Forecasting Earnings Using k-Nearest Neighbors

Peter D. Easton
University of Notre Dame†
peaston@nd.edu

Martin M. Kapons
University of Amsterdam††
M.M.Kapons@uva.nl

Steven J. Monahan
University of Utah‡
steven.monahan@eccles.utah.edu

Harm H. Schütt
Tilburg University‡‡
H.H.Schutt@uvt.nl

Eric H. Weisbrod
University of Kansas§§
eric.weisbrod@ku.edu

This Draft: July 26, 2021

We thank Jeff Coulton, Michael Ettredge, 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 2021 European Accounting Association Annual Congress; City, University of London, Cass Business School; Egyptian Online Seminars in Business, Accounting, and Economics; University of New South Wales; University of St. Gallen; Tilburg University; and the University of Kansas for helpful comments on earlier drafts.

† Mendoza College of Business, University of Notre Dame, IN 46556, USA.
†† Amsterdam Business School, University of Amsterdam, 1018 TV Amsterdam, Netherlands.
‡ David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, USA.
‡‡ Tilburg School of Economics and Management, Tilburg University, P.O. Box 90153 5000 LE Tilburg, Netherlands.
§§ University of Kansas, School of Business, Lawrence, KS 66045, USA.

Forecasting Earnings Using k-Nearest Neighbors

ABSTRACT

We use a simple k-nearest neighbors (k-NN) model to forecast a subject firm's annual earnings by matching its recent earnings history to earnings histories of comparable firms, and then extrapolating the forecast from the comparable firms' lead earnings. Out-of-sample forecasts generated by our model are more accurate than forecasts generated by the random walk; more complicated k-NN models; the matching approach developed by Blouin, Core, and Guay (2010); and popular regression models. These results are robust. Our model's superiority holds for different error metrics, for firms that are followed by analysts and firms that are not, and for different forecast horizons. Our model also generates a novel *ex ante* indicator of forecast inaccuracy. This indicator, which equals the interquartile range of the comparable firms' lead earnings, is reliable and useful. It predicts forecast accuracy and it identifies situations when our forecasts are strong (weak) predictors of future stock returns.

Keywords: earnings, forecasting, machine learning.

JEL Classifications: C21, C53, G17, M41.

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

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, and writing contracts. *Et cetera*. 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 models. Nonetheless, several fundamental issues are not well understood. In this study, we consider one of these issues: The use and usefulness of comparable-firm-based forecasts.

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 in-sample 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. These studies are important but they leave key questions unanswered such as: How do forecasts based on comparable firms perform relative to one another and to alternative models? Are forecasts based on comparable firms useful in the context of security analysis?

We use k-nearest neighbors (i.e., k-NN) to study the questions described above. k-NN is a simple, effective, and longstanding forecasting approach.[1] It involves matching the subject firm-

[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).

year to a set of nearest neighbors, and then basing the forecast of the subject firm-year's earnings on these neighbors' earnings. Matches are formed on the basis of one or more firm characteristics (or, in the parlance of machine learning, *features*). Hence, k-NN is a natural and objective way of integrating comparable firms into the earnings forecasting process: Select a set of nearest neighbors, and then extrapolate the forecast from their earnings.

When implementing k-NN the way that nearest neighbors are selected is central. We base our selection algorithm on an intuitive argument: Earnings reflect economic performance and firms with similar past performance are more likely to perform similarly in the future. Consequently, we implement our k-NN model by matching each subject firm-year to firm-years with similar earnings histories.[2] Then, we set our forecast equal to the median of lead earnings for the matched firm-years. We do not require that the matched firm-years be contemporaries. And, 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).[3]

We begin our empirical analyses by establishing some basic properties of our model. 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 overfitting, we use the "tuning sample," which contains only firm-years drawn from 1979 to 1995. Using minimum mean absolute forecast error (i.e., MAFE) as our criterion, we find that the optimal value of M is two and that the optimal value of k is 90. Hence, colloquially speaking, learning more about a firm is less important than learning

[2] We find matches on the basis of earnings deflated by equity market value. We obtain similar results (untabulated) when we use alternative deflators such as equity book value, total assets, or revenues.

[3] 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 firm in this set. In a set of untabulated results, we find that searching for matches from the set of *all* preceding firm-years that end h years before the subject firm-year leads to only a small improvement in forecast accuracy.

about more firms.

Second, we compare our model to four "naïve" k-NN models 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' current (instead of lead) earnings; (3) the median of lead earnings for all firm-years that end during the ten-year period that ends h years before the subject firm-year; and (4) the median of lead earnings for a randomly selected sample of k firm-years. We use the "testing sample," which contains firm-years drawn from 1998 to 2018. We find that, regardless of the value of k, our model is superior to all of the naïve models. This implies that both selecting *nearest* neighbors (instead of *all* or *random* neighbors) and extrapolating the earnings of these neighbors (i.e., using their *lead* earnings instead of their *current* earnings) are both important determinants of our model's performance.

Finally, we compare the subject firm-years to their nearest neighbors per our model. We make some interesting discoveries. Per the Fama and French 12 industry classification scheme, the average (median) percentage of nearest neighbors in the same industry as the subject firm is 16 (16) percent; and, when industry is defined on the basis of two-digit SIC code, this amount falls to 7 (4) percent. The average (median) fraction of nearest neighbors in the same lifecycle group (as defined in Dickinson 2011) as the subject firm is 37 (38) percent. The typical subject firm is 53 percent larger (in terms of equity market value) than its typical nearest neighbor. And, the average and median difference between the subject firm-year and the year corresponding to its typical nearest neighbor is approximately 5.54 years. Hence, our matching approach 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 strata. Whether this difference leads to more accurate forecasts is an interesting empirical question that we evaluate next.

We compare our model to a broad set of models including: (1) k-NN models that find matches using alternative (e.g., profit margin and asset turnover) or additional (e.g., accruals) features; (2) the 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 we show that our simple k-NN model generates more accurate forecasts than all of the alternative models. This result is robust. It holds regardless of the forecast horizon, which we vary from one to three years, and regardless of the criterion we use to evaluate accuracy—i.e., MAFE, median absolute forecast error (i.e., MDAFE), or mean square error (i.e., MSE).

After establishing that our model is more accurate than the alternative models on average and for the typical firm, we evaluate how its accuracy varies across firms. We evaluate unsigned forecast errors and we study both absolute and relative (to other models) accuracy. We evaluate a number of standard variables such as analyst following, the book-to-market ratio, equity market value, etc. We emphasize two variables: (1) the interquartile range of the nearest neighbors' lead earnings, $IQR_{i,t}$, and (2) the absolute value of forecasted earnings growth, $AbsFEG_{i,t}$. We emphasize these variables because they are novel and they measure important properties of our forecast. Specifically, given our forecast equals the median value of the nearest neighbors' lead earnings, $IQR_{i,t}$ is an *ex ante* indicator of forecast inaccuracy. Whereas the variable $AbsFEG_{i,t}$ measures the extremity of the forecast.

We find that our model's accuracy is strongly associated with both $IQR_{i,t}$ and $AbsFEG_{i,t}$. Although there is a marginally significant, small, negative association between absolute accuracy and $AbsFEG_{i,t}$, there is a highly significant, large, positive association between $AbsFEG_{i,t}$ and relative accuracy. This implies that when our model generates more extreme forecasts, it is *more* accurate than other models. Regarding $IQR_{i,t}$, we find that when it is low (high), our model performs significantly better (worse) on both an absolute and relative basis. Hence, $IQR_{i,t}$ is a reliable *ex ante* indicator of forecast inaccuracy.

In our final set of tests, we assess the usefulness of our forecasts within the context of security analysis. To do this, we evaluate the association between the forecast of earnings growth implied by our model, $FEG_{i,t}$, and future stock returns. We find that, on average, there is a marginally significant, small, positive relation between $FEG_{i,t}$ and future returns. However, when $IQR_{i,t}$ is low (i.e., when forecast accuracy is high), the relation between $FEG_{i,t}$ and future returns is highly significant, large, and positive. Hence, our model generates two complementary and useful outputs: The variable $FEG_{i,t}$ is a useful predictor of future stock returns and $IQR_{i,t}$ is a useful indicator of when $FEG_{i,t}$ is a better (worse) predictor.

We make four related contributions. First, we develop a k-NN forecasting model 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 $IQR_{i,t}$, which is a useful *ex ante* indicator of forecast inaccuracy.

Second, we provide initial evidence about the usefulness of k-NN models 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 non-parametric manner. Consequently, compared to regression models, k-NN models accommodate non-linearities better and are less sensitive to extreme values. Moreover, k-NN is a natural and objective way of integrating comparable firms into the forecasting process. Hence, the lack of evidence about the reliability of earnings forecasts generated by k-NN models is a noticeable gap in the literature that we begin to fill.

Third, we make a methodological contribution regarding the best way of identifying matched firms/samples, which is an important research design issue. Many studies assume that: (1) it is desirable to match on certain structural factors that are either unobservable or difficult to measure and (2) these factors vary systematically across industries or lifecycle groups. Hence, matching on either industry or lifecycle 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 features leads to *worse* forecasts. Whether these results have implications beyond the subject of earnings forecasting is an intriguing question that is beyond the scope of this study.

Last, but certainly not least, we demonstrate that a firm's recent earnings history is very informative about what its future earnings will be. The trick to uncovering this information is to put this history into the correct context and this can be done by identifying firms with similar histories.

II. FORECASTING MODELS

For each subject firm-year i, t, we compute an h-year-ahead forecast of earnings before special items and we refer to firm $i's$ realized earnings before special items for year t as $EBSI_{i,t}$. In this section, we describe the different models that we use to generate forecasts of $EBSI$. We begin by describing our k-NN model, and then we describe the models we compare it to. These models include alternative k-NN models; the model used by BCG; and the regression models used by HVZ, LM, FRY, and VY. We also compare our model to the random walk. We do not elaborate on the random walk except to state that its forecast equals the subject firm-year's current earnings.

Our k-NN Model

To form our k-NN forecast we follow a five-step process.[4] First, we identify the most recent M-year earnings history for subject firm-year i, t, and then we deflate all the dollar amounts in this history by firm i's equity market value at the end of year t.[5] We refer to these deflated amounts as scaled earnings before special items, $SEBSI$. Second, we identify the set of firm-years that have complete histories of $EBSI$ of length M ending in *any* year $s \in [t-h, t-9-h]$. These are the neighbors of firm-year i, t. Then, for each neighbor firm-year j, s, we deflate all of the dollar amounts in its earnings history by its equity market value at the end of year s.

Third, for each neighbor firm-year j, s, we calculate the variable $DIST^M_{i,t,j,s}$, which is the 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.

$$DIST^M_{i,t,j,s} = \sqrt{\sum_{m=1}^{M} \left(NSEBSI_{i,t-m+1} - NSEBSI_{j,s-m+1}\right)^2} \quad [1]$$

In equation [1], $NSEBSI_{i,t-m+1}$ ($NSEBSI_{i,t-m+1}$) is the *normalized* value in year $t-m+1$ ($s-m+1$) of $SEBSI$ for subject firm i (neighbor firm j).[6]

Fourth, we identify the k neighbors with the smallest values of $DIST^M_{i,t,j,s}$. These are the *nearest* neighbors of subject firm-year i, t. Fifth, we compute k intermediate forecasts by deflating each nearest neighbor's $EBSI$ for year $s+h$ by its equity market value at the end of year s. Finally, we set the variable $IQR_{i,t+h}$ equal to the interquartile range of the intermediate forecasts; and we

---

[4] For an illustration of how we implement k-NN, please refer to Appendix B. In this appendix, we provide an example in which we make a forecast in 2010 of Walmart's earnings for 2011.

[5] We obtain similar results (untabulated) when we use alternative deflators such as equity book value, total assets, or revenues.

[6] We normalize $SEBSI$ by subtracting its contemporaneous cross-sectional average from its raw value, and then dividing this difference by the contemporaneous cross-sectional standard deviation. (This is a common way of implementing k-nearest neighbors.) The resulting normalized value has a mean of zero and a standard deviation of one. Consequently, both the current and lagged normalized values have the same scale, and thus have the same influence on $DIST^M_{i,t,j,s}$.

compute our forecast by multiplying the median of the intermediate forecasts by the subject firm's (i.e., firm $i's$) equity market value at the end of year t.

Alternative k-NN Models

Our k-NN model uses only one feature (i.e., scaled earnings) to identify the subject firm-year's nearest neighbors. The advantage of using a single feature is that it leads to the best possible matches *with respect to this feature*, whereas when more features are used the matches with respect to *every* feature are worse. Bellman (1957) refers to this phenomenon as the curse of dimensionality. The disadvantage of using a single feature is that we ignore information embedded in other features. Hence, whether it is better to add a feature depends on whether its incremental information content exceeds the error introduced by having to find matches across more dimensions. This is ultimately an empirical question.

With the above question in mind, we compare our k-NN model to eight alternative k-NN models. To implement each model, we follow a five-step process that is similar to the process described in the previous subsection. The primary difference is that we use different sets of features to identify nearest neighbors, and thus we use a modified distance measure, $MDIST_{i,t,j,s}^{F,M}$, that reflects multiple features.

$$MDIST_{i,t,j,s}^{F,M} = \sqrt{\sum_{f=1}^{F} \sum_{m=1}^{M} \left(NFEAT_{i,t-m+1}^{f} - NFEAT_{j,s-m+1}^{f}\right)^2} \qquad [2]$$

In equation [2], F denotes the number of features and $NFEAT_{i,t-m+1}^{f}$ ($NFEAT_{j,s-m+1}^{f}$) is the *normalized* value in year $t-m+1$ ($s-m+1$) of feature f for subject firm i (neighbor firm j).

We separate the eight sets of alternative features into two categories: (1) the DuPont category and (2) the HVZ category. The DuPont category is inspired by the well-known DuPont decomposition of return on equity (e.g., Fairfield and Yohn 2001; Nissim and Penman 2001; and Soliman 2008). Its root feature set consists of profit margin, PM, and asset turnover, ATO; and then there are three additional sets in which percentage sales growth, $SGrow$, and/or leverage, LEV, are added to the root feature set. The feature sets in the HVZ category consist of different combinations of the predictor variables used by HVZ. It's root feature set consists of $SEBSI$ and accruals, ACC; and then three additional feature sets are formed by progressively adding total assets, TA; dividends, DIV; and a loss indicator, $LOSS$. The variables ACC, TA, and DIV are deflated by equity market value at the end of either year t (for subject firm i) or year s (for neighbor firm j).

BCG's Model

BCG's model is similar to a k-NN model in the sense that they base their forecast of the subject firm's earnings on the earnings of a set of matched firms. To implement it, we follow the four-step process outlined 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 performance-size bins: Ten negative $SEBSI$-size bins and twenty positive $SEBSI$-size bins. Third, we randomly select a sample of 50 observations from the performance-size bin to which subject firm-year i, t belongs and we calculate the median earnings growth from $t-2$ to $t-1$ for this sample. 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

[7] BCG rank observations on the basis of their return on assets, ROA, for year $t-2$ and they use the average growth rate instead of the median growth rate. We rank on $SEBSI$ for year $t-2$ so that our BCG model is comparable to the other models that we evaluate. We use the median growth rate because the average growth rate is sensitive to extreme values and generates forecasts that are much less accurate than those based on the median growth rate. 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 forecasts generated by our k-NN model.

model described in LM. The HVZ model is widely adopted and is often referred to as the benchmark model for regression-based forecasts (e.g., Evans, Njoroge, and Yong 2017 and So 2013). The EP model is also popular and it forms the basis for the industry- and lifecycle-based forecasts described in FRY and LV, respectively.

The HVZ (EP) forecasts are obtained using the estimated coefficients from the ordinary least squares (i.e., OLS) regression shown in equation three (four) below.[8]

$$SEBSI_{i,t+h} = \alpha_0 + \alpha_1 \times TA_{i,t} + \alpha_2 \times DD_{i,t} + \alpha_3 \times DIV_{i,t} + \alpha_4 \times SEBSI_{i,t} + \alpha_5 \times LOSS_{i,t} + \alpha_6 \times ACC_{i,t} + \varepsilon_{i,t} \quad [3]$$

$$SEBSI_{i,t+h} = \beta_0 + \beta_1 \times SEBSI_{i,t} + \beta_2 \times LOSS_{i,t} + \beta_3 \times \left(SEBSI_{i,t} \times LOSS_{i,t}\right) + \epsilon_{i,t} \quad [4]$$

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 one (zero) 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 one (zero) if $SEBSI_{i,t}$ is (is not) negative; $ACC_{i,t}$ denotes firm i's scaled accruals for year t; and $\varepsilon_{i,t}$ ($\epsilon_{i,t}$) 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.

As discussed in the next subsection, the estimated coefficients that form the basis of our regression-based forecasts are obtained by estimating regressions on rolling samples of "training data." To estimate the HVZ regression coefficients we use the entire sample of training data. Following FRY and VY, we estimate three versions of the EP model. In the first version, we use

[8] In a set of untabulated results, we find that forecasts generated by median (instead of OLS) regressions are less accurate than the forecasts generated by our k-NN model.

the entire sample of training data. In the other two versions, we use a subset of the training data that is restricted to those firm-years that either have the same Global Industry Classification Standard (i.e., 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

When developing our k-NN forecasts and our regression-based forecasts, 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, following HVZ, we define the cross-section of data for year t as all of the firm-years that ended their fiscal year between April 1 of calendar year $t-1$ and March 31 of calendar year t.[9] We refer to *all* of these firm-years (including those with fiscal years that ended in calendar year $t-1$) as year t firm-years; and we assign the time subscript t to the corresponding variables. We then identify nearest neighbors and estimate regressions coefficients using the sample of training data that consists of the cross-sections of data for years $t-9-h$ through year $t-h$ (h is the length of the forecast horizon in years).

III. SAMPLE CONSTRUCTION, VARIABLE DEFINITIONS, AND DESCRIPTIVE STATISTICS

Sample Construction

We obtain data about U.S.-incorporated companies from the Compustat Fundamentals Annual file. Our testing sample consists of two subsets: (1) the training data and (2) the forecast comparison sample. Because we need data from the cash flow statement to implement the lifecycle model, our first ten-year sample of training data begins in 1988. Because our rolling-window

[9] This assures that earnings for year t and our year t forecast of $EBSI$ for year $t+1$ are available by July 1 of calendar year t, which is the first month of the 12 consecutive months that we evaluate when testing the association between our year t forecasts and future monthly stock returns.

forecasting procedure uses ten years of training data, the forecast comparison sample spans the years 1998 through 2018. Consequently, the last ten-year sample of training data ends in 2017.

In Table 1, we summarize how we construct the forecast comparison sample, which consists of 62,710 firm-years.[10] In order to be able to implement and evaluate all of the forecasting models, we eliminate firm-years with missing: (1) lagged, current, or lead $EBSI$; (2) cash flow data; or (3) balance sheet accruals. For the same reason, we eliminate observations with missing or non-positive: (1) equity market value; (2) current or lagged sales; (3) total assets; or (4) equity book value. We also remove financial firms and regulated firms. To minimize the effect of database errors and small deflators, we eliminate firm-years with equity market value that is less than 10 million U.S. dollars or for which the absolute value of $SEBSI_{i,t}$ exceeds one. Finally, per FRY, we eliminate industry-years that have fewer than 100 observations with non-missing values of the variables in equation [4] or that have fewer than ten observations with negative current $SEBSI$.

We also construct a sample of data that we use to determine the optimal values of k and M for each of the k-NN models. This process is referred to as parameter tuning, so we refer to these data as the tuning sample. To construct the tuning sample, we use an algorithm that is similar to the algorithm described above. However, to avoid hindsight bias, we limit the forecast comparison subset of the tuning sample to firm-years between 1978 and 1995; thus, the first (last) ten-year sample of tuning training data begins in 1968 (ends in 1994). In addition, in order to be able to evaluate values of M between one and five years, we only include a firm-year in the tuning sample if its $EBSI$ for the current and previous four years are non-missing. The forecast comparison subset of the tuning sample contains 55,905 firm-years (untabulated).

[10] We use a similar algorithm to construct the rolling samples of training data. However, we tailor the data requirements to the model that is being trained. For example, none of the variables in the HVZ model are a function of sales. Hence, when constructing the data that we use to train the HVZ model, we do not eliminate firm-years with missing or non-positive current or lagged sales.

Variable Definitions

We define earnings before special items, $EBSI_{i,t}$, as the difference between Compustat data item $ib_{i,t}$ and Compustat data item $spi_{i,t}$. We set missing values of $spi_{i,t}$ to zero. With the exception of the variables that relate to the DuPont category of k-NN models, we deflate all firm-year i,t (j,s) variables used in the forecast models by firm $i's$ ($j's$) equity market value at the end of fiscal year t (s), $MVE_{i,t}$. (The variable $MVE_{i,t}$ equals the product of Compustat data items $prcc\_f_{i,t}$ and $csho_{i,t}$.) For the k-NN models in the DuPont category, we require the following variables: Profit margin, $PM_{i,t}$, which equals the ratio of $EBSI_{i,t}$ to firm $i's$ sales for year t (Compustat data item $sale_{i,t}$); asset turnover, $ATO_{i,t}$, which equals the ratio of $sale_{i,t}$ to contemporaneous total assets (Compustat data item $at_{i,t}$); sales growth, $SGrow_{i,t}$, which equals the ratio of $sale_{i,t}$ to its lagged value; and leverage, $LEV_{i,t}$, which equals the ratio of $TA_{i,t}$ to equity book value, (Compustat data item $ceq_{i,t}$).

For the k-NN models in the HVZ category, the HVZ regression model, and the EP regression model, we require the following variables: The indicator variable $LOSS_{i,t}$ is set equal to one (zero) if $SEBSI_{i,t} < 0$ ($SEBSI_{i,t} \geq 0$). We use the balance sheet method to calculate accruals. Consequently, $ACC_{i,t} = \{\Delta(act_{i,t} - che_{i,t}) - \Delta(lct_{i,t} - dlc_{i,t} - txp_{i,t}) - dp_{i,t}\}/MVE_{i,t}$ (the acronyms shown in brackets refer to Compustat data items).[11] Scaled total assets, $TA_{i,t}$, is Compustat data item $at_{i,t}$ divided by $MVE_{i,t}$. We set the dividend indicator, $DD_{i,t}$, to one (zero) if Compustat data item $dvc_{i,t} > 0$ ($dvc_{i,t} = 0$) and $DIV_{i,t} = dvc_{i,t}/MVE_{i,t}$. We set missing values of $dvc_{i,t}$ to zero. In Appendix A, we provide a complete list of all of the variables that we use and we describe how we compute each variable.

[11] We set missing values of Compustat items che, lct, dlc, txp and dp to zero.

Descriptive Statistics

In Panel A of Table 2, we provide descriptive statistics for the predictors in the HVZ regression model (which includes the predictors in the EP model and the features in the HVZ category of k-NN models) and the features in the DuPont category of k-NN models. We refer to these two sets of variables as the HVZ variables and the DuPont variables, respectively. 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 data that we use to train either a regression model or one of the k-NN models. Three comments are warranted. First, when constructing the training data for the models 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 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.[12]

In Panels B and C of Table 2, 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

[12] Our results are not attributable to extreme values. We base this conclusion on three facts. First, when estimating the regressions shown in equations [3] and [4], we first "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. (We also estimate the regressions on "unclean" data and, in untabulated results, we find that our model is the most accurate.) Second, as discussed in Section V, when evaluating forecast accuracy, we consider four different error metrics each of which weighs extreme forecast errors differently; and, regardless of the error metric used, our model is the most accurate. Finally, we conduct a battery of robustness tests (results untabulated) in which we define extreme values differently (e.g., removing observations with $|PM_{i,t}| > 1$, or $|SGrow_{i,t}| > 1$, or total assets or sales less than 10 million U.S. dollars, etc.); and we find that, regardless of the way we define extreme values, our model is the most accurate.

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-squareds) 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 zero and the level of significance for three of the coefficients is high. Specifically, the average coefficients (t-statistics) on scaled earnings, $SEBSI_{i,t}$, the loss indicator, $LOSS_{i,t}$, and the dividend indicator, $DD_{i,t}$, are 0.52 (86.85), -0.06 (-74.71), and 0.02 (87.77), respectively. On the other hand, the t-statistic on dividends, $DIV_{i,t}$, is much smaller (-4.13); and the coefficients on total assets, $TA_{i,t}$, and accruals, $ACC_{i,t}$, are insignificant. Finally, the average r-squared is 0.45.[13]

In Panel C, we summarize the results of the EP-model regressions that are estimated on all of the training data (i.e., the "full sample") as well as those that are estimated on the subsets of the training data that relate to the different lifecycle groups. For the full sample, the average coefficients (t-statistics) on $SEBSI_{i,t}$ and $LOSS_{i,t}$ are 0.50 (28.32) and -0.07 (-56.03), respectively. These coefficients vary considerably across lifecycle groups. For example, the coefficient on $SEBSI_{i,t}$ achieves its lowest value of 0.34 (t-statistic of 8.02) for the Decline group and its highest value of 0.58 (t-statistic of 22.14) for the Introduction group. The coefficient on the interaction between $SEBSI_{i,t}$ and $LOSS_{i,t}$ is insignificant for the full sample and three of the five lifecycle groups. However, it is negative and significant for the Mature group (average coefficient and

[13] When comparing our parameter estimates to those shown in HVZ, it is important to note that we deflate each of our continuous variables by equity market value (we check the sensitivity of our results to different deflators). HVZ, on the other hand, report results for regressions that are based on un-deflated variables. All of our inferences about forecast accuracy remain unchanged if we use forecasts generated by HVZ regressions that are based on un-deflated variables.

t-statistic of -0.22 and -11.21, respectively) and positive and significant for the Decline group (average coefficient and t-statistic of 0.17 and 3.20, respectively). Finally, the r-squared also varies considerably. For the full sample it is 0.45; and, when it is allowed to vary across lifecycle groups, it ranges between 0.21 (Mature group) and 0.42 (Introduction group).

Taken together, the results in Panels B and C suggest that regression models that either use multiple predictors or that allow the coefficients on the predictors to vary across lifecycle groups generate more accurate forecasts. However, whether these forecasts are more accurate than forecasts generated by our k-NN model is unclear. OLS regressions are predicated on strong assumptions about linearity and within-sample homogeneity. Our k-NN model, on the other hand, avoids these assumptions. Consequently, whether our k-NN model is more accurate than these popular regression models is an important empirical question.

IV. BASIC PROPERTIES OF OUR MODEL

Parameter Tuning

In order to implement our k-NN model, we must make three choices: (1) the feature(s) we use to identify neighbors; (2) the length of the history in the features (i.e., M) that we match on; and (3) the number of neighbors (i.e., k) that we match to the subject firm. Regarding the first choice, we make three assumptions: (1) earnings are a good indicator of past performance; (2) firms with similar past performance will have similar future performance; and (3) the incremental information content of other variables is too low to justify adding them to the set of features that we use to find matches. Based on these assumptions, we choose only one feature: $SEBSI$. The validity of these assumptions is an empirical question that we evaluate in Section V.

We are agnostic about the optimal values of M and k, and thus we use an empirical approach to choose them. Specifically, to avoid lookahead bias, we use the sample of tuning data; and then we evaluate the accuracy of our k-NN model for combinations of M and k. We consider values of M between one and five years and values of k between 10 and 200 neighbors. We vary M by increments of one and k by increments of 10. We define the best model as the combination of M and k that minimizes the mean of the absolute value of the scaled forecast errors (i.e., the MAFE). Each scaled forecast error equals 100 multiplied by the ratio of the difference between realized and forecasted $EBSI_{i,t+1}$ to $MVE_{i,t}$. That is, we express forecast errors as percentages of the subject firm-year's equity market value at the end of year t.

We document our parameter tuning results in the graph shown in Figure 1. In this graph, we show MAFE (k) on the vertical (horizontal) axis. For each of the five different values of M, we plot the MAFEs corresponding to the different values of k. Then we identify the optimal value of k, which we refer to as k^*.[14] The graph reveals four facts. First, and foremost, the k-NN model in which $M = 2$ and $k = 90$ is best. Second, regardless of the value of M, there is 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.[15] 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 Extrapolation

In this subsection, we evaluate the benefits of matching on $SEBSI$ and of extrapolating the forecast from the nearest neighbors' implied earnings growth rate. To do this, we use the forecast

[14] We define k^* as the value of k for which an increase in k does not lead to a statistically significant decrease in the MAFE (at the 5 percent level).

[15] Within the context of k-NN, dimensionality is a function of both the number of features that are used to find matches and the length of the history in the features (i.e., M).

comparison subset of the testing sample; and we compare the forecasts from our k-NN model to forecasts generated by alternative models in which we either use naïve matching approaches or do not extrapolate. To evaluate the importance of matching on $SEBSI$, we compare our k-NN model to two naïve models that are identical to our model except that their forecasts are based on the earnings of different sets of neighbors. Specifically, the Random (Economy-wide) k-NN model 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 extrapolating, we compare our model to a k-NN model that is identical except that it bases the forecast on the median of the k nearest neighbors' current (i.e., year s) earnings instead of their lead (i.e., year $s+1$) earnings. We refer to this as the No-extrapolation model. Finally, we also compare our model to the random walk, which is a naïve k-NN model that matches the subject firm to itself and does not extrapolate. Consequently, this comparison is informative about the benefits of both matching and extrapolation.

We document the results of the comparisons described above in the graph shown in Figure 2, which has the same format as the graph shown in Figure 1. Our model is noticeably better than all of the naïve models.[16] This has two implications. First, the way matches are formed matters. Although the MAFE of the Random k-NN model decreases as k increases, it converges to the MAFE of the Economy-wide k-NN model, which is greater than 10 and much larger than the MAFE of our k-NN model. Second, extrapolation also matters. Regardless of the value of k, forecasts based on the median of the nearest neighbors' current earnings are less accurate than forecasts generated by our k-NN model, which are based on the nearest neighbors' lead earnings.

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

Comparison of Subject Firm-years to their Nearest Neighbors

In Table 3, we compare the subject firm-years to their nearest neighbors. We evaluate the forecast comparison subset of the testing sample. The statistics shown in Panel A are computed via a three-step process. First, for each subject firm-year, we compute the variable shown in the first column of the panel for each of the 90 nearest neighbors. Second, we compute the median of these 90 amounts. This yields a sample of 62,710 medians—i.e., one for each subject firm-year. Finally, we compute and tabulate the mean, standard deviation, etc. of this sample of medians. To compute the statistics shown in Panel B, we first determine the percentage of nearest neighbors that are in the same group as the subject firm—i.e., same Fama-French industry, same two-digit Standard Industrial Classification (i.e., SIC) code, or same lifecycle group. We then compute and tabulate the mean, standard deviation, etc. of these percentages.

As shown in Panel A, the average, median, and inter-quartile range of the median differences in $SEBSI$ are each equal to 0.00. Thus, subject firm-years have 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 dissimilarities between the subject firm-years and their neighbors. For example, the average, median, and inter-quartile range of the median percentage differences in equity market value are -84, 53, and 122 percent, respectively. Hence, the size of the subject firm-year is usually very different from the size of its nearest neighbors. Moreover, per the distribution of the medians of the variable $(Year_{i,t} - Year_{j,s})$, the nearest neighbor firm-years typically precede the subject firm-year by more than five years. Finally, as shown in Panel B, the median percentage of nearest neighbors in the same Fama-French 12-industry group (with the same two-digit SIC code) as the subject firm-year is 16 percent (4 percent) and the median percentage of nearest neighbors in the same lifecycle group as the subject firm-year is 38 percent.

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 strata, or same lifecycle group.

V. COMPARISON TO ALTERNATIVE FORECASTING MODELS

In this section, we compare the forecasts generated by our k-NN model (i.e., the model in which we match on $SEBSI$ and set $M = 2$ and $k = 90$) to a number of alternative models. We use the forecast comparison subset of the testing sample and we evaluate four measures of forecast accuracy each of which weighs extreme errors differently: (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. To compute TMSE, we first delete the top and bottom 0.1 percent of the scaled forecast errors, and then we compute the mean square error. Each scaled forecast error equals 100 multiplied by the ratio of the difference between realized and forecasted $EBSI_{i,t+1}$ to $MVE_{i,t}$. That is, we express forecast errors as percentages of the subject firm-year's equity market value at the end of year t.

Comparison to Alternative k-NN Models

When implementing our k-NN model, we identify nearest neighbors by matching on $SEBSI$. This raises the question: What if we use more and/or different features? For example, empirical results in Fairfield and Yohn (2001) and Soliman (2008) imply that the DuPont decomposition is useful within the context of forecasting profitability. And, as shown in Table 2, four of the six predictors used in the HVZ model are associated with future earnings. Hence, if we use only $SEBSI$ to find neighbors, we discard information. That said, adding features implies higher dimensionality, which, in turn, implies worse matches. Consequently, whether using alternative features and/or adding features leads to more accurate forecasts is an empirical question.

With the above question in mind, we compare our k-NN model to eight alternative k-NN models. To implement these models, we follow the same five-step algorithm that we use to implement our k-NN model except we use different sets of features to identify nearest neighbors. In addition, for each model, we use the tuning sample and the approach described in Section IV to choose the optimal values of M and k.

We put the eight models into two categories: (1) the DuPont category and (2) the HVZ category. Both categories consist of four models and we refer to the models in the DuPont (HVZ) category as k-NN_{DUP1}, k-NN_{DUP2}, etc. (k-NN_{HVZ1}, k-NN_{HVZ2}, etc.). When implementing the k-NN_{DUP1} model, we identify nearest neighbors by using a root feature set that consists of profit margin, PM, and asset turnover, ATO. The k-NN_{DUP2} (k-NN_{DUP3}) model adds sales growth, $SGrow$, (leverage, LEV) to the root feature set; and the k-NN_{DUP4} model uses all four features. The k-NN_{HVZ1} model uses $SEBSI$ and ACC to find matches; and the models k-NN_{HVZ2} through k-NN_{HVZ4} are formed by progressively adding the features AT, DIV, and $LOSS$.

We present the results of these analyses in Table 4. In the top row of the table, we show the error metrics (e.g., MAFE, MDAFE, etc.) for our k-NN model. In the remaining rows we tabulate the differences between the error metrics for the corresponding alternative model and the error metrics for our k-NN model. The results lead to four conclusions. First, and foremost, our k-NN model is never less accurate and usually more accurate than all of the alternative k-NN models. Second, our k-NN model dominates the models in the DuPont category. For example, regardless of the error metric used, k-NN_{DUP1} is the best DuPont model. Nonetheless, its MAFE (MDAFE) is 19.1 (40.6) percent larger than the MAFE (MDAFE) of our k-NN model and its MSE (TMSE) is 17.0 (28.5) percent larger than the MSE (TMSE) of our k-NN model.

Third, the models in the HVZ category perform relatively well. For example, the k-NN_{HVZ1} and k-NN_{HVZ2} models are roughly as accurate as our k-NN model. That said, the k-NN_{HVZ3} and k-NN_{HVZ4} models are less accurate than our model and the differences are nontrivial. Finally, adding features reduces accuracy. Both the MAFE and MDAFE are increasing in the number of features used to find matches. Moreover, for the models in the DuPont category, the MSE and TMSE are also increasing in the number of features. Consequently, the curse of dimensionality is a genuine concern.

Comparison to Extant Models

In this subsection, we compare the forecasts generated by our k-NN model to forecasts generated by six well-known extant models: (1) the random walk (i.e., RW); (2) the BCG model (i.e., BCG); (3) the HVZ regression model (i.e., HVZ); (4) the EP regression model that is trained on an unrestricted sample of training data (i.e., EP); (5) the EP regression model that is trained on a set of training data that is restricted to firm-years with the same GICS code as the subject firm-year (i.e., EP-GICS); and (6) the EP regression model that is trained on a set of training data that is restricted to firm-years in the same lifecycle group as the subject firm-year (i.e., EP-LIFE).

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: Our k-NN model is superior to all of the extant models. Specifically, regarding MAFE, MDAFE, and TMSE, our model is always the best and all the differences are statistically significant. Moreover, our model always has the lowest MSE; however, one of the six differences is not statistically significant.

Analyst Following and Forecast Horizon

Given the necessity of having an accurate forecasting model 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 our k-NN model to the random walk, the HVZ model, and the EP-LIFE model. We include the random walk because it is a common benchmark. We include the remaining models because each is the best of its class—i.e., our k-NN model is the most accurate k-NN model and the HVZ (EP-LIFE) model is the most accurate regression model that uses an unrestricted (a restricted) sample of training data.

We present the results of the analyses described above 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 our k-NN model vary with analyst coverage. Two conclusions warrant discussion. First, each of the models is more accurate when it is used to forecast the earnings of firms that are covered by analysts. For example, the MAFE of our k-NN forecasts is 5.36 percent for the subsample of firms that are followed by analysts and 10.21 percent for the subsample of firms that are not. Second, the relative performance of our k-NN model does not depend on analyst coverage. That is, our k-NN model generates more accurate forecasts of the earnings of firms that are covered by analysts and those that are not. This is important because it implies that our k-NN forecasts are the preferred alternative when analysts' forecasts are unavailable.

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 model performs worse as the forecast horizon lengthens, which is not surprising. For example, the MAFE of our k-NN forecasts of earnings for years $t + 1$, $t + 2$, and $t + 3$ are 6.80, 8.73, and 10.15 percent, respectively. Nonetheless, and perhaps more important, regardless of the forecast horizon, our k-NN model continues to generate forecasts that are more accurate than those generated by the other models.

Cross-sectional Variation in Absolute and Relative Accuracy

In this subsection, we evaluate the association between 11 firm characteristics and the absolute and relative (to other models) accuracy of our k-NN model. To do this, we estimate four separate regressions. Each regression has a different dependent variable and the same 11 independent variables. In the first regression, the dependent variable is the absolute value of the scaled forecast error generated by our k-NN model for subject firm-year i, t, $\left|kNNFE_{i,t}\right|$. The dependent variables in regressions two through four are equal to the difference between $\left|kNNFE_{i,t}\right|$ and the absolute value of the scaled forecast error for subject firm-year i, t generated by the random walk model, $\left|RWFE_{i,t}\right|$, the HVZ regression model $\left|HVZFE_{i,t}\right|$, and the EP-LIFE regression model, $\left|LIFEFE_{i,t}\right|$, respectively.

We estimate OLS panel regressions and we report standard errors that are clustered by firm and year. For ease of interpretation, we standardize all the variables to have a mean of zero and a standard deviation of one. Hence, the variables have the same scale, and thus the coefficients are directly comparable. Specifically, each coefficient reflects the number of standard deviations that the dependent variable changes for a change of one standard deviation in the corresponding independent variable. Consequently, the absolute values of the coefficients can be used to rank the independent variables with respect to their relative importance.

We present the results in Table 7. In Panel A, we show descriptive statistics for the raw amounts of the dependent and independent variables. We make three comments. First, because we include return volatility, $RetVol_{i,t}$, in the set of independent variables, the sample we estimate our regressions on is smaller than the forecast comparison sample described in Table 1. Second, when calculating descriptive statistics, we express all the dependent variables in percentage terms.

Finally, the independent variables are: (1) The absolute value of forecasted earnings growth, $AbsFEG_{i,t}$; (2) the interquartile range of the intermediate forecasts, $IQR_{i,t}$; (3) An indicator that equals one (zero) if the firm is (is not) followed by at least one analyst, $FOLLOW_{i,t}$; (4) the book-to-market ratio, $BP_{i,t}$; (5) the natural log of equity market value, $LnMVE_{i,t}$; (6) an indicator that equals one (zero) if the firm reported (did not report) special items in year t; $SPI_{i,t}$; (7) research and development expense divided by total assets, $R\&D_{i,t}$; (8) $LOSS_{i,t}$; (9) the absolute value of accruals divided by equity market value, $AbsACC_{i,t}$; (10) $RetVol_{i,t}$; and (11) the absolute value of lagged dollar earnings growth divided by equity market value, $AbsLEG_{i,t}$.

In panel B of table 7, we show the coefficient estimates from regressions involving each of the dependent variables and the independent variables described in Panel A. Regarding absolute accuracy, three variables stand out in terms of their association with $\left|kNNFE_{i,t}\right|$: (1) $AbsACC_{i,t}$, which has a coefficient (t-statistic) of 0.17 (10.33); (2) $BP_{i,t}$, which has a coefficient (t-statistic) of 0.13 (6.49); and (3) $IQR_{i,t}$, which has a coefficient (t-statistic) of 0.12 (7.12). We draw three conclusions from these results. First, extreme accruals may be an indicator of low-quality earnings in the sense that when the absolute magnitude of the accrual component of earnings is large, matches based on the subject firm's recent earnings history are less informative about its future earnings. Second, our k-NN model performs worse for value stocks and better for glamour stocks. Finally, a one standard deviation increase in $IQR_{i,t}$ is associated with a 0.12 standard deviation increase in $\left|kNNFE_{i,t}\right|$, which implies that $IQR_{i,t}$ is a reliable *ex ante* indicator of forecast inaccuracy. That is, when the dispersion of the nearest neighbors' lead earnings is high, the forecast based on these earnings is less accurate.

Regarding relative accuracy, we draw three key conclusions. First, relative to the other models, our model is at its best when it makes extreme forecasts. Specifically, the coefficient (t-statistic) on $AbsFEG_{i,t}$ is -0.43 (-10.17) in the regression that compares our k-NN model to the random walk model. Moreover, the coefficients (t-statistics) on $AbsFEG_{i,t}$ in the remaining two regressions are -0.14 and -0.16 (-5.08 and -4.51), respectively. Second, in addition to being an important determinant of absolute accuracy, $IQR_{i,t}$ is also an important determinant of relative accuracy. For example, when we compare our model to the EP-LIFE and HVZ models, the coefficients (t-statistics) on $IQR_{i,t}$ are 0.15 and 0.16 (2.70 and 2.60), respectively. This buttresses our conclusion that $IQR_{i,t}$ is a reliable *ex ante* indicator of forecast inaccuracy. Finally, when compared to the random walk model (regression models) our model performs worse (better) for firms that are experiencing losses.

VI. RELATION WITH FUTURE STOCK RETURNS

As shown in the previous section, our k-NN model is useful in the sense that its forecasts are more accurate than the forecasts generated by each of the alternative models that we evaluate. In this section, we evaluate its usefulness within another context: Security analysis. We examine two outputs of our model: (1) the implied forecast of earnings growth, $FEG_{i,t}$, and (2) $IQR_{i,t}$, which, as shown in the previous section, is a reliable *ex ante* indicator of forecast inaccuracy. We focus on these two variables for two complementary reasons. First, results in Ball and Brown (1968), who show that there is a positive association between *realized* earnings growth and *contemporaneous* stock returns, provoke the question: Is there a positive association between *forecasts* of earnings growth and *future* stock returns? Second, to the extent there is a positive association, *a priori* logic suggests that the strength of this association is a decreasing function of forecast inaccuracy. Hence, we hypothesize that accurate forecasts of earnings growth are more useful for predicting future stock returns than inaccurate forecasts. To test this prediction, we estimate the following regression:

$$RET_{i,t+1} = \gamma_0 + \gamma_1 \times FEG_{i,t} + \gamma_2 \times IQR_{i,t} + \gamma_3 \times (FEG_{i,t} \times IQR_{i,t}) + \mathbf{\Gamma} \cdot \mathbf{C}_{\boldsymbol{i,t}} + \zeta_{i,t+1} \quad [5]$$

In equation [5], $RET_{i,t+1}$ is firm $i's$ market-adjusted stock return for each of the 12 months between July 1 of the calendar year in which the forecast is made and June 30 of the subsequent year.[17] The variable $FEG_{i,t}$, which is forecasted earnings growth, is the difference between our year t forecast of $EBSI$ for year $t+1$ and realized $EBSI$ for year t divided by $MVE_{i,t}$. The variable $IQR_{i,t}$ is defined above; $\mathbf{C}_{\boldsymbol{i,t}}$ **(Γ)** is a 4×1 (1×4) vector of control variables (regression coefficients); and $\zeta_{i,t+1}$ is the error term. Each independent variable is the scaled decile rank of the underlying raw variable, and thus it equals one of the following ten numbers $\left\{0, \frac{1}{9}, \cdots, \frac{8}{9}, 1\right\}$. Hence, each slope coefficient equals the return on a hedge portfolio that takes a long (short) position in stocks in the top (bottom) decile of the corresponding independent variable.

In Table 8, we summarize the estimates of the regression coefficients obtained from equation [5]. The coefficients (r-squareds) are the time-series averages generated by the monthly cross-sectional regressions. Each t-statistic equals the average coefficient divided by its time-series standard error. As shown in Column 1, when we do not allow the coefficient on $FEG_{i,t}$ to vary with $IQR_{i,t}$, there is a marginally significant, small, positive association between $FEG_{i,t}$ and $RET_{i,t+1}$. Hence, for the "average" firm, the forecasts generated by our model are relatively weak predictors of future stock returns. However, as shown in Column 2, when we allow the coefficient

[17] To assure that our year t forecasts of $EBSI$ in year $t+1$ are available, we follow HVZ and evaluate returns for the 12-month period that starts on July 1 of calendar year t and ends on June 30 of calendar year $t+1$; and, per the discussion in Section II, $EBSI$ for year t and our year t forecast of $EBSI$ for year $t+1$ are based on accounting reports that are available on or before June 30 of calendar year t. When estimating equation [5], we do not exclude observations with missing realized *ex post* earnings. Rather, we include all observations for which there is an observable forecast in year t of $EBSI$ for year $t+1$. We include CRSP delisting returns and we adjust for missing delisting returns following Shumway (1997) and Shumway and Warther (2002). Specifically, we assume a delisting return of -30 percent for NYSE and AMEX firms (-55 percent for NASDAQ firms) with missing performance-related delisting returns. We include delisting returns following the procedure in Beaver et al. (2007); and we assume that, starting on the delisting date, the proceeds from firms that delist are re-invested in the value-weighted market portfolio.

to vary with $IQR_{i,t}$, we find that when $IQR_{i,t}$ is low (high), the coefficient on $FEG_{i,t}$ is positive and significant (negative and insignificant). Moreover, in the regressions shown in Columns 2, the coefficient on $FEG_{i,t}$ is the hedge-portfolio return for low $IQR_{i,t}$ firms. Hence, per the results in Column 2, this hedge portfolio generates an average monthly market-adjusted return of 0.66 percent, which is equivalent to an annual market-adjusted return of 8.21 percent. These amounts are economically significant.

The results described above lead to two important conclusions. First, they support our prediction that accurate forecasts of earnings growth are more useful predictors of future stock returns than inaccurate forecasts. Second, and perhaps more important, they imply that our model has a unique built-in self-assessment feature. Specifically, the variable $IQR_{i,t}$ is informative about forecast accuracy, and thus it a useful indicator of when the forecasts generated by our model are more (less) useful.

VII. 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. Nonetheless, as discussed in Monahan (2018), most models either do not beat the random walk or do not beat it by much. Moreover, although practitioners often use comparable firms, 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 models and use a simple k-NN model in which we match the 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 models 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, our k-NN model performs very well. Its forecasts are significantly more accurate than forecasts generated by the random walk, more complicated k-NN models, the model developed by BCG, and extant regression models. These results are robust. Our k-NN model's superiority holds for different error metrics, for firms that are followed by analysts and for firms that are not, and for different forecast horizons. Moreover, our model is unique in the sense that it self-assesses via the variable $IQR_{i,t}$, which is a reliable *ex ante* indicator of forecast inaccuracy. This ability to self-assess is useful. For example, as we show, it can be used to identify situations in which our forecasts are strong (weak) predictors of future stock returns.

Our results offer new insights into the usefulness of reported earnings. Our simple k-NN model, which matches on the most recent two years of earnings, is best. Adding more features or using a longer earnings history does not lead to better forecasts. This implies that a firm's recent earnings history is very informative about what its future earnings will be. The trick to uncovering this information is to put this history into the correct context and this can be done by identifying firms with similar histories.

REFERENCES

Ball, R., and Brown, P., 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research*. 6(2): 159-178.

Barber, B., and Lyon, J., 1996. Detecting abnormal operating performance: the empirical power and specification of test statistics. *Journal of Financial Economics*. 41 (3), 359 – 399.

Beaver, W., McNichols, M. and Price, R., 2007. Delisting returns and their effect on accounting-based market anomalies. Journal of Accounting and Economics. 43 (2-3): 341-368.

Bellman, R., 1957. Dynamic programming. Princeton University Press.

Blouin, J., Core, J. and Guay, W., 2010. Have the Tax Benefits of Debt Been Overestimated? *Journal of Financial Economics.* 98 (2): 195-213.

Chen, G.H. and Shah, D., 2018. Explaining the success of nearest neighbor methods in prediction. Now Publishers.

Dickinson, V., 2011. Cash Flow Patterns as a Proxy for Firm Life Cycle. *The Accounting Review*. 86(6): 1969-1994.

Evans, M.E., Njoroge, K. and Yong, K.O., 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.

Fairfield, P.M., Ramnath, S. and Yohn, T., 2009. Do Industry-Level Analyses Improve Forecasts of Financial Performance? *Journal of Accounting Research*. 47(1): 147-178.

Fairfield, P.M. and Yohn, T., 2001. Using Asset Turnover and Profit Margin to Forecast Changes in Profitability. *Review of Accounting Studies*. 6: 371-385.

Hou, K., van Dijk, M.A. and Zhang, Y., 2012. The Implied Cost of Capital: A New Approach. *Journal of Accounting & Economics*. 53: 504-526.

Li, K. and Mohanram, P., 2014. Evaluating Cross-Sectional Forecasting Models for Implied Cost of Capital. *Review of Accounting Studies*. 19: 1152-1185.

Monahan, S.J., 2018. Financial Statement Analysis and Earnings Forecasting, *Foundations and Trends® in Accounting*. 12 (2): 105-215.

Nissim, D. and Penman, S.H., 2001. Ratio Analysis and Equity Valuation: From Research to Practice. *Review of Accounting Studies.* 6: 109-154.

Shumway, T., 1997. The delisting bias in CRSP data. *The Journal of Finance*. 52 (1): 327-340.

Shumway, T. and Warther, V.A., 2002. The Delisting Bias in CRSP's Nasdaq Data and Its Implications for the Size Effect. *The Journal of Finance*. 54 (6): 2361-2379.

Silver, N., 2003. Introducing PECOTA. *Baseball Prospectus*. 2003: 507 – 514.

Silver, N., 2008. Frequently Asked Questions. *FiveThirtyEight.com*. Accessed 11/27/2020 at https://fivethirtyeight.com/features/frequently-asked-questions-last-revised/.

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

Soliman, M.T., 2008. The Use of DuPont Analysis by Market Participants. *The Accounting Review*. 83(3): 823-853.

Vorst, P. and Yohn, T., 2018. Life Cycle Models and Forecasting Growth and Profitability. *The Accounting Review*. 93(6): 357-381.

APPENDIX A
Variable Definitions

Variable	Definition	Construction
Panel A: Variables and subscripts to describe matching models		
i,t	Subscript for firm (i) and time (t) for subject firm	
M	Years of earnings history	
j,s	Subscript for firm (j) and time (s) for neighbor firm	
$DIST_{i,t,j,s}^{M}$	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	
k	Number of nearest neighbors with the smallest value of $DIST_{i,t,j,s}^{M}$	
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)	

Variable	Definition	Construction
Panel B: Financial variables		
$EBSI_{i,t}$	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 $EBSI_{i,t+h}$	$FSEBSI_{i,t+h}$ * $MVE_{i,t}$
$ACC_{i,t}$	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*	$l(DIV_{i,t} > 0)$
$LOSS_{i,t}$	Indicator variable equal to 1 for firms with negative $SEBSI_{i,t}$ and 0 otherwise	$l(SEBSI_{i,t} < 0)$
$PM_{i,t}$	Profit margin for firm *i* at time *t*	$EBSI_{i,t} / sale_{i,t}$
$ATO_{i,t}$	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*	$sale_{i,t} / sale_{i,t-1} - 1$
$IQR_{i,t}$	Inter quartile range of forward *SEBSI* of comparable firms	
$FEG_{i,t}$	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 non-zero special items and 0 otherwise at time *t*	$l(abs(spi_{i,t}) > 0)$
$R\&D_{i,t}$	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*	
$Profitability_{i,t}$	$EBSI_{i,t}$ scaled by ending equity book value for firm *i* at time *t*	$(ib_{i,t} – spi_{i,t}) / ceq_{i,t}$
$Investment_{i,t}$	Percentage growth of total assets at the fiscal year end for firm *i* from time *t-1* to time *t*	$(at_{i,t} – at_{i,t-1}) / at_{i,t-1}$

Variable	Definition	Construction
Panel C: Forecast evaluation metrics		
MAFE	Mean absolute forecast error (% of $MVE_{i,t}$)	Mean($\lvert EBSI_{i,t+h}$ - $FEBSI_{i,t+h}\rvert$ / $MVE_{i,t}$) * 100
MDAFE	Median absolute forecast error (% of $MVE_{i,t}$)	Median($\lvert EBSI_{i,t+h}$ - $FEBSI_{i,t+h}\rvert$ / $MVE_{i,t}$) * 100
MSE	Mean of squared forecast error	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% signed forecast errors	

Lowercase variables in the construction column refer to Compustat identifiers.

APPENDIX B
Illustration of k-NN for Walmart 2010

In this appendix, we provide an illustration of k-NN. Specifically, we use it to make a forecast in 2010 of Walmart's $EBSI$ for 2011. In this example, we set $M = 5$ and $k = 10$. Hence, we compare the five-year history of Walmart's $SEBSI$ that ends in 2010 to *all* the observable five-year histories of $SEBSI$ that *end* in *any* year $s \in \{2000, 2001, \cdots, 2009\}$, and then we use the variable $DIST^{M}_{i,t,j,s}$, which is described in Equation [1], to identify Walmart's ten nearest neighbors.

In Panel A of Figure B.1, we plot, in event time, the scaled earnings of Walmart along with those of its ten nearest neighbors. In Panel B, we show the different calendar time periods that relate to each of these neighbors. As shown in Figure B.1, Walmart's ten nearest neighbors come from a broad range of industries and are drawn from as early as $s = 2001$ and as late as $s = 2009$. Moreover, Walmart itself is one of its own nearest neighbors—i.e., the Walmart history that ends in 2009 is one of the histories that is closest to the Walmart history that ends in 2010.

To compute the variable IQR and our forecast of Walmart's $EBSI$ for 2011, we first compute ten intermediate forecasts by deflating each of the nearest neighbor's realized $EBSI$ for year $s + 1$ by its equity market value at the end of year s. As shown in Panel A of Figure B.1, these intermediate forecasts range between 0.015 to 0.084. Next, we set IQR equal to 0.022, which is the interquartile range of the intermediate forecasts. Finally, we compute our forecast by multiplying the median of the intermediate forecasts, which equals 0.076, by 202,286 million U.S. dollars, which is Walmart's equity market value at the end of 2010. Hence, in this example, the 2010 forecast of Walmart's $EBSI$ (in millions of U.S. dollars) for 2011 is $15{,}354 = 0.076 \times 202{,}286$.

Figure B.1
Walmart 2010

Figure B.1 illustrates the nearest neighbor matching approach. It shows the example of forecasting Walmart's 2011 earnings as of $t_0 = 2010$. For exposition purposes we plot the 10 nearest firm sequences with their respective ending years. Sub-figure A plots the earnings sequences aligned in sequence time. Earnings sequences that are more similar to Walmart's 2006 to 2010 period are colored darker. Sub-figure B shows the calendar year timing of the 10 nearest earnings sequences inside the rolling 10-year window used to train the k-NN model. The dashed line for each neighbor's sequence represents the $s + 1$ period used to compute the median forecast for Walmart's $t + 1$ earnings.

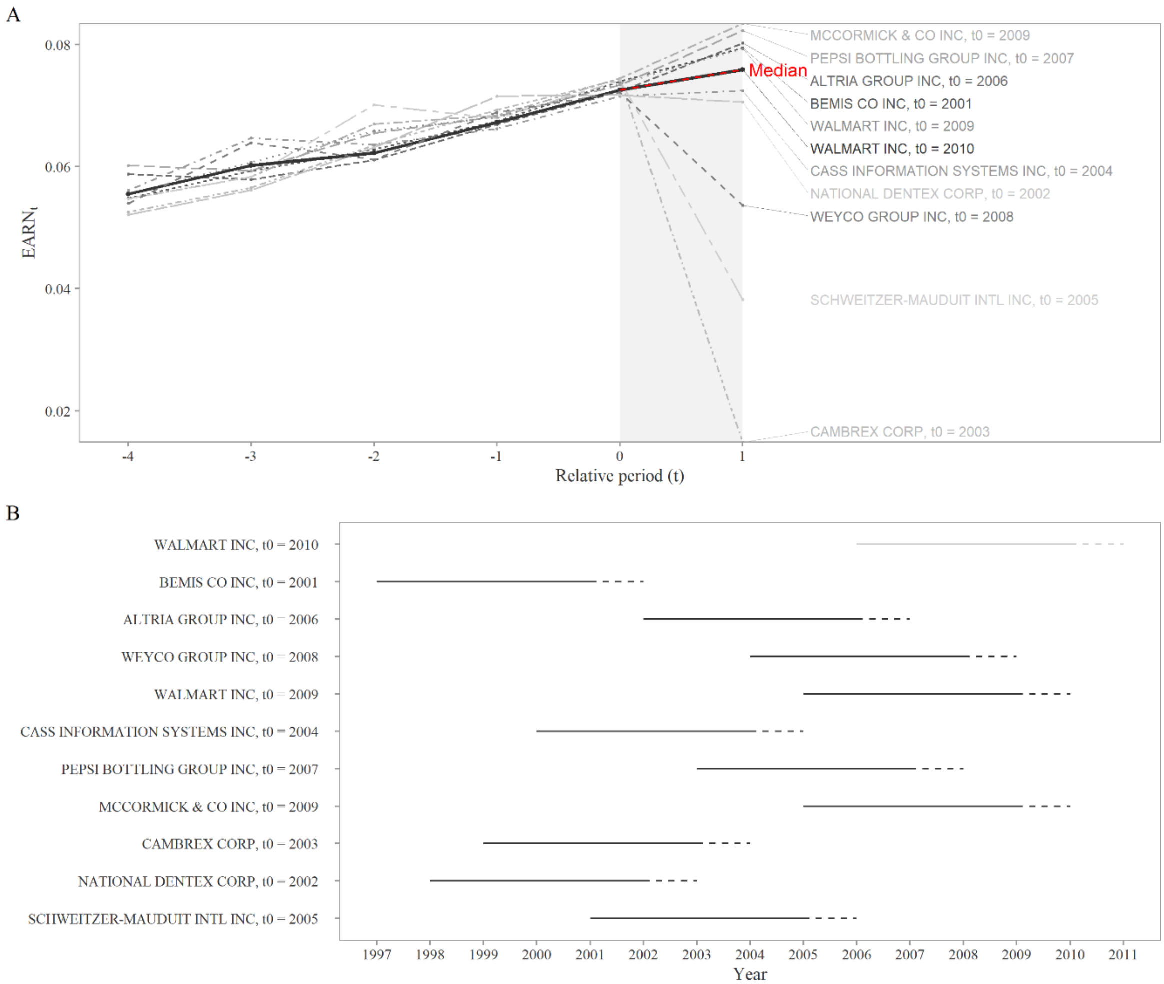

Figure 1
Mean Absolute Forecast Error by K and M

Figure 1 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 neighborss (k*) at which decreases in MAFE become insignificant for increasing K by another 10 neighbors, given a value of M.

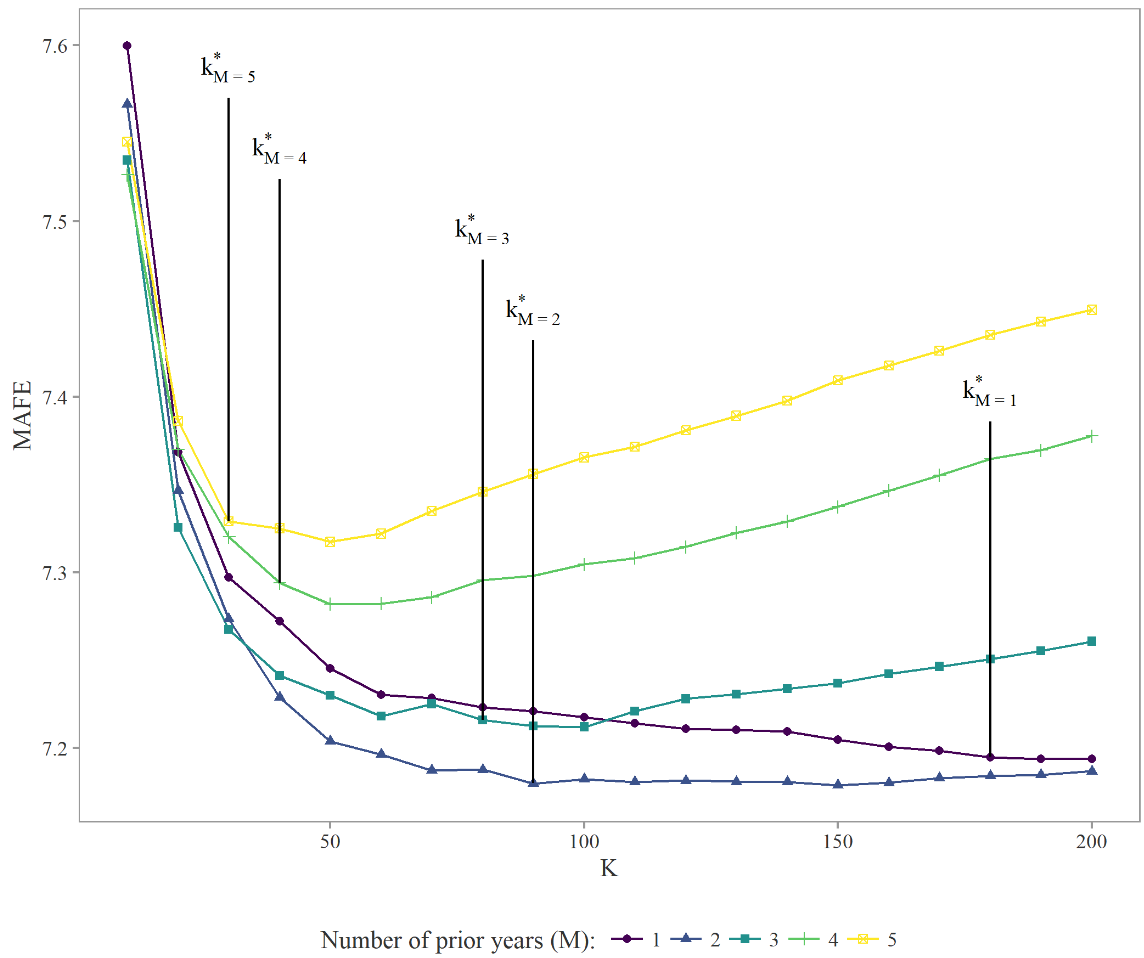

Figure 2
Decomposition of MAFE by Components of the Matching Model

Figure 2 compares the mean absolute $i, t+1$ forecast error (MAFE) of a random walk forecast (RW) with various simple forecast models that use the median $s+1$ or s year's earnings of K selected firms j. The number of selected firms K is varied to show the influence of the number of nearest neighbors. We select firms either randomly or using the k-NN method to show the impact of selecting nearest neighbors. Nearest neighbors are selected using earnings as the matching variable and the model considers most recent earnings and lagged earnings ($M = 2$). We examine choices of a forecast median ($s+1$ or s year's earnings) to show the impact of extrapolating the selected firms' trends.

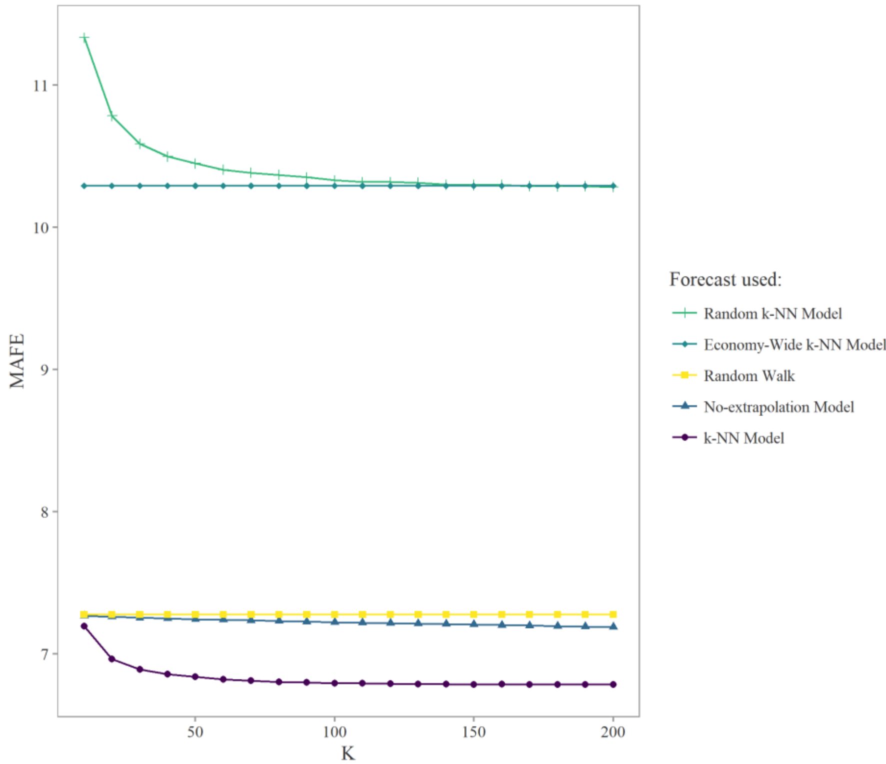

Table 1
Sample Composition

Data Filter	Firm-Years
Total Compustat Observations 1998 - 2018	**182,167**
Less missing *EBSI*	-32,835
Random walk forecast sample	**149,332**
Less missing lagged *EBSI*	-9,224
Less missing and non-positive *MVE*	-12,221
Less financial firms and regulated firms	-31,309
KNN (M=2) forecast sample	**96,578**
Industry forecast sample	**96,343**
Less missing Cash Flow Data	-149
Lifecycle/HVZ forecast sample	**96,194**
Less missing and non-positive sales, total assets, and equity book value	-16,478
Less missing and non-positive lagged sales	-870
Less missing balance sheet accruals	-1,705
Less *MVE < $10M*	-6,702
Less absolute scaled earnings greater than one	-724
Less missing future *EBSI*	-5,769
Less Industry-years with <100 (<10) total (loss) observations	-1,125
Less Missing BCG Forecast Data	-111
Forecast Comparison Sample	**62,710**

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

Variable	N	Mean	StD	P05	P25	Med	P75	P95
ACC	117,734	-0.04	14.17	-0.41	-0.09	-0.02	0.00	0.13
DD	117,734	0.26	0.44	0.00	0.00	0.00	1.00	1.00
DIV	117,734	0.01	0.08	0.00	0.00	0.00	0.00	0.04
SEBSI	117,734	-0.03	0.19	-0.41	-0.06	0.03	0.07	0.14
LOSS	117,734	0.37	0.48	0.00	0.00	0.00	1.00	1.00
TA	117,734	1.53	31.44	0.11	0.44	0.89	1.68	4.44
PM	106,986	-3.94	162.70	-1.84	-0.03	0.03	0.07	0.19
ATO	106,986	1.18	0.93	0.14	0.59	1.02	1.54	2.76
LEV	106,986	4.93	275.93	1.13	1.40	1.91	2.80	7.19
SGrow	106,986	1.24	60.38	-0.31	-0.02	0.10	0.27	1.25

Panel B: Coefficients of the in-sample estimation of the HVZ regression model

Average N	Intercept	ACC	DD	DIV	SEBSI	LOSS	TA	Adj. R^2
39,907	0.01	-0.00	0.02	-0.01	0.52	-0.06	-0.00	0.45
	[19.55]	[-1.22]	[87.77]	[-4.13]	[86.85]	[-74.71]	[-1.63]	

Panel C: Coefficients of the in-sample estimation of the EP regression model

Lifecycle	Average N	Intercept	SEBSI	LOSS	LOSS*SEBSI	Adj. R^2
	41,192	0.02	0.50	-0.07	0.02	0.45
		[16.96]	[28.32]	[-56.03]	[0.93]	
Introduction	8,537	-0.01	0.58	-0.05	0.01	0.42
		[-7.49]	[22.14]	[-28.64]	[0.20]	
Growth	11,218	0.02	0.44	-0.04	0.03	0.24
		[29.39]	[26.44]	[-52.30]	[1.48]	
Mature	13,557	0.03	0.54	-0.04	-0.22	0.21
		[40.70]	[41.47]	[-45.79]	[-11.21]	
Shake-Out	4,567	0.01	0.53	-0.06	-0.03	0.36
		[10.17]	[28.15]	[-31.18]	[-1.18]	
Decline	3,314	-0.01	0.34	-0.04	0.17	0.35
		[-7.50]	[8.02]	[-28.22]	[3.20]	

Table 2, Panel A provides pooled summary statistics for the variables included in the regression and k-NN models. Panel B (Panel C) shows the average coefficients of the 10-year rolling window regressions for the HVZ (EP) model. T-statistics are derived from Fama-MacBeth standard errors. See Table A.1 for remaining variable definitions.

Table 3
Comparison of Subject Firms to their Nearest Neighbors

Panel A: Descriptive statistics for comparable and subject firms

Variable	N	Mean	StD	P05	P25	Med	P75	P95
$SEBSI_{i,t}$ - $SEBSI_{j,s}$	62,710	0.00	0.02	0.00	0.00	0.00	0.00	0.00
$(MVE_{i,t} - MVE_{j,s})/MVE_{i,t}$	62,710	-0.84	5.85	-6.05	-0.38	0.53	0.84	0.98
$(acc_{i,t} - acc_{j,s})/MVE_{i,t}$	62,710	-0.02	0.24	-0.27	-0.06	-0.01	0.02	0.22
$(FEG_{i,t} - FEG_{j,s})$	62,710	0.00	0.09	-0.01	0.00	0.00	0.00	0.01
$(Year_{i,t} - Year_{j,s})$	62,710	5.54	1.08	4.00	5.00	5.50	6.00	7.00
$(Age_{i,t} - Age_{j,s})$	62,710	5.54	13.08	-11.00	-3.00	2.00	12.00	33.00

Panel B: Industry and lifecycle membership

Variable	N	Mean	StD	P05	P25	Med	P75	P95
percent same FF12	62,710	16.46	10.48	2.22	8.89	15.56	22.22	36.67
percent same Sic2	62,710	7.07	7.69	0.00	1.11	4.44	10.00	23.33
percent same lifecycle	62,710	36.98	18.53	6.67	22.22	37.78	52.22	65.56

Table 3, Panel A presents descriptive statistics about the comparable firms chosen by our comparable matching procedure. The suffix *i,t* denotes the subject firm-year and the suffix *j,s* denotes the relevant matched neighbor firm-year. Panel B reports the percentage of matched comparable firms that are in the same industry (lifecycle) as the subject firm. FF12 is the Fama-French-12 Industry classification, and Sic2 is the 2-digit SIC industry code. *t* is the first year of the two-year earnings sequence. See Table A.1 for remaining variable definitions.

Table 4
Comparison of Our k-NN Model to Alternative k-NN Models (t+1)

Model	N	MAFE	MDAFE	MSE	TMSE
KNN	62,710	6.80	2.56	3.42	1.72
KNN_{DUP1}	62,710	1.30***	1.04***	0.58***	0.49***
KNN_{DUP2}	62,710	1.53***	1.08***	0.72***	0.63***
KNN_{DUP3}	62,710	2.02***	1.24***	0.97***	0.87***
KNN_{DUP4}	62,710	2.05***	1.27***	0.98***	0.88***
KNN_{HVZ1}	62,710	0.00	-0.01	0.02	0.01
KNN_{HVZ2}	62,710	0.01	0.02	-0.02	0.00
KNN_{HVZ3}	62,710	0.14***	0.15***	0.04	0.04***
KNN_{HVZ4}	62,710	0.11***	0.16***	0.00	0.02

Table 4 tabulates forecast error metrics from the nearest neighbor (k-NN) model for a variety of alternative model specifications. See Table A.1 for definitions of the forecast evaluation metrics tabulated in each column. Below each forecast error metric, the table provides the difference between the error metric of the k-NN model and that of the alternate specifications. The table compares the accuracy of the k-NN model compared to other models that use a comparable-firm-based approach. KNN_{DUP1} matches on PM and ATO. KNN_{DUP2} matches on PM, ATO, and LEV. KNN_{DUP3} matches on PM, ATO, and SGrow. KNN_{DUP4} matches on PM, ATO, LEV, and SGrow. KNN_{HVZ1} matches on ACC. KNN_{HVZ2} matches on ACC and TA. KNN_{HVZ3} matches on ACC, TA, and DIV. KNN_{HVZ4} matches on ACC, TA, DIV, and LOSS. 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 models. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

Table 5
Comparison of Our k-NN Model to Extant Models (t+1)

Model	N	MAFE	MDAFE	MSE	TMSE
KNN	62,710	6.80	2.56	3.42	1.72
RW	62,710	0.48***	0.09***	0.24***	0.25***
BCG	62,710	0.52***	0.28***	0.23***	0.23***
HVZ	62,710	0.34***	0.51***	0.04**	0.04**
EP	62,710	0.35***	0.48***	0.05**	0.05***
EP-GICS	62,710	0.35***	0.51***	0.07***	0.05***
EP-LIFE	62,710	0.20***	0.39***	0.03	0.03**

Table 5 tabulates forecast error metrics from the nearest neighbor (k-NN) model and extant earnings forecasting models. See Table A.1 for definitions of the forecast evaluation metrics tabulated in each column. Below each forecast error metric, the table provides the difference between the error metric of the k-NN model and that of the extant forecasting models. 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 models. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

Table 6
Performance After Partitioning by Analyst Coverage and for Longer Forecast Horizons

Panel A: Split by analyst coverage

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error with analyst coverage					
KNN	44,108	5.36	2.03	2.19	1.10
RW	44,108	0.42***	0.09***	0.22***	0.18***
HVZ	44,108	0.39***	0.52***	0.05***	0.03***
EP-LIFE	44,108	0.23***	0.37***	0.04**	0.01
(b) t+1 forecast error without analyst coverage					
KNN	18,602	10.21	4.52	6.34	3.35
RW	18,602	0.61***	0.12	0.27**	0.36***
HVZ	18,602	0.24***	0.31***	0.04	0.04
EP-LIFE	18,602	0.13***	0.29***	0.01	0.02

Panel B: Forecast errors for different forecast horizons

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+2 forecast error					
KNN	55,548	8.73	3.72	4.20	2.58
RW	55,548	0.79***	0.14***	0.45***	0.49***
HVZ	55,548	0.24***	0.33***	0.07**	0.07***
EP-LIFE	55,548	0.12***	0.23***	0.00	0.02
(b) t+3 forecast error					
KNN	49,007	10.15	4.46	7.32	3.41
RW	49,007	0.94***	0.21***	0.54***	0.58***
HVZ	49,007	0.15***	0.19***	0.01	0.06**
EP-LIFE	49,007	0.09***	0.18***	-0.01	0.00
(c) Aggregate forecast error (t+1 + t+2 + t+3)					
KNN	48,941	21.50	9.65	24.00	14.64
RW	48,941	2.30***	0.46***	3.68***	3.76***
HVZ	48,941	0.71***	0.95***	0.57**	0.60**
EP-LIFE	48,941	0.38***	0.70***	0.15	0.20

Table 6 tabulates forecast error metrics from the nearest neighbor (k-NN) model within subsamples. See Table A1 for definitions of the forecast evaluation metrics tabulated in each column. Below each forecast error metric, the table provides the difference between the error metric of the k-NN model compared with that from the random walk (RW) and regression-based (HVZ, EP-LIFE) models. Panel A examines forecast errors for firm-years with vs. without analyst coverage. Panel B tabulates forecast error metrics from the nearest neighbor matching (k-NN) model for various forecast horizons. Below each forecast error metric, the table provides the difference between the error metric of the k-NN model compared with that from the random walk (RW) and regression-based (HVZ, EP-LIFE) models. 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 models. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

Table 7
Determinants of Absolute and Relative Forecast Accuracy

Panel A: Summary Statistics

Variable:	N	Mean	StD	P05	P25	Med	P75	P95
(a) Dependent variables								
|kNNFE|	54,440	6.51	16.13	0.14	0.84	2.41	6.48	24.90
|kNNFE| - |RWFE|	54,440	-0.51	4.98	-5.51	-0.80	-0.15	0.62	2.67
|kNNFE| - |LIFEFE|	54,440	-0.20	3.02	-4.38	-1.26	-0.17	0.85	3.95
|kNNFE| - |HVZFE|	54,440	-0.34	2.89	-4.47	-1.68	-0.28	0.89	3.90
(b) Explanatory variables								
AbsFEG	54,440	0.02	0.04	0.00	0.00	0.01	0.01	0.10
IQR	54,440	0.09	0.10	0.01	0.03	0.05	0.11	0.29
FOLLOW	54,440	0.74	0.44	0.00	0.00	1.00	1.00	1.00
BP	54,440	0.58	0.49	0.09	0.26	0.45	0.74	1.50
LnMVE	54,440	6.03	2.00	2.94	4.49	5.94	7.36	9.58
SPI	54,440	0.63	0.48	0.00	0.00	1.00	1.00	1.00
R&D	54,440	0.06	0.11	0.00	0.00	0.01	0.07	0.28
LOSS	54,440	0.28	0.45	0.00	0.00	0.00	1.00	1.00
AbsACC	54,440	0.09	0.14	0.00	0.02	0.04	0.10	0.35
RetVol	54,440	0.15	0.09	0.05	0.09	0.12	0.18	0.33
AbsLEG	54,440	0.06	0.11	0.00	0.01	0.02	0.06	0.27

Panel B: Regression results

Dep. Variable:	|kNNFE|	|kNNFE| - |RWFE|	|kNNFE| - |LIFEFE|	|kNNFE| - |HVZFE|
Intercept	-0.00	0.00	-0.00	-0.00
	[-0.00]	[0.00]	[-0.00]	[-0.00]
AbsFEG	0.04 ***	-0.43 ***	-0.14 ***	-0.16 ***
	[2.76]	[-10.17]	[-5.08]	[-4.51]
IQR	0.12 ***	0.06	0.15 ***	0.16 ***
	[7.12]	[1.39]	[2.70]	[2.60]
FOLLOW	-0.01 ***	-0.01	-0.01	-0.02 ***
	[-2.65]	[-1.18]	[-1.46]	[-3.13]
BP	0.13 ***	0.05 ***	0.00	0.01
	[6.49]	[3.38]	[0.20]	[1.05]
LnMVE	-0.02	0.00	-0.00	0.00
	[-1.04]	[0.03]	[-0.52]	[-0.48]
SPI	0.00	-0.02 ***	0.02 ***	0.03 ***
	[0.80]	[-3.79]	[4.65]	[5.42]
R&D	0.01	0.02 **	-0.02	-0.00
	[1.03]	[2.40]	[-1.43]	[-0.10]
LOSS	0.03 ***	0.04 ***	-0.06 ***	-0.09 ***
	[4.96]	[3.44]	[-6.44]	[-8.40]
AbsACC	0.17 ***	0.01	-0.01	0.01
	[10.33]	[0.70]	[-0.57]	[1.01]
RetVol	0.05 ***	0.00	0.04 ***	0.04 ***
	[4.53]	[0.33]	[4.54]	[4.57]
AbsLEG	0.08 ***	-0.05	-0.01	-0.02
	[3.50]	[-1.31]	[-0.20]	[-0.43]
Adj. R^2	0.21	0.15	0.01	0.02
N	54,440	54,440	54,440	54,440

Table 7 reports summary statistics and coefficients of regressions of absolute forecast errors of the k-NN model and of absolute forecast error differences of the k-NN model and the benchmark models on factors associated with forecast errors. All continuous explanatory variables are winsorized at the 1st and 99th percentile. Panel A provides summary statistics. Panel B provides the regression results. For Panel B, all variables, dependent and independent variables, are standardized (subtracting the mean of the variable and dividing the difference by the standard deviation of the variable). The number of observations differs from 62,170 due to requiring return volatility data. T-values are reported in parentheses and are based on standard errors clustered by firm and by year. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

Table 8
Return Predictability Tests

	1	2
1 FEG	0.17*	0.66***
	[1.68]	[3.17]
2 IQR		0.26
		[0.85]
3 $FEG * IQR$		-0.79**
		[-2.48]
4 BP	0.17	0.19
	[0.53]	[0.56]
5 Profitability	0.39	0.26
	[0.95]	[0.73]
6 Investment	-0.81***	-0.82***
	[-4.67]	[-4.58]
7 MVE	-0.20	-0.24
	[-0.61]	[-0.73]
8 Intercept	0.45	0.35
	[0.78]	[0.73]
Effect of 1 + 3		-0.13
		[-0.79]
Adj. R^2	0.02	0.03
Average N per cross-sect.	2,730	2,730
No. of cross-sect. coeff.	252	252

Table 8 reports coefficients and t-values from Fama-MacBeth regressions to predict monthly market-adjusted returns. The coefficients (r-squareds) are the time-series averages generated by 252 monthly cross-sectional regressions. Each t-statistic equals the average coefficient divided by its time-series standard error. All explanatory variables are decile ranks with values from 0 to 1 in increments of 1/9 which are formed in June of each year. Returns span from July until June of the next year. Market-adjusted returns are raw returns minus the value-weighted market return. Firms that delist within the 12-month holding period from July to June are invested in the value-weighted market return until the end of the holding period. T-values are reported in parentheses. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.