observation pertains to the k-NN (alternate) approach.	
$UE_{f,i,t+1}$	The unexpected earnings (deflated by $MVE_{i,t}$) implied by either the k-NN algorithm (i.e., when *f* = k-NN*) or the alternate approach (i.e., when *f* = alternate).	
$MAD4_{i,t}$	The annual quartile rank of $MAD_{i,t}$ scaled to range between 0 and 1. That is, we subtract one from the annual quartile rank of $MAD_{i,t}$, and then we divide this amount by three. Hence, $MAD4_{i,t} \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$.	

Panel C: Forecast Evaluation Metrics

Variables	Definition	Construction
MAFE	Mean absolute forecast error (percent of $MVE_{i,t}$).	$\text{Mean}(\lvert EBSI_{i,t+h} - FEBSI_{i,t+h}\rvert / MVE_{i,t}) * 100$
MDAFE	Median absolute forecast error (percent of $MVE_{i,t}$).	$\text{Median}(\lvert EBSI_{i,t+h} - FEBSI_{i,t+h}\rvert / MVE_{i,t}) * 100$
MSE	Mean of squared forecast error.	$\text{Mean}(((EBSI_{i,t+h} - FEBSI_{i,t+h}) / MVE_{i,t})^2) * 100$
TMSE	Mean of squared forecast error after truncating the top and bottom 0.1 percent signed forecast errors.	
IMPROVE	The percentage of observations for which the approach is more accurate than the alternate approach and the absolute value of the difference in the forecast error exceeds 0.50 percent of the subject firm-year's equity market value.	
|*k-NN*FE*|	The absolute value of the scaled forecast error generated by the k-NN* algorithm for subject firm-year *i*,*t*.	

(continued on next page)

APPENDIX A (continued)

Variables	Definition	Construction
$\|k\text{-}NN^*FE\| - \|RWFE\|$	$\|k\text{-}NN^*FE\|$ minus the absolute value of the scaled forecast error generated by the random walk model for subject firm-year *i,t*.	
$\|k\text{-}NN^*FE\| - \|LIFEFE\|$	$\|k\text{-}NN^*FE\|$ minus the absolute value of the scaled forecast error generated by the EP-LIFE model for subject firm-year *i,t*.	
$\|k\text{-}NN^*FE\| - \|HVZFE\|$	$\|k\text{-}NN^*FE\|$ minus the absolute value of the scaled forecast error generated by the HVZ model for subject firm-year *i,t*.	
$\|k\text{-}NN^*FE\|$ (I/B/E/S)	The absolute value of the scaled forecast error generated by the k-NN* algorithm for subject firm-year *i,t* for I/B/E/S EPS forecasts.	
$\|k\text{-}NN^*FE\| - \|ANLYSTFE\|$	$\|k\text{-}NN^*FE\|$ (IBES) minus the absolute value of the error in the consensus analyst forecast for subject firm-year *i,t*.	

Lowercase variables in the construction column refer to Compustat identifiers.