Forecasting Earnings Using k-Nearest Neighbor Matching

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

Martin M. Kapons
Tilburg University[††]
M.M.Kapons@uvt.nl

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

Harm H. Schütt
Tilburg University[1]
H.H.Schutt@uvt.nl

Eric H. Weisbrod
University of Kansas[2]
eric.weisbrod@ku.edu

This Draft: November 30, 2020

We thank Jeff Coulton, Michael Ettredge, Andrew Jackson, Stephannie Larocque, Richard Morris, Arthur Stenzel, Stephen Taylor, Andrew Yim, David Veenman, and workshop participants at the 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.
[††] Tilburg School of Economics and Management, Tilburg University, P.O. Box 90153 5000 LE Tilburg, Netherlands.
[‡] David Eccles School of Business, University of Utah, Salt Lake City, Utah, 84112, USA.
[1] Tilburg School of Economics and Management, Tilburg University, P.O. Box 90153 5000 LE Tilburg, Netherlands.
[2] University of Kansas, School of Business, Lawrence, KS, 66045, USA.

Forecasting Earnings Using k-Nearest Neighbor Classification

ABSTRACT

We use the k-nearest neighbors (i.e., k-NN) algorithm to forecast a firm's annual earnings by matching its recent trend in annual earnings to historical earnings sequences of "neighbor" firms. Our forecasts are more accurate than forecasts obtained from the random walk, the regression model developed by Hou, van Dijk and Zhang (2012), other regression models and the matching approach described in Blouin, Core and Guay (2010). The k-NN model is superior to these alternative models both when analysts' forecasts are available and when they are not. Further, for firm-years with I/B/E/S earnings data available, the accuracy of k-NN forecasts of I/B/E/S earnings is similar to the accuracy of analysts' forecasts. The k-NN model is also superior to a random forest classifier that we use to choose the best model ex-ante. Finally, we find that our forecasts of earnings changes have a positive association with 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.

1 INTRODUCTION

We examine the accuracy of earnings forecasts formed by k-nearest neighbor classification (i.e., k-NN). The practice of finding similar episodes, or "k-nearest neighbors," on which to base predictions is a simple yet effective forecasting approach that 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). We develop a method for forecasting a subject firm's future earnings by matching its recent trend in annual earnings to similar historical earnings sequences of other firms.

We focus on expected future earnings because they are a key determinant of investment decisions. And, expected year-ahead (or lead) earnings play a central role. For example, the abnormal earnings growth valuation model of Ohlson and Juettner-Nauroth (2005) uses capitalized expected lead earnings as its anchor. In a similar vein, practitioners often evaluate a firm's current equity price by comparing it to the firm's expected lead earnings (i.e., the price-to-forward-earnings ratio). Thus, understanding the link between past and future earnings and using this understanding to develop accurate earnings forecasts are important endeavors for researchers and practitioners alike.

Nevertheless, the forecasting methods commonly used by practitioners are quite different from those used by researchers in accounting and finance. Most financial statement analysis courses and valuation textbooks teach practitioners to forecast a subject firm's earnings by extrapolating from trends observed among comparable firms. Textbooks generally suggest using trends in key financial ratios of comparable firms as inputs in the process of forecasting the earnings of the subject firm. A benefit of this tailored forecasting approach is that it weighs the historical

performance of both the subject firm and potential peer firms in a non-parametric manner. However, it is challenging to apply this approach in a large-scale empirical study because it relies on subjective decisions such as selecting the appropriate peer firms and choosing the relevant financial ratios for each subject firm.

Rather than attempt to develop their own large samples of peer-based forecasts, researchers have generally relied on earnings forecasts obtained from sell-side financial analysts, regression models, or some combination of the two. Although a vast literature in accounting and finance uses earnings forecasts of sell-side analysts, recent studies increasingly rely on cross-sectional regressions to forecast annual earnings (Hou, van Dijk and Zhang, 2012; Gerakos and Gramacy, 2013; Li and Mohanram, 2014; So, 2013). One explanation for the growing popularity of regression-based forecasts is that they are available for many more firms than analysts' forecasts. However, regression-based approaches are limited in their ability to differentially weigh the historical performance of more versus less relevant peers. This makes them susceptible to the influence of extreme observations (i.e., outliers). Accordingly, regression-based forecasts are not much better than forecasts based on the random walk (e.g., Gerakos and Gramacy, 2013; Li and Mohanram, 2014; Monahan, 2018; Easton, Kapons, Kelly and Neuhierl, 2020).

We examine whether modern k-NN methods can be used to develop large-sample non-parametric earnings forecasts that overcome some of the limitations of the common forecasting approaches used by researchers. Motivated by the intuition observed in common practice, we introduce a simple k-NN forecasting approach that forecasts a subject firm's earnings by extrapolating from trends observed among matched nearest neighbors. We show that our approach provides considerably more accurate forecasts than those from the popular Hou, van Dijk and Zhang (2012) (hereafter, HVZ) regression model and can be applied to a broader set of firm-years.

For example, analyzing a large sample of annual earnings data from 1978 to 2018 we find that the mean absolute forecast error (i.e., MAFE) from our one-year-ahead k-NN forecasts is approximately 27.1 percent smaller than that from the HVZ model, and that k-NN forecasts can be computed for approximately 18.8 percent (roughly 40,000) more firm-years.

Specifically, we examine three research questions. First, what is the best way to use k-NN prediction methods to forecast earnings? Second, how accurate are k-NN forecasts compared to those from competing approaches? Third, when are k-NN forecasts more likely to outperform other approaches?

With regards to the first question, we introduce the idea that comparable firm-years need not be contemporaneous; so, for any given year t, we search for matches from the previous ten years of historical data. When searching for nearest neighbors, we examine how the accuracy of the resulting forecasts varies with three parameters: (1) the number, F, and type of financial variables (i.e., features) that we use to find matches; (2) the length of time, M, years, that we use to match the trend in each variable (we vary M from one to five); and, (3) the number of nearest neighbors, k, we select as matches. The k nearest neighbors are then the firm-years in the prior ten-years with the most similar histories of feature values vis-à-vis the recent history of the subject firm.[1]

For example, suppose in 2011 we want to forecast the 2012 earnings of subject firm i. Further suppose that the only variable in our feature set is earnings (i.e., $F = 1$) and that we are matching on sequences with a length of two years (i.e., $M = 2$). We begin by identifying the sample of all available firm-years in our dataset with both current and lagged earnings data reported during 2001

[1] We also find that our k-NN forecasts have smaller forecast errors than forecasts obtained from the random walk, other regression-based models and the matching model developed by Blouin, Core and Guay (2010). In addition, our forecast errors are similar to the forecast errors documented in Cao and You (2020). However, our machine learning model is much simpler than the machine learning models that they evaluate. We use simple k-NN models and consider only a small set of features. Cao and You (2020), on the other hand, evaluate several complex models and consider 56 different features.

to 2010.[2] Next, we identify the k nearest neighbors – i.e., the k firm-years in this sample that are the closest match to firm i in terms of current and lagged earnings. Finally, we set our forecast of firm $i's$ earnings in 2012 equal to the the median of the k nearest neighbors' lead earnings.[3]

The definition of a comparable firm-year naturally depends on the length of the sequence (i.e., M) and the set of features. We explore various features that may be considered standard, including matches based on industry, size and past accruals. We also vary the number of years (i.e., M) and the number of nearest neighbors (i.e., k). Our, perhaps surprising, conclusion is that the optimal number of years (based on minimum mean absolute forecast error, MAFE) is two and the optimal number of nearest neighbors is approximately eighty. We also find that a simple k-NN model that uses only one feature – i.e., earnings scaled by equity market value or "scaled" earnings – works well.[4] Adding more features does not improve accuracy. These results imply that: (1) a firm's earnings history is highly informative when put into the context of similar histories and (2) a high number of nearest neighbors is necessary to obtain a sufficiently precise forecast of the future trajectory of earnings.[5]

Based on these findings, in our analyses of the second question, "How accurate are k-NN forecasts compared to those from competing approaches?", we consider k-NN forecasts using $K = 80$ nearest neighbors matched on $M = 2$ years of scaled earnings. We compare our k-NN forecasts to those obtained from the random walk and the regression model proposed by Hou, van Dijk and Zhang (2012). We evaluate forecasts of one-, two-, and three-year-ahead earnings. We

[2] The two-year sequence of firm $i's$ 2010 and 2011 earnings can be matched to any two-year "neighbor" sequence as early as 2000 – 2001 and as recently as 2009 – 2010. The two-year "neighbor" sequence can be taken from the reported earnings of any available firm, including firm i itself.

[3] We elaborate on this example in Figure 1 and Section 2.2.1.

[4] As discussed in Section 4.1.3, we obtain similar results when we use alternative deflators such as equity book value, total assets or revenues.

[5] This does not mean that other financial ratios are uninformative about future earnings, only that their information content is outweighed by the additional error introduced by having to find similar firms across many dimensions.

find that, based on many summary metrics for forecast accuracy, our k-NN forecasts are significantly more accurate than the HVZ forecasts and the random walk forecasts.

Our k-NN forecasting procedure is flexible. It can be applied to a broad cross-section of firms and to alternative earnings metrics. Although our primary analyses are based on earnings before special items, which is a commonly-used metric, we also evaluate I/B/E/S forecasts of "street" earnings. Specifically, for the sub-sample of observations for which I/B/E/S forecasts of street earnings are available, we apply our k-NN forecasting approach to I/B/E/S street earnings, and then we compare our forecasts to analysts' consensus forecasts. We find that analysts' consensus forecasts are significantly more accurate than random walk forecasts of street earnings. However, the accuracy of our k-NN forecasts of street earnings is similar to the accuracy of analysts' consensus forecasts in the sense that, for a number of accuracy metrics, it is not statistically different. Moreover, the difference between the accuracy of analysts' forecasts and the accuracy of our k-NN forecasts is less for forecasts of two-year-ahead earnings than for forecasts of one-year-ahead earnings.

We address our last research question, "When are k-NN forecasts more likely to outperform other approaches?", in two ways. First, we show that k-NN forecasts are more accurate than HVZ or random walk forecasts for smaller firms, firms without analyst coverage, firms with positive growth, and firms with either high or low current earnings. k-NN forecasts are slightly less accurate than random walk forecasts for firms with negative earnings growth. We also compare forecast errors by industry and find that our k-NN forecasts are consistently more accurate than HVZ or random walk forecasts.

Second, we construct hedge portfolios to test whether our k-NN forecasts are associated with future stock returns. On June 30 of each year, we separate firms into two portfolios: Firms for

which the sign of the change in earnings implied by the forecasting model (i.e., k-NN or HVZ) is positive and firms for which the sign is negative. We compute hedge-portfolio returns by subtracting the average monthly returns generated by firms in the negative earnings-change portfolio from the returns generated by the firms in the positive earnings-change portfolio, and then we compute average monthly hedge-portfolio returns for each of the subsequent twelve months. We find that the k-NN hedge portfolio generates positive returns whereas the returns generated by the HVZ hedge portfolio are negative. We also find that the k-NN hedge-portfolio generates higher returns for firms that are not covered by analysts. These results lead us to conclude that, in addition to being more accurate than other forecasts, our k-NN forecasts are also useful in the sense that they are informative about future stock price changes.

We make three contributions to the literature on earnings properties and earnings forecasting. First, we introduce a new forecasting model that: (1) is simple, flexible and easy to implement; (2) can be applied to a much larger sample of firms than regression-based models; (3) outperforms competing approaches over both short and long forecasting horizons; and, (4) generates earnings forecasts that are associated with future stock returns. Second, because our method is easily modified, we open up several avenues for future research. For example, future research might evaluate whether the distribution of the forecasts of the k nearest neighbors that are matched to a subject firm serves as a useful measure of the degree of uncertainty about that firm's earnings; or, whether the k-NN approach we develop to forecast a firm's earnings can be used to forecast other financial metrics such as sales, cash flow or accruals. Finally, our results offer new insights into the link between future earnings and historical earnings. The simple k-NN model that only matches on the most recent two years of earnings works best. Adding more features to identify "better"

matches does not lead to better forecasts. This result implies that, when put into the correct context, a firm's recent earnings history is highly informative about what its future earnings will be.

2 APPROACHES TO EARNINGS FORECASTING

2.1 k-Nearest Neighbors

2.1.1 IBM Example

Before formally describing our k-NN model, we provide an example, which is diagramed in Figure 1. In this example, we make a forecast in 2011 of IBM's earnings before special items, $EBSI$, for 2012. We refer to $EBSI$ as *unscaled* earnings. We match on only one feature (i.e., $F = 1$): the ratio of $EBSI$ to end-of-year t equity market value, which we refer to as *scaled* earnings, $SEBSI$. We set $M = 5$ and $k = 10$. Hence, we compare IBM's $SEBSI$ for the years 2007 to 2011 to all the observable five-year sequences of $SEBSI$ that end on *any* year $s \in [2001,2010]$, and then we identify the ten "neighbor" sequences that are closest to IBM's $SEBSI$ for the years 2007 to 2011. As shown in Figure 1, in this example, IBMs' ten nearest neighbors are drawn from as early as $s = 2001$ (Irwin Financial Corp.) to as late as $s = 2010$ (Envision Healthcare Corp.).[6]

Next, using each of the ten nearest neighbors, we forecast IBM's unscaled earnings (i.e., $EBSI$) for year $t + 1$. We do this in two steps. In the first step, we form a set of intermediate forecasts by multiplying each of the ten nearest neighbor's scaled earnings for year $s + 1$ by *IBM's* equity market value at the end of 2011. Hence, in this example, we multiple Irwin Financial Corp.'s observed $SEBSI$ for 2002 (i.e., $s + 1$) and the year $s + 1$ $SEBSI$ of the other nearest neighbors by 180,221 million USD, which equals IBM's equity market value at the end of 2011. As shown in

[6] In Panel A of Figure 1 we plot, in event time, the scaled earnings of IBM along with those of its matched neighbors. In Panel B we show the different calendar time periods that relate to each of the ten nearest neighbors.

Figure 1, $SEBSI$ in year $s+1$ for the ten nearest neighbors ranges from 0.062 to 0.110, and thus the intermediate forecasts range from 11,174 million USD to 19,824 million USD. Finally, in the second step, we set our k-NN forecast of IBM's earnings for 2012 equal to the median of the ten intermediate forecasts, which, in this example, is 14,478 million USD.

2.1.2 Formal Description of the k-NN Model

For each firm-year i,t, we use k-NN to compute an h-year-ahead forecast of earnings before special items. We refer to firm i's realized unscaled (scaled) earnings before special items for year t as $EBSI_{i,t}$ ($SEBSI_{i,t}$) and we refer to our k-NN-based forecast of $EBSI_{i,t+h}$ as $FEBSI_{i,t+h}^{kNN}$. First, we determine the most recent history of features for firm-year i,t. We refer to this history as sequence $i,t\colon i,t-M+1$ (M denotes the length of the sequence in years). Second, we identify the set of firm-years that have complete sequences of features of length M ending in *any* year $s \in [t-h, t-9-h]$. These are the neighbors of firm-year i,t. [7]

Third, for each firm-year j,s in the set of neighbors, we calculate the variable $DIST_{i,t,j,s}^{F,M}$, which is the Euclidean distance between sequence $i,t\colon i,t-M+1$ and sequence $j,s\colon j,s-M+1$.

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

In equation [1], $FEAT_{i,t-m+1}^{f}$ ($FEAT_{j,s-m+1}^{f}$) is the *normalized* value in year $t-m+1$ ($s-m+1$) of feature f for subject firm i (neighbor firm j).[8] To illustrate, suppose we match on

[7] We match on scaled features. We scale all the dollar amounts in the subject firm's sequence of features by the year t value of the deflator for the subject firm. Similarly, we scale all the dollar amounts of a neighbor's sequence of features by the year s value of the deflator for that neighbor. As discussed in Section 4.1.3, our results do not depend on the choice of deflator.

[8] We normalize each feature by subtracting the contemporaneous cross-sectional average from the raw value of the feature, and then dividing this difference by the contemporaneous cross-sectional standard deviation. (This is common way of implementing k-nearest neighbors.) The resulting normalized feature has a mean of zero and a standard deviation of one. Consequently, all the features have the same scale, and thus $DIST_{i,t,j,s}^{F,M}$ is not dominated by a feature with an outsized scale.

two features – e.g., $SEBSI$ and scaled accruals – then $FEAT^{1}_{i,t-m+1}$ and $FEAT^{2}_{i,t-m+1}$ ($FEAT^{1}_{j,s-m+1}$ and $FEAT^{2}_{j,s-m+1}$) are the normalized values of firm i's (j's) $SEBSI$ and scaled accruals in year $t-m+1$ ($s-m+1$), respectively.

Fourth, we identify the k neighbors with the smallest values of $DIST^{F,M}_{i,t,j,s}$. These are the *k-nearest* neighbors of firm-year i,t. Finally, we form k intermediate forecasts of firm $i's$ $EBSI$ for year $t+h$ by multiplying each neighbor's $SEBSI$ for year $s+h$ by the value of the deflator for *subject firm* i in year t. We then set $FEBSI^{kNN}_{i,t+h}$ equal to the median value of the k intermediate forecasts. In most of our analyses, we choose 80 nearest neighbors (i.e., $k=80$) by matching on two years (i.e., $M=2$) of scaled earnings (i.e., $F=1$) and we define $SEBSI$ as earnings before special items scaled by equity market value. We provide reasons for these choices later in the paper.

2.2 Description of Regression-Based Forecasts

We compare our k-NN forecasts to random walk forecasts, regression-based forecasts, forecasts obtained from the matching approach described in Blouin, Core and Guay (2010) (BCG hereafter) and analyst forecasts.

For the sake of brevity, in the main tables, we focus on only one regression model: The model proposed by HVZ. This model is widely adopted and is often referred to as the benchmark model for regression-based earnings forecasts (e.g., Evans, Njoroge and Yong, 2017; Li and Mohanram, 2014; So, 2013). In Appendix B, we compare our k-NN forecasts to forecasts obtained from several other regression models. The inferences we obtain for the HVZ model hold for the other models.

Forecasts from the HVZ model are obtained using the estimated coefficients from the regression shown below:

$$SEBSI_{i,t+h} = \alpha_0 + \alpha_1 \times TA_{i,t} + \alpha_2 \times DD_{i,t} + \alpha_3 \times DIV_{i,t} + \alpha_4 \times SEBSI_{i,t} + \alpha_5 \times LOSS_{i,t} + \alpha_6 \times ACC_{i,t} + \epsilon_{i,t}. \quad [2]$$

In equation [2], $SEBSI_{i,t+h}$ 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; and, $ACC_{i,t}$ denotes firm $i's$ scaled accruals for year t. (When calculating the numerator of $ACC_{i,t}$, we use the same definition of accruals as HVZ.) With the exception of the indicator variables $DD_{i,t}$ and $LOSS_{i,t}$, the variables in equation [2] are scaled by firm $i's$ equity market value at the end of year t. We elaborate on how we compute all of our variables in Section 3.2 and Table A.1.

We estimate median regressions not ordinary least squares (i.e., OLS) regressions. We do this because median regressions are less sensitive to extreme observations. And, as shown in Evans, Njoroge and Yong (2017); Easton, Kapons, Kelly and Neuhierl (2020); and, Tian, Yim and Newton (2020), forecasts based on median regressions are significantly more accurate than forecasts based on OLS regressions.[9]

2.3 Rolling Window Forecasting Procedure

When developing our k-NN forecasts and our regression-based forecasts, we apply the rolling window forecasting procedure that is illustrated in Figure 2. For each year t in our sample, we identify nearest neighbors for our k-NN model and estimate coefficients for the regression model based on the last $t-h$ to $t-9-h$ years of panel data. In the parlance of machine learning, this

[9] As shown in Appendix B, inferences remain unchanged if we estimate the HVZ model using OLS.

is our training data. To eliminate look-ahead bias, we: (1) define the cross-section of data for year t as the firm-years that ended their fiscal year between April first of year $t-1$ and March 31 of year t and (2) we identify nearest neighbors and estimate regressions coefficients using the panel of data that begins on year $t-9-h$ and ends on year $t-h$. For example, to develop our 1998 regression forecast of firm $i's$ $EBSI$ for 1999, we use a panel of data in which the dependent (independent) variables are drawn from the years 1990 through 1998 (1989 through 1997). We multiply the estimated coefficients from this panel regression by the values of the predictors for firm i in 1998, and then, to obtain our unscaled (or dollar) forecast of firm $i's$ $EBSI$ in 1999, we multiply this predicted value by firm $i's$ equity market value at the end of 1998. We refer to the HVZ forecast of $EBSI_{i,t+h}$ as $FEBSI_{i,t+h}^{HVZ}$.

3 SAMPLE CONSTRUCTION AND VARIABLE DEFINITIONS

3.1 Sample Construction

We obtain company data from the Compustat Fundamentals Annual file. We delete firms that are not incorporated in the U.S. Our initial sample spans the years 1968 through 2019 inclusive. We evaluate the accuracy of our forecasts of $EBSI$ for year $t+h$ by comparing it to realized earnings for year $t+h$. Therefore, our largest forecasting sample spans the years 1979 to 2018. To minimize the effect of database errors and small deflators, we require all firm-years to report positive values of: (1) total assets; (2) equity book value; (3) equity market value; and, (4) sales. For observations included in the forecast comparison sample, we also require equity market value to be greater than ten million U.S. dollars.

We describe our sample construction in Panel A of Table 1. Missing data for earnings before special items, $EBSI$, and non-positive deflators reduces the available Compustat sample from 339,171 to 213,071, which is the number of observations for which we can form a random walk

forecast – i.e. the random walk sample. Only 5,315 of the firm-years in the random walk sample have missing lagged values of $EBSI$, which is the minimum requirement for the $M = 2$ k-NN forecast model. Hence, the k-NN sample overlaps with 97.5 percent of the random walk sample. After removing observations with missing accruals, missing realizations of future $EBSI$, which is required for assessing forecast accuracy, and with equity market value less than \$10 million, we have 132,039 firm-year observations for which we can compare one-year-ahead realizations to one-year-ahead forecasts generated by the random walk, our k-NN model and the HVZ model.

The data described above and the amount of overlap between the random walk sample, the k-NN sample, the HVZ sample and the sample of analysts' forecasts are summarized in Figure 3 as a Venn diagram. The k-NN sample covers 97.5 percent of the random walk sample. However, the HVZ sample overlaps with only 78.9 percent of the random walk sample, and analyst forecasts are available for only 53.8 percent of the random walk sample.

In later analyses we evaluate relative forecast accuracy for each decade within our sample. Hence, in Panel B of Table 1, we describe the data summarized in Figure 3 by decade.

In later analyses, we compare forecasts of I/B/E/S street earnings obtained from our k-NN model and the random walk model with analysts' consensus forecasts. To create our k-NN and random walk forecasts of street earnings, observations in the street earnings sample must have non-missing lagged street earnings; and, they must also have a contemporaneous consensus analyst forecast available for comparison. We obtain actual raw street earnings from the unadjusted eps actual file and consensus eps forecasts from the unadjusted summary file. We align the I/B/E/S data in calendar time to mimic that of the Compustat forecasts.[10] We merge the I/B/E/S data to our

[10] For example, for the fiscal year ended December 31, 2001, we obtain the $t + 1$ (I/B/E/S $fpi = 1$) consensus forecast of 2002 annual eps as of March 2002. The I/B/E/S historical consensus is recorded once a month on the third Thursday of each month. Hence, if no $t + 1$ consensus is available as of the third month after the fiscal year end, we use the consensus for the following month.

Compustat sample using the Compustat security file; and, we require all observations to have adjustment factors available in the I/B/E/S adjustment file. We scale I/B/E/S eps forecasts by the end-of-fiscal-year closing stock price from Compustat, adjusted for stock splits and dividends. We require that observations included in the street earnings sample have a year t adjusted closing stock price greater than \$1. These data requirements result in sample sizes of 96,345 and 74,679 observations in the $t+1$ and $t+2$ street earnings samples, respectively.

3.2 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. To keep the deflator constant throughout the analysis, we scale all firm i (j) variables used in the forecast models by the value of the deflator at period t (s). In most of our analyses, we scale by equity market value. Firm $i's$ equity market value at the end of fiscal year t, $MVE_{i,t}$, equals the product of Compustat data items $prcc\_f_{i,t}$ and $csho_{i,t}$ – i.e., $MVE_{i,t} = prcc\_f_{i,t} \times csho_{i,t}$. For the comparison with the HVZ 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] 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 Table A.1 we provide a complete list of all variables 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.

3.3 Descriptive Statistics

In Panel A of Table 2, we provide descriptive statistics for the predictors we use in our regression models, which are also the candidate features that we use to identify nearest neighbors. (As discussed later, we ultimately use only one feature, scaled earnings, $SEBSI$.) The medians of these variables are similar to the amounts shown in other studies, but the means, standard deviations and tails of the distribution are different. The reason for this is that for the analyses in these tables, we neither winsorize nor delete extreme values.

In Panel B of Table 2, we summarize the estimates of the regression coefficients for the HVZ model. The coefficients (pseudo r-squared) are the time-series averages of the estimated coefficients (pseudo r-squared) generated by the rolling-window median regressions. t-statistics are derived from Fama-MacBeth standard errors. The estimate of the coefficient on lagged earnings is highly significant (coefficient estimate of 0.709 with a t-statistic of 9.654). The other coefficient estimates that are statistically significantly different from zero are those on total accruals (-0.024 with a t-statistic of -4.78), the dividend variable (1.052 with a t-statistic of 4.491) and the indicator for non-zero dividend payments (-0.01 with a t-statistic of -2.234).[12] The estimates of the coefficients on total assets and the loss indicator are not significantly different from zero.

In Panel C of Table 2, we provide some initial descriptive statistics comparing the subject firms with the firms that comprise the nearest neighbors (k-NN using $SEBSI$ as the only feature and setting $M = 2$ and $k = 80$). $SEBSI$ of the nearest neighbors are very similar, which is not surprising given $SEBSI$ is the variable we use to find matches. For instance, the interquartile range

[12] When comparing our coefficient estimates to those in HVZ, it is important to note that we use median regressions to reduce the undue influence of a small proportion of the observations while HVZ use OLS regressions. We also scale each of our variables by market capitalization (and we check the sensitivity of our results to different deflators). HVZ, on the other hand, report results for regressions based on unscaled variables.

of the difference between $SEBSI$ of subject firms and $SEBSI$ of the k-nearest neighbors is 0.001 and the median difference is 0.000. On the other hand, the difference in size measured as either equity market value or total assets is large. The median percentage of nearest neighbors in the same Fama-French 12-industry group as the subject firm is 12.5 percent and the median percentage of nearest neighbors with the same 2-digit SIC as the subject firm is 3.8 percent. These statistics highlight that the nearest neighbors are indeed very similar in terms of the matched features, but otherwise quite heterogeneous.

4 RESULTS AND MODEL EVALUATION

4.1 Implementing the k-NN Model

4.1.1 The Role of Matching and Extrapolating

We begin by answering our first research question: "What is the best way to use k-NN prediction methods to forecast earnings?" We first ask some foundational questions. Does nearest neighbor matching yield lower forecast errors than random matching? The answer is yes. And: How much of the reduction in the forecast error is attributable to finding nearest neighbors with matched earnings sequences and how much is attributable to applying the growth rate implied by the nearest neighbors' earnings from year s to year $s + h$ to the year t earnings of the subject firm? We benchmark these analyses against the $t + 1$ forecast errors from the random walk. The results are shown in Figure 4.

The MAFE for the random walk model is shown as the green horizontal line in Figure 4, which plots the MAFE (shown on the vertical axis) against the number of nearest neighbors (shown on the horizontal axis). Of course, the line for the random walk runs parallel to the horizontal axis (and the MAFE is 7.82 percent) because the forecast is based on the subject firm's year t earnings, and thus does not depend on k. The yellow line with squares is the MAFE we obtain when we

randomly choose k neighbors – i.e., k *random* neighbors instead of k *nearest* neighbors – and $k \in [10, 20, \cdots, 200]$. The MAFE, which is 10.36 percent with ten random neighbors, declines rapidly as k increases but the incremental accuracy begins to taper off at $k = 100$ and the MAFE converges to 9.83 percent. Clearly, the forecast errors based on k random neighbors are greater than the errors when forecasts are based on the random walk. The blue and purple lines plot the MAFEs for values of $k \in [10, 20, \cdots, 200]$ *nearest* neighbors. The blue line with triangles plots the MAFEs we obtain when we assume that the earnings of the nearest neighbors follow a random walk – i.e., we assume zero growth and use the median of the nearest neighbors' earnings in year s as the forecast. The black line with circles plots the MAFEs that we obtain when we set the forecast equal to the median of the nearest neighbors' realized earnings in year $s + 1$. Each of these lines is far below the plot of the MAFEs obtained using k random neighbors, demonstrating the benefit of nearest neighbor matching over random selection. They are also well below the MAFE of the random walk. Moreover, the purple line is clearly the closest to zero and noticeably lower than the blue line. This demonstrates the importance of using the growth rate implied by the earnings trends of the nearest neighbors. For example, the MAFE based on the median of year t earnings of 100 nearest neighbors is 7.58 percent while the MAFE based on the median of year $t + 1$ earnings is 7.05 percent. Thus, it is the combination of both the nearest neighbor matching and growth extrapolation that matters.

4.1.2 Choosing the Optimal Matching Model

Three considerations, which potentially make a sizeable difference in the effectiveness of the k-NN model, are: (1) the length of the sequence of the subject-firm's data that is matched to sequences of other firms (i.e., M); (2) the number of firms that are included in the set of nearest neighbors (i.e., k); and, (3) the number and type of features on which we match (i.e., F). We

investigate the effects of changing these parameters in this section. Based on these analyses, we choose values of M, k and F that we use in our subsequent analyses. The choice is based on comparisons of mean absolute forecast errors (MAFE). That said, this choice is not sensitive to the error metric used – i.e., using different error metrics lead to the same general inferences.

We summarize the effects of varying M and k in Table 3 and Figure 5. In Table 3, we tabulate the mean absolute one-year-ahead forecast error (MAFE) from the k-NN model for various combinations of M and k. The analysis is performed on a constant sample of observations with data available for all values of $M = 1$ to $M = 5$. Each column examines the effect of increasing k by increments of ten for a given value of M. Beside each tabulated value of MAFE, we tabulate the difference in that MAFE relative to the MAFE using ten fewer peers. For each value of M, we stop tabulating differences in MAFE at the value of k, which we refer to as k^*, for which adding ten additional matches does not lead to a statistically significant reduction in the MAFE at the five percent level.

Figure 5 plots the values tabulated in Figure 3. Two results are apparent. First, the MAFE is lowest for every value of k when $M = 2$. Second, when $M = 2$, the last 10-neighbor increment in k that leads to a statistically significant decrease in the MAFE is the increment to a value of $k^* = 80$. Based on these two results, in our subsequent analyses, we choose 80 nearest neighbors by matching on sequences of length two – i.e., $k = 80$ and $M = 2$.

As an interesting aside, the plots for $M = 3, 4$ and 5 show that increasing k does not necessarily increase accuracy. For example, when $M = 4$, accuracy peaks at $k^* = 30$, and then decreases as k increases. This is attributable to the phenomenon known in statistics as the bias-variance trade-off. Within the context of our study, the intuition for this trade-off is as follows. When k is small, k-NN exhibits low bias because the nearest neighbors are very similar. However, the median

forecast obtained from a small set of neighbors is very noisy because it is sensitive to the idiosyncrasies of each neighbor. As we increase k, we include more dissimilar neighbors, which tend to be less representative of the subject firm, and thus the bias increases. However, our estimate of the median is less affected by idiosyncratic noise, and thus the variance of the forecast error decreases. The fact that the highest accuracy is generally achieved with large k suggests that, when forecasting earnings, reducing variance – i.e., achieving a more precise estimate of the typical earnings trajectory – is more important than reducing bias by choosing a small but very similar set of neighbors.

4.1.3 Alternative Specifications of the k-NN Model

In the previous analyses, we evaluated forecasts obtained from nearest neighbors that were identified by matching on only one feature: earnings before special items scaled by equity market value (i.e., $SEBSI$). In this section, we evaluate the sensitivity of our results to the use of: (1) earnings before special items that are scaled by other variables (equity book value, total assets and sales) and (2) additional features. We also examine forecasts obtained by using matches within the same size decile or within the same industry.

We show the results of these analyses in Table 4. We consider four measures of forecast accuracy, which weigh large errors differently: (1) the mean of the absolute value of the scaled forecast errors, MAFE; (2) the median of the absolute value of the scaled forecast errors, MDAFE; (3) the mean of the squared scaled forecast errors, MSE; and, (4) the mean of the squared scaled trimmed forecast errors, TMSE. When computing TMSE, we first delete the top and bottom 0.1 percent of the forecast errors, and then we compute the mean squared error.[13] We evaluate one-year-ahead, two-year-ahead, and three-year-ahead forecasts. And, in all cases, forecast errors are

[13] Inferences are identical when we winsorize, rather than trim, the top and bottom 0.1 percent of the signed forecast error distribution.

computed as $\left(EBSI_{i,t+h} - FEBSI_{i,t+h}\right)/MVE_{i,t}$ – i.e., we always scale the forecast *errors* by equity market value.

In Panel A, we summarize the forecast errors we obtain when we scale the features that we use to identify nearest neighbors by three alternative deflators: equity book value, total assets and sales. The MAFE, MDAFE and TMSE are significantly larger when we replace equity market value with any of these deflators. Moreover, MSE is also significantly larger when we replace equity market value with sales. For example, the MAFE for one-year-ahead earnings when the deflator is equity market value is 6.965 percent, which is 0.330, 0.452 and 0.507 percent larger than the MAFE we obtain when we scale the matching features by equity book value, total assets and sales, respectively. We also note that trimming the extreme forecast errors has a considerable effect on the mean squared error. For example, the MSE for the one-year-ahead earnings forecast that is based on nearest neighbors that are matched to the subject firm on the basis of $EBSI$ scaled by equity market value is 7.631 whereas the TMSE is 1.973.

Of course, when identifying nearest neighbors, there are several features that we could use in addition to scaled earnings. Considering the central role of earnings in valuation and given that we are forecasting earnings, the obvious choice for a single feature is earnings. Nonetheless, we examine several sets that add additional features to scaled earnings.[14] We refer to these as expanded sets. Although we only tabulate the results for two of these expanded sets, we obtain similar results for all the expanded sets that we evaluate. The first expanded set consists of scaled earnings and scaled accruals. We start by adding accruals because, after lagged earnings, the estimate of the coefficient on accruals in the HVZ model is the most significant. We refer to this

[14] Given the results discussed in the previous paragraph, we scale all feature sets by equity market value. We exhaustively examine all possible combinations of both deflators and feature sets and the un-tabulated results of these tests lead to the same inferences.

k-NN model as NN2. In the second expanded set, we add all the statistically significant variables included in the HVZ regression to the set underlying NN2.[15] We refer to this k-NN model as NN3. The results, which are shown in Panel B, demonstrate that the forecast errors increase as we expand the set of features. For example, when scaled earnings is the only feature, the MAFE for one-year-ahead earnings is 6.893 percent. When we add accruals, the MAFE is higher (the difference is 0.76 percent); and, when we add the additional HVZ features, the MAFE is higher yet (the difference is 2.32 percent). These results are based on $k = 80$ nearest neighbors; however, in un-tabulated results, we find that the ranking of the models is not sensitive to the choice of k.

In Panel C, we examine the effect of estimating the k-NN model within strata based on either the Fama and French (1997) (FF12) twelve industry classification or deciles of equity market value. When analyzing industry-based (sized-based) matching, for each year, we first group firms into FF12 industry groups (size deciles), and then, within each group, we use scaled earnings to identify the nearest 80 neighbors.[16] Our k-NN forecasts based on matching on only scaled earnings significantly outperform the k-NN forecasts in which we first group on industry (size) and then match on scaled earnings. For example, the MAFE obtained by matching on only scaled earnings is 6.965 percent, which is 0.074 (0.103) percent lower than the MAFE we obtain when we first match on industry (size) and then match on scaled earnings.

These results raise two questions: First, why is it that we obtain such accurate forecasts by matching on only scaled earnings? That is, what are the underlying economics? Second, why is

[15] Inferences are unchanged if we also include $LOSS_{i,t}$, which has an insignificant coefficient in the median HVZ regressions.

[16] We set $k = 80$ neighbors so that these results are comparable to the earlier results in the paper; but, we note that, for some industries, 80 nearest neighbors includes a large portion of the members of that industry. Hence, we conduct sensitivity tests in which we evaluate values of $k \in [20,30,\cdots,120]$. We do this for both the industry- and size-based matching. The un-tabulated results of these tests show that that our conclusions do not depend on k. Inferences also remain unchanged when we use the two-digit SIC industry classification to classify firms.

two years sufficient? The answer to these questions lies in the observation that earnings and earnings growth are the most fundamental indicators of value as evidenced by the work of Ohlson and Juettner-Nauroth (2005) as well as the use of PE and PEG ratios as fundamental ratios used by analysts to compare similar firms.[17] The fact that other financial indicators do not seem to help is due to the so-called "curse of dimensionality," which in layman's terms means that the distance measure becomes noisier as the number of features F and years M increases, which, in turn, raises the forecast error. This result is well known in the machine learning literature. k-NN approaches are known to asymptotically converge to the conditional expectation (e.g., Biau, Cérou, and Guyader, 2010; Chaudhuri and Dasgupta 2014), but the speed of convergence is drastically reduced when the number of dimensions ($F \times M$) along which distances are measured increases (Hastie et al., 2009). What remains to be seen is how well the k-NN model performs given its appeal (approximating the conditional expectation) and limits (being susceptible to the curse of dimensionality because of the few assumptions it makes). As the proverb prescribes, *the proof is in the pudding.*

4.2 Accuracy of k-NN Forecasts

4.2.1 Comparison of Model Forecast Accuracy

Having settled on a preferred k-NN model, we proceed with our second research question: "How accurate are k-NN forecasts compared to those from competing approaches?" In Table 5, we compare the forecasts generated by the k-NN model, the HVZ regression model and the random walk.

[17] We note that even in the HVZ regression model, variables other than earnings have little incremental explanatory power.

The results in Table 5 lead to one overarching conclusion: The k-NN model is clearly superior to either the HVZ model or the random walk. Specifically, for every error metric and for all three forecast horizons, the forecast errors generated by the k-NN model are the smallest. For example, the MAFE for two-year-ahead forecasts based on the nearest neighbor match model is 8.867 percent, which is significantly lower than the MAFE for random walk forecasts and the MAFE from the HVZ model (by 1.210 and 1.277 percent, respectively). The only comparison for which the errors generated by the k-NN model are not significantly smaller is when we compare the MSE of the k-NN model's forecasts of one-year-ahead $EBSI$ to the MSE of the forecasts generated by the random walk. This reflects the influence of outlying observations; and, we note that this result no longer holds when we eliminate the top and bottom 0.1 percent of the forecast errors – i.e., the TMSE of the nearest neighbor match model is smaller and the difference is statistically significant.

4.2.2 Comparison with the BCG Matching Procedure

Extending Barber and Lyon (1996), BCG develop a matching procedure in a context that is different from ours; but, their reason for matching is similar inasmuch as they use matched firms to help them understand (i.e., make an estimate for) the subject firm. Since BCG provide an alternative method of matching, we compare our k-NN forecasts to those generated by their procedure. The essence of the BCG matching procedure is similar to ours in the sense that they match the earnings of the subject firm for year t to the earnings of matched firms for year $t-2$, and then they use the matched-firms' earnings for year $t-1$ as their forecast of subject firm's earnings for year $t+1$.

To obtain firms with attributes similar to those of the subject firm at year t, BCG first separate all firms with available data at time $t-2$ into 30 performance-size bins. Specifically, using return on assets for year $t-2$, ROA, BCG rank firms into two negative ROA groups and four positive

ROA groups. Within each of these six groups, they further separate firms into five size quintiles based on the ranking of average assets for year $t-2$. This yields 30 performance-size bins: ten negative ROA-size bins and twenty positive ROA-size bins. For each firm in a bin, they collect data on its change in ROA and growth in total assets in year $t-1$. Thus, given earnings in year $t-2$, they have an empirical distribution of income for year $t-1$.

To find matches for a particular subject firm-year i,t, BCG match the subject firm's ROA and average assets for year t to a year $t-2$ performance-size bin. Then, they randomly select 50 observations from this bin and they use the mean value of the earnings growth from $t-2$ to $t-1$ of these 50 observations to determine their forecast of the subject firm's earnings for year $t+1$.[18] A more detailed description of the procedure can be found in BCG.

In Table 6, we compare our k-NN forecasts to forecasts obtained from BCG's approach. We consider two versions of BCG's approach; (1) the original approach and (2) a modified approach in which we match on $EBSI$ scaled by equity market value instead of $EBSI$ scaled by total assets. Regardless of which error metric we use or which forecast horizon we consider, we find that our k-NN forecasts are more accurate than the BCG forecasts. For example, the MAFE for the one-year-ahead forecasts generated by our k-NN model is 6.95 percent, which is 0.923 (0.702) percent lower than the MAFE generated by the original (modified) BCG approach.

4.2.3 Comparison of k-NN Forecasts with Analysts' Forecasts

Because I/B/E/S forecasts are used by both academics and practitioners, they are a compelling benchmark. However, analysts' forecasts are not available for all firms. Further, analysts generally forecast "street" earnings, rather than $EBSI$. Thus, to examine how k-NN forecasts compare with

[18] BCG use the mean of the earnings growth of the 50 randomly selected firms. We use the median because we find that it generates a more accurate forecast than the mean.

analysts' forecasts using a like-for-like comparison, we re-estimate the k-NN model using I/B/E/S actual earnings (rather than $EBSI$) for both the subject and matched firms. We also make random walk forecasts using I/B/E/S actual earnings.

The results in Table 7 show that, when MDAFE is the error metric, forecasts generated by the k-NN model are less accurate than analysts' forecasts of both one-year-ahead and two-years-ahead earnings. Specifically, the MDAFE of the k-NN model is 0.322 and 0.130 percent higher than analysts' forecasts of one-year-ahead and two-years-ahead earnings, respectively. However, there is no significant difference in forecast accuracy using any other error metric. Further, it is important to note that the k-NN model is superior to the random walk model for this sample. For example, when evaluating one-year-ahead (two-years-ahead) forecasts, the MDAFE and TMSE of the random walk is 0.113 and 0.532 (0.210 and 0.626) percent higher than the MDAFE and TMSE of the k-NN model, respectively.

4.3 When are Nearest Neighbor Matched Forecasts More or Less Accurate?

4.3.1 k-NN Forecasts for Samples with and without Analyst Coverage

Our next research question is: "When do k-NN forecasts outperform other approaches?" Given our previous results regarding k-NN's accuracy vis-a-vis analysts' forecasts and the necessity of having an accurate forecast model for firms not covered by analysts, we examine forecast accuracy across both the sub-sample of firms that are covered by analysts and those that are not. We show the results of these tests in Panel A of Table 8. 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 the k-NN forecasts is 5.482 percent for the sub-sample of firms that are followed by analysts and 9.505 percent for the sub-sample of firms that are not. In both sub-samples, the k-NN model outperforms both the random walk model and the HVZ model. For example, for the sample of firms that are

not followed by analysts, the MAFE of the HVZ (random walk) forecasts is 0.892 (3.493) percent higher than the MAFE of the k-NN forecasts. Based on these results we draw a key conclusion: k-NN forecasts are the preferred alternative when analysts' forecasts are not available.

4.3.2 Changes in the Accuracy of k-NN Forecasts Over Time

In this section, we examine changes in the performance of the k-NN model over time. We split the sample into four decades. The results are reported in Panel B of Table 8. Regardless of which forecast model we consider, we find that the MAFE and MSE are on average higher for the two most recent decades compared to the two earlier decades. The MDAFE, on the other hand, is relatively constant across decades. The k-NN model generates lower forecast errors in every decade. When we consider the MAFE and MSE, we find that, when compared to the random walk model, the relative superiority of the k-NN model is substantially greater during the last two decades vis-à-vis the first two decades.

4.3.3 Cross-sectional Comparison

To gain a sense of the robustness of our results across different economic circumstances, we partition the observations in several different ways. For ease of exposition, we graph the results of these analyses in Figure 6 and Figure 7. In Figure 6, we show the difference between the MAFE of either the random walk model and the k-NN model or the HVZ model and the k-NN model. That is, for a particular partition, we: (1) calculate the MAFE of the random walk (HVZ) model and the k-NN model; (2) subtract the MAFE of the k-NN model from the MAFE of the random walk (HVZ) model; and, (3) plot the difference. We evaluate four different ways of partitioning the data: (A) on the basis of (look-ahead) realized growth in $EBSI$ in year $t+1$ (i.e., $\left(EBSI_{i,t+1} - EBSI_{i,t}\right)/MVE_{i,t}$); (B) on the basis of the ratio of $EBSI$ in year t to contemporaneous equity market value (i.e., the E/P ratio); (C) on the basis of the ratio of $EBSI$ in year t to

contemporaneous total assets (i.e., profitability); and, (D) on the basis of equity market value at the end of year t (i.e., firm size). In every graph, the darker (lighter) plot reflects the difference between the MAFE of the HVZ (random walk) model and the MAFE of the k-NN model.

Two facts are readily apparent. First, for every quintile of every partitioning variable, the MAFE of the k-NN model is lower than the MAFE of the HVZ model. Second, with the exception of firms with negative growth in $EBSI$ in year $t + 1$, the MAFE of the of the k-NN model is always lower than the MAFE of the random walk model. Moreover, we note that k-NN forecasts are especially more accurate than either the HVZ model or the random walk model for small firms, firms with high growth in $EBSI$ in year $t + 1$ and firms with extreme (high or low) ratios of year t $EBSI$ to contemporaneous equity market value.

In Figure 7, we examine the relative performance of k-NN forecasts by industry. We find that for every one of the FF12 industry groups, the k-NN model has the lowest MAFE; the random walk model has the second lowest MAFE; and, the HVZ forecasts have the highest MAFE. These analyses provide descriptive evidence that the k-NN model is generally better than HVZ and the random walk model regardless of the type of firm that is being analyzed.

4.4 The Relation between Relative Forecast Accuracy and Future Stock Returns

Given the accuracy of the k-NN forecasts, we next examine whether they are associated with future stock returns. To do this, we use a simple procedure that is motivated by Ball and Brown (1968), who show that the sign of the *realized* change in earnings has a positive association with *contemporaneous* stock returns. We extend this idea by evaluating the relation between *forecasts* of earnings changes and *future* stock returns. Specifically, on June 30th of each year, we form two sets of portfolios. To form the first set, we separate firms into two portfolios: P and N. Portfolio P (N) contains all firms for which the k-NN model predicts a positive (negative) change in $EBSI$. To

form the second set of portfolios, we sort firms on the predicted change in $EBSI$ implied by the k-NN model scaled by contemporaneous equity market value (i.e., $\left(FEBSI_{i,t+1}^{kNN} - EBSI_{i,t}\right)/MVE_{i,t}$), and then we form a portfolio for each quintile. We refer to these portfolios as Q5, Q4, …, Q1. Q5 corresponds to the top quintile (i.e., the top 20 percent of the distribution), Q4 corresponds to the next highest quintile, and so on.

Next, for each firm, we compute mean monthly returns for the 12-month period starting on July 1 (i.e., the portfolio formation date) and ending on June 30th of the subsequent year. 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 -0.30 (-30 percent) for NYSE and AMEX firms with missing performance-related delisting returns; and, for NASDAQ firms, we assume a delisting return of -0.55 (-55 percent) for 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.

After forming the portfolios for each calendar-year, we compute the difference between the average mean return of the P and the N (Q5 and the Q1) portfolios. We refer to these as the P-N (Q5-Q1) hedge-portfolio returns. We then average these hedge-portfolio returns across calendar years. Finally, we compare the average hedge-portfolio returns based on our k-NN forecasts (i.e. the k-NN hedge-portfolio returns) to the hedge portfolio returns that are based on HVZ forecasts (i.e., the HVZ hedge-portfolio returns).[19] We do not evaluate the random walk because, by construction, its forecasts imply that earnings will not change.

[19] When forming the positive and negative portfolios (quintiles) for the HVZ model, we separate (sort) firms on the sign of the predicted change in $EBSI$ implied by the HVZ model (the predicted change in $EBSI$ implied by the HVZ model scaled by contemporaneous equity market value (i.e., $\left(FEBSI_{i,t+1}^{HVZ} - EBSI_{i,t}\right)/MVE_{i,t}$)).

In Panel A of Table 9, we show the average monthly returns generated by the hedge portfolios described above. The k-NN hedge-portfolio returns are positive, statistically significant and relatively large. Specifically, the average monthly returns of 19 and 34 basis points generated by the P-N and Q5-Q1 hedge portfolios, respectively, represent annualized average returns of 2.30 and 4.20 percent. Moreover, the average monthly returns of the quintiles are a monotonically increasing function of the quintile number – i.e., quintiles that contain firms with higher forecasted growth generate higher future returns. On the other hand, the HVZ hedge-portfolio returns are *negative* and statistically significant; and, the monthly average returns are a monotonically *decreasing* function of the quintile number. Consequently, the difference between the k-NN hedge-portfolio returns and the HVZ hedge-portfolio returns are positive and statistically significant. They are also economically significant: The average monthly difference in returns for the P-N (Q5-Q1) portfolios is 0.41 percent (0.73 percent), which represents an annualized average return of 5.03 (9.10) percent.

In light of the above, we conclude that forecasts of changes in earnings implied by the k-NN model are associated with future stock returns. This is an intriguing result. Although delving deeply into this result is beyond the scope of the current study, we do provide some initial evidence. Specifically, we evaluate the difference between the k-NN hedge-portfolio returns for firms that are followed by analysts and for firms that are not. We show the results of these tests in Panel B of Table 9. We find that the k-NN hedge-portfolio returns are larger for firms without analyst following. These results buttress the conclusion we draw from the results shown in Panel A of Table 7 – i.e., the k-NN forecasts are the preferred alternative when analysts' forecasts are not available.

Taken together, the results discussed in this sub-section augment the results of our tests of forecast accuracy. Specifically, not only are the earnings predictions generated by the k-NN model more accurate than those generated by other models, they are also more useful in the sense that they are associated with future returns. Moreover, the fact that this association is strongest for firms that are not followed by analysts *suggests* that this association may reflect situations in which the k-NN model identifies growth opportunities that are not reflected in current security prices. Whether this is the case is an interesting avenue for future research.

4.5 Selecting the Best Forecast Model Ex-Ante

The cross-sectional patterns shown in Figures 6 and 7 suggest that it might be possible to predict ex-ante the forecasting model that is best for a particular subject firm. To examine this possibility, we use a random forest classifier, which is a popular machine learning algorithm, to predict, for each subject firm-year, the model with the lowest absolute forecast error. We describe this algorithm and how we implement it in Appendix C. Overall, the random forest classifies 51 percent of the observations correctly out-of-sample. This is considerable, given the no-information rate is 39 percent. However, we find that using the model chosen by the classifier leads to a higher MAFE and MDAFE, and a similar MSE to that which we obtain by using the k-NN model for every firm. The results detailed in Appendix C suggest that more sophisticated approaches such as the random forest, which is among the most popular machine learning approaches, are not better than our simple k-NN model.[20]

[20] Cao and You (2020) also evaluate a random forest model. Our random forest model differs from theirs. We use the random forest to choose the best model ex-ante whereas Cao and You (2020) use the random forest to directly forecast earnings.

5 SUMMARY AND CONCLUSIONS

Expected earnings are important. Managers consider earnings consequences when making investment decisions, and forecasting future earnings is at the heart of equity valuation. Moreover, earnings forecasts are a key variable of interest in many academic studies. Nonetheless, extant earnings forecasting models are a bit dissatisfying. For example, as discussed in Monahan (2018), most models either do not beat the random walk or do not beat it by much.

In this study, we examine the efficacy of non-parametric peer-based earnings forecasts. We eschew complicated models and use a simple nearest neighbor matching (i.e., k-NN) model instead. We do this for two reasons. First, k-NN models allow us to objectively operationalize the advice that financial statement analysis textbooks give to practitioners: Select a set of comparable firms, and then forecast the subject firm's earnings by extrapolating the comparable firms' earnings trends. Second, we view simplicity as a virtue. Simple models are easy to use, understand, replicate and modify; and, they are less subject to overfitting.

Despite its simplicity, our k-NN model performs well. Its forecasts are significantly more accurate than forecasts generated by the random walk, extant regression models, the matching approach developed by BCG and a random forest classifier that uses a sophisticated machine learning algorithm. These results are robust. The k-NN model's superiority holds for sub-samples with and without analyst following, for different sub-periods, etc. Moreover, the earnings growth forecasts implied by our k-NN model are associated with future stock returns, which suggests that the improvements in accuracy we document are economically meaningful.

Our model can be easily modified, and thus it offers various avenues for future research. One interesting avenue is to evaluate whether the distribution of the k-nearest neighbors' realized earnings can be used to measure the degree of uncertainty about the subject firm's earnings. In

addition, our results offer new insights into the link between future earnings and historical earnings properties. The simple k-NN model that only matches on the most recent two years of earnings works best. Adding more features does not lead to better forecasts. This implies that a firm's recent earnings history contains a significant amount of information about what that firm's future earnings will be. The trick to uncovering this information is to put it into the correct context and this can be done by identifying firms with similar earnings 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.

Biau, G., Cérou, F., and Guyader, A., 2010. Rates of Convergence of the Functional K-nearest Neighbor Estimate. *IEEE Transactions on Information Theory* 56 (4): 2034-2040.

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

Bramer, M., 2007. *Principles of Data Mining*. New York: Springer.

Brier, G.W., 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*. 78: 1-3.

Cao, K. and You, H., 2020. Fundamental Analysis Via Machine Learning. Working Paper. Available at SSRN: https://ssrn.com/abstract=3706532

Chaudhuri, K., and Dasgupta, S., 2014. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*.

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

Easton, P., Kapons, M., Kelly, P and Neuhierl, A., 2020. Attrition Bias and Inferences Regarding Earnings Properties: Evidence from Compustat Data. Working Paper: CARE.

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.

Fama, E. and French, K., 1997. Industry Costs of Equity. *Journal of Financial Economics.* 43: 153-193.

Gareth, J., Witten, D., Hastie, T. and Tibshirani, R., 2015. *An Introduction to Statistical Learning*. New York: Springer.

Gerakos, J. and Gramacy, R.B., 2013. Regression-Based Earnings Forecasts. Working Paper: University of Chicago.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learnings – Data Mining, Inference, and Prediction*. New York: Springer.

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.

Makridakis, M. and Hibon, M., 1979. Accuracy of Forecasting: An Empirical Investigation. *Journal of the Royal Statistical Society*. 142 (2): 97-145.

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

Ohlson, J.A. and Juettner-Nauroth, B.E., 2005. Expected EPS and EPS Growth as Determinants of Value. *Review of Accounting Studies*. 10 (2-3): 349-365.

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.

Tian, H., Yim, A., and Newton, D. 2020. Tail heaviness, asymmetry, and profitability forecasting by quantile regression. *Management Science.* 55 (12).

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W., 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14 (1): 137.

FIGURES AND TABLES

Figure 1: IBM Example Study

Figure 1 illustrates the nearest neighbor matching approach. It shows the example of forecast IBM's earnings in 2011 in $t_0 = 2010$. 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 IBM's 2007 to 2010 (*not including 2011*) period are colored darker. Sub-figure B shows the position of the 10 nearest earnings sequences inside the rolling 10-year window used to train the k-NN model. The dashed line represents the $s + 1$ periods used to compute the median forecast for IBM's $t + 1$ earnings.

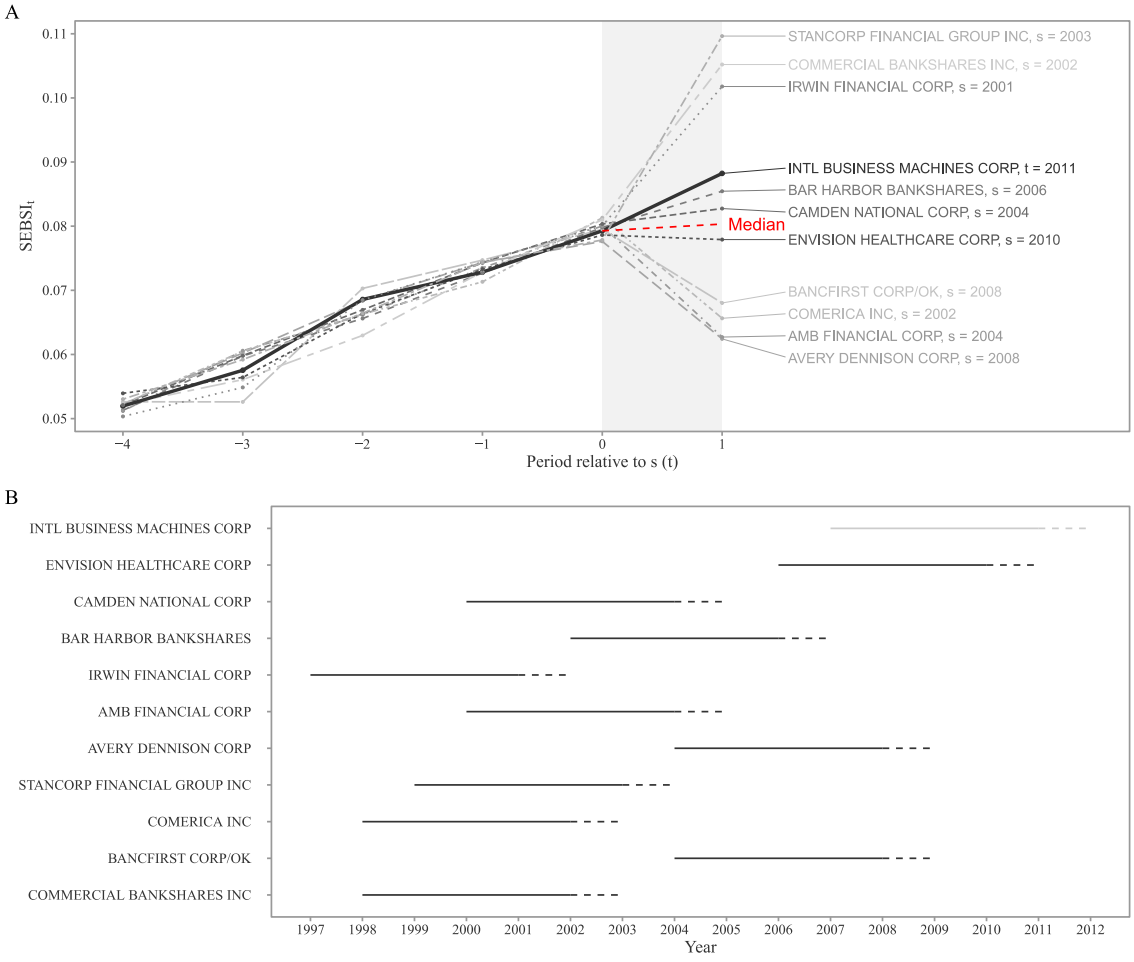

Figure 2: Rolling Out-of-sample Forecasting Procedure

Figure 2 illustrates the setup of the earnings forecasts for a forecast horizon $h = 1$. At the end of June of each year t, we compute earnings forecasts $\hat{E}_{t+1}$ for all firms with fiscal year ends (FYE) from April of year $t - 1$ to March of year t. For regression models, we do so by combining the firms' accounting variables X_t with the fitted coefficients $\hat{\beta}_t$. The coefficients $\hat{\beta}_t$ are fitted by estimating the earnings model on a pooled cross-section that includes all firm-years with FYE between April of year $t - 10$ and March of year t. The same logic applies for estimating K nearest neighbors. We then compare our forecasts, $FEBSI_{t+1}$, with the actual earnings, $EBSI_{t+1}$, of the next fiscal year.

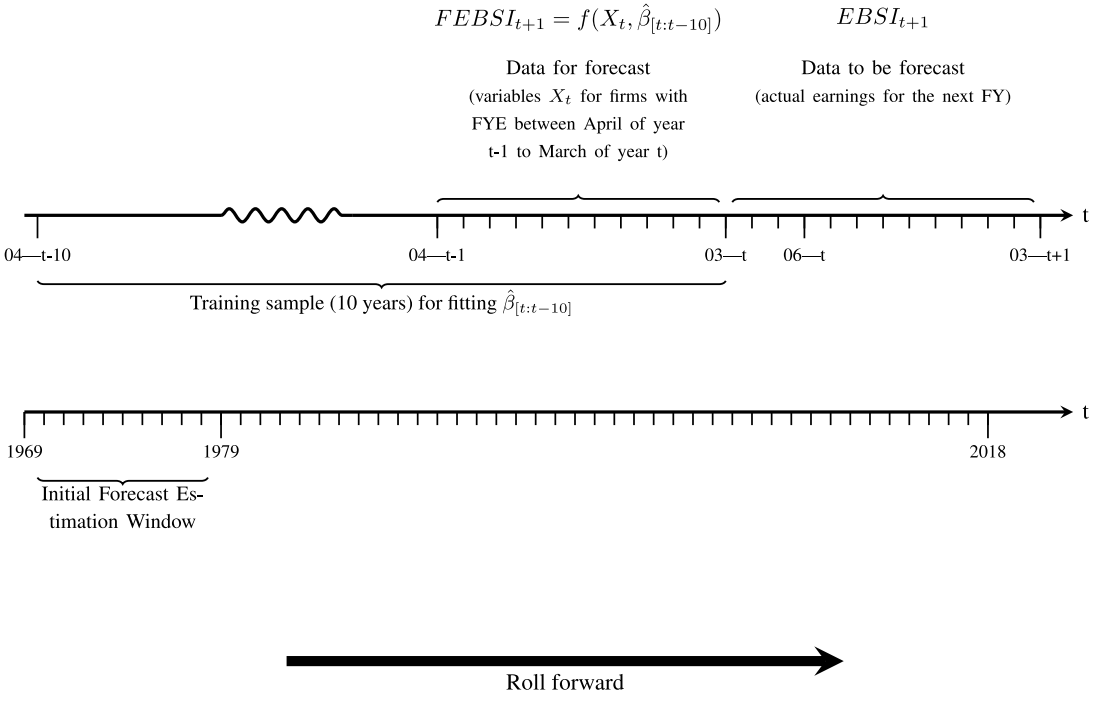

Figure 3: Forecast Coverage by Model

Figure 3 illustrates the forecast coverage for each type of forecast examined in the study. The overall sample described in the figure is the random walk (RW) forecast sample, as defined in Table 1, Panel A. For each firm-year in the RW forecast sample, the figure denotes whether a nearest neighbor match (k-NN), HVZ model (HVZ) or I/B/E/S analyst consensus (ANALYST) forecast is also available for that firm-year. Percentages in parenthesis denote the percentage of the total random walk forecast sample covered by each combination of additional forecast types. See Table 1, Panel B for the total coverage of each forecast type, as well as forecast coverage over ten-year subperiods.

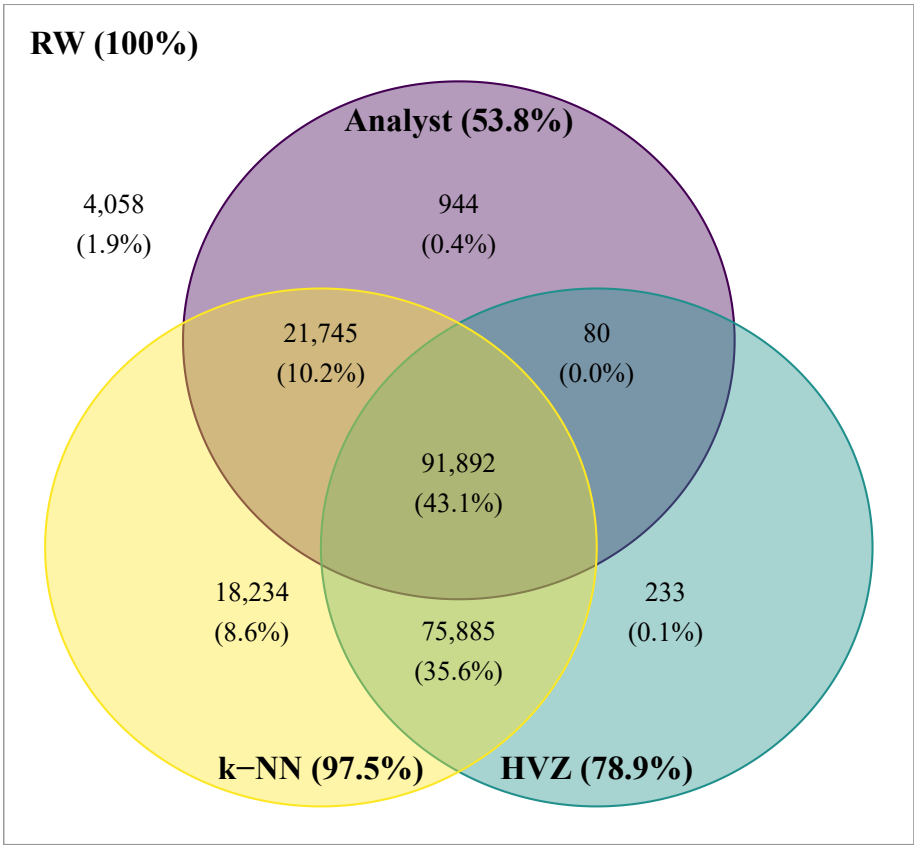

Figure 4: Decomposition of MAFE by Components of the Matching Model

Figure 4 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 only considering most recent earnings ($M = 1$). We examine choices of a forecast median ($s+1$ or s year's earnings) to show the impact of extrapolating the selected firms' trends.

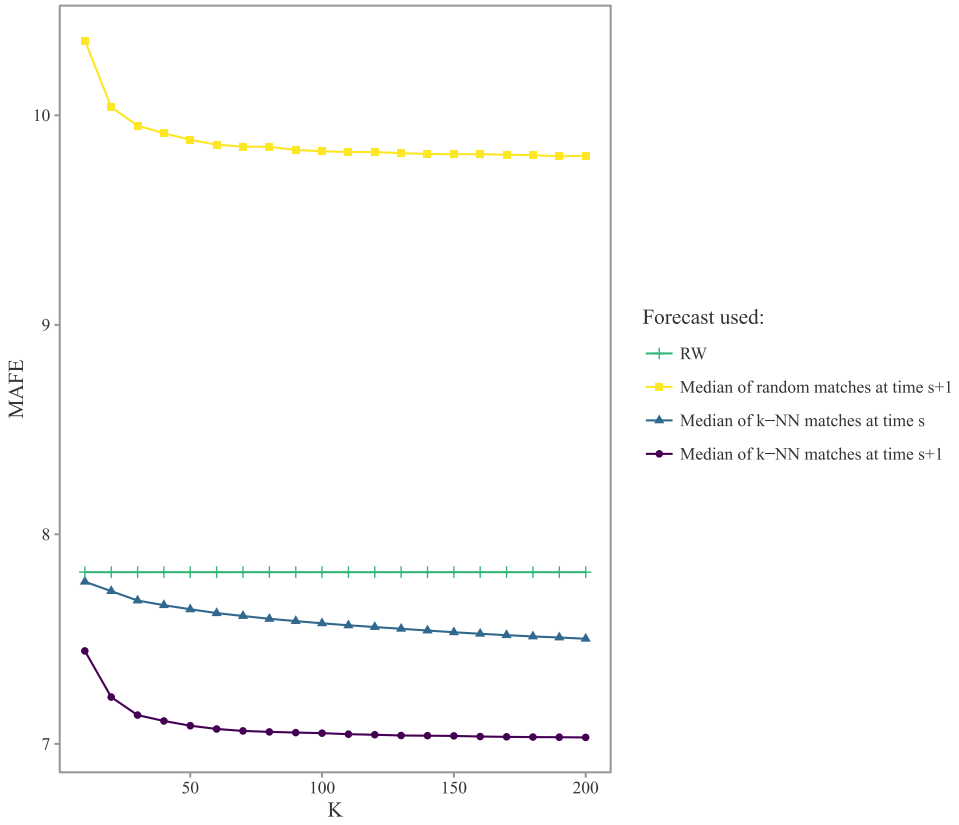

Figure 5: Mean Absolute Forecast Error by K and M

Figure 5 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 peers (k^*) at which decreases in MAFE become insignificant for increasing K by another 10 peers, given a value of M

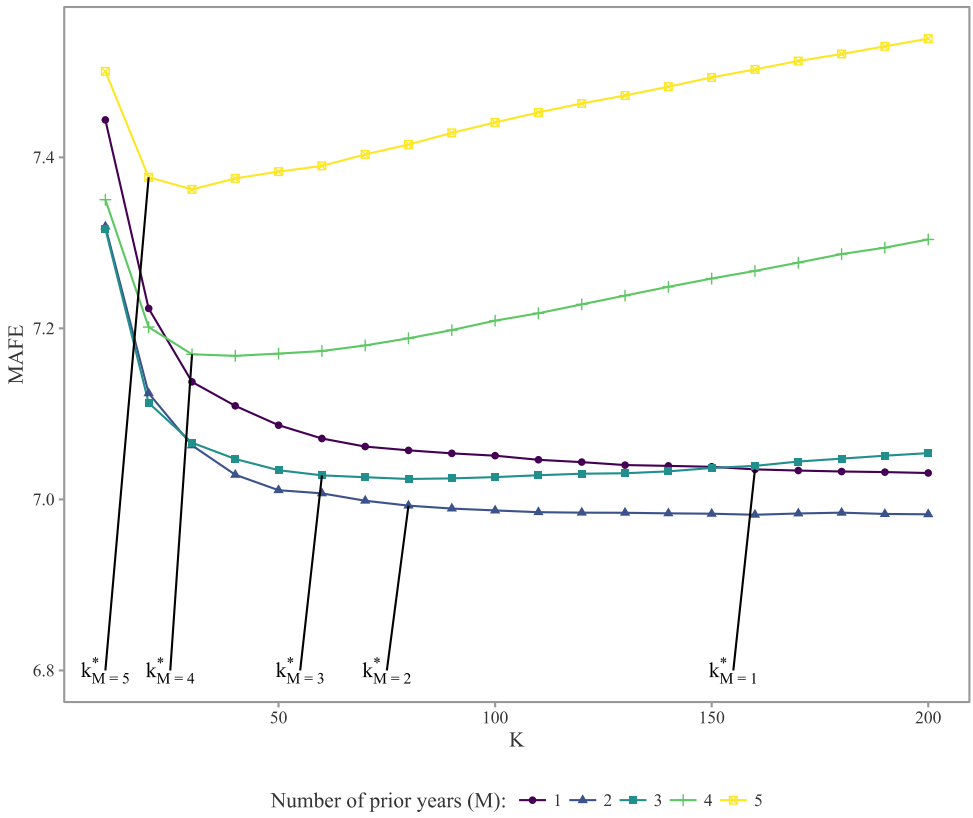

Figure 6: Difference in Mean Absolute Forecast Errors Across Partitions

Figure 6 shows differences in mean absolute forecast errors between the k-NN model and the HVZ model and random walk, respectively. The differences are computed for different quintile sorts. For each sorting variable, quintiles are formed in June of each calendar year. Then, average absolute forecast errors are computed for each calendar-year-quintile portfolio. Those are averaged over all calendar years. Forecast errors are deflated by market value of equity. Models compared are the random-walk model (RW), the regression-based model (HVZ) relative to the nearest neighbor matched model (k-NN)

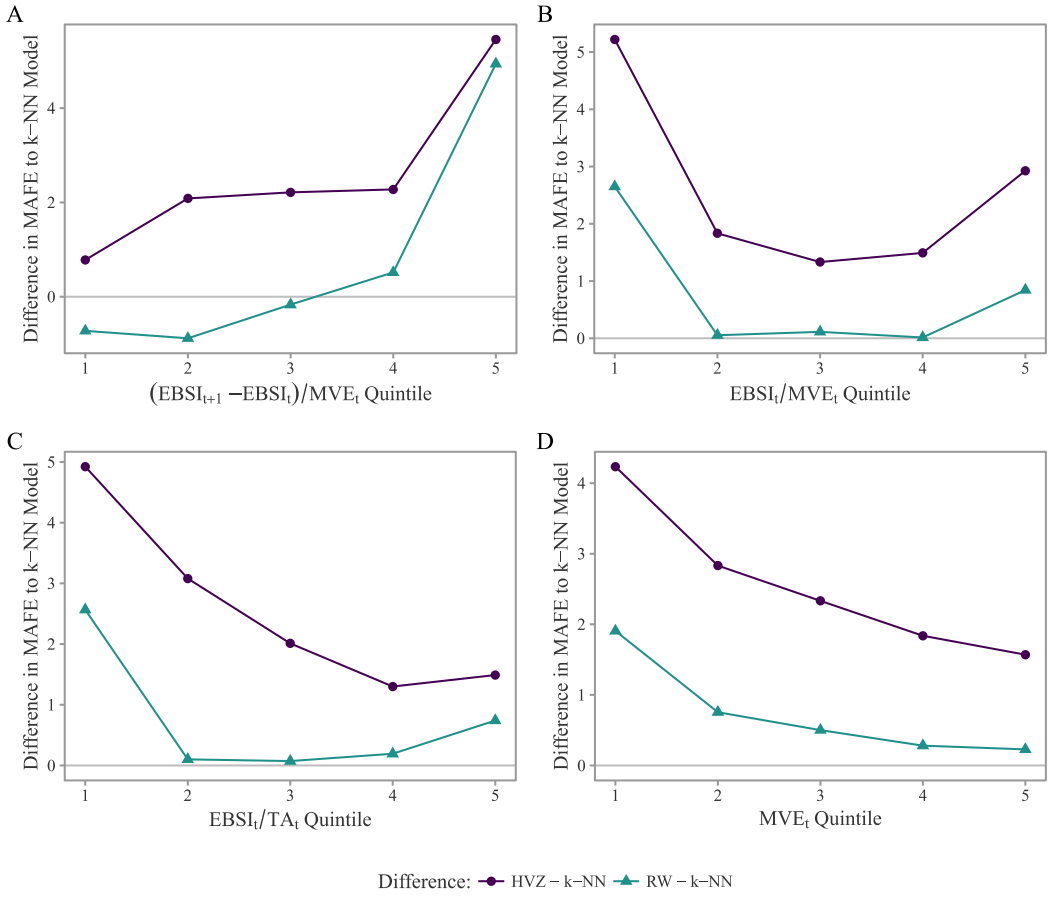

Figure 7: Forecast Errors (t+1) by FF12 Industry

Figure 7 shows median absolute forecast errors for each Model, sorted by Fama-French-12 industry classification. Forecast errors are deflated by market value of equity. Models compared are the random-walk model (RW), the regression-based model (HVZ), and the nearest neighbor matched model (k-NN).

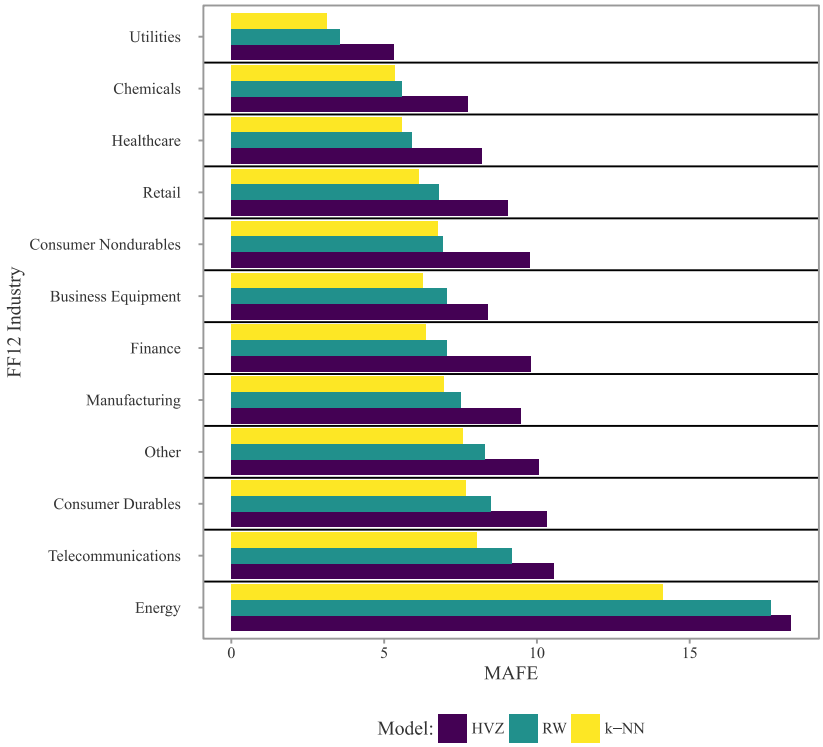

Table 1: Sample Composition

Panel A: Earnings forecast estimation samples

Data Filter	Firm-Years
Total Compustat Observations 1979 – 2018	**339,171**
Less missing *EBSI*	-53,214
Less missing and non-positive deflators	-72,886
Random walk forecast sample	**213,071**
Less missing lagged *EBSI*	-5,315
Nearest neighbor matching forecast sample	**207,756**
Less missing accruals	-39,979
Less missing future *EBSI*	-11,092
Less *MVE* < $10M	-24,646
Forecast Comparison Sample	**132,039**

Panel B: Data requirement comparison by model

Variable availability	1979 - 1988	1989 - 1998	1999 - 2008	2009 - 2018	Total
RW Forecasts Only:	1,466	1,085	920	587	4,058
RW, k-NN Forecasts:	2,830	4,938	6,147	4,319	18,234
RW, HVZ Forecasts:	55	103	59	16	233
RW, Analyst Forecasts:	226	433	168	117	944
RW, k-NN, HVZ Forecasts:	24,591	22,090	19,501	9,703	75,885
RW, k-NN, Analyst Forecasts:	2,596	5,642	6,938	6,569	21,745
RW, HVZ, Analyst Forecasts:	12	40	24	4	80
RW, k-NN, HVZ, Analyst Forecasts:	16,861	25,893	26,711	22,427	91,892
Total sample size					
Total k-NN Forecasts:	46,878	58,563	59,297	43,018	207,756
Total HVZ Forecasts:	41,519	48,126	46,295	32,150	168,090
Total Analyst Forecasts:	19,695	32,008	33,841	29,117	114,661
Total RW Forecasts:	48,637	60,224	60,468	43,742	213,071

Table 1, Panel A shows the effect of our data requirements on the final sample composition. Panel B provides the forecast coverage for each type of forecast examined in the study. For each firm-year in the RW forecast sample, the table denotes whether a nearest neighbor matching (k-NN), HVZ model (HVZ) or I/B/E/S analyst consensus (ANALYST) forecast is also available for that firm-year. The table also provides the total coverage for each type of forecast.

Table 2: Descriptive Statistics

Panel A: Summary statistics for the regression estimation sample

Variable	N	Mean	StD	P05	P25	Med	P75	P95
ACC	198,767	-2.452	441.998	-0.551	-0.111	-0.029	0.011	0.193
DD	198,767	0.395	0.489	0.000	0.000	0.000	1.000	1.000
DIV	198,767	0.357	44.034	0.000	0.000	0.000	0.020	0.068
SEBSI	198,767	-0.611	448.801	-0.433	-0.012	0.049	0.090	0.204
LOSS	198,767	0.272	0.445	0.000	0.000	0.000	1.000	1.000
TA	198,767	53.929	4456.494	0.210	0.624	1.245	2.418	6.464

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

Stat	Intercept	*TA*	*DD*	*DIV*	*EARN*	*LOSS*	*ACC*	Pseudo R^2
Estimate	0.036	-0.035	-0.010	1.052	0.709	-0.003	-0.024	0.424
t-value	[9.874]	[-0.718]	[-2.234]	[4.491]	[9.654]	[-0.18]	[-4.78]	

Panel C: Descriptive statistics for peers of the peer-based models

Variable	N	Mean	StD	P05	P25	Med	P75	P95
(a) Matching and regression variables								
$SEBSI_{t,s}$ - $SEBSI_{t,p}$	207,756	-28.232	12915.6	-0.010	-0.001	0.000	0.000	0.004
$(MVE_{t,s}-MVE_{t,p})/MVE_{t,s}$	207,756	-2.415	163.555	-10.499	-1.254	0.309	0.795	0.972
$(TA_{t,s}-TA_{t,p})/TA_{t,s}$	207,756	-3.274	86.907	-12.198	-1.281	0.330	0.825	0.981
(b) Industry membership of matched observations								
percent same FF12	207,756	0.148	0.102	0.025	0.075	0.125	0.200	0.350
percent same Sic2	207,756	0.059	0.070	0.000	0.013	0.038	0.075	0.212

Table 2, Panel A provides summary statistics for the variables included in the regression and extended peer models. The tabulated statistics are the time-series averages calculated from each rolling 10-year regression sample. Panel B shows the average coefficients of the 10-year rolling window regressions for the HVZ model (Estimate). T-statistics (t−value) are derived from Fama-MacBeth standard errors. Panel C presents descriptive statistics about the peer firms chosen by our peer matching procedure. The suffix *s* denotes the firm to be forecast, suffix *p* denotes the median chosen peer 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 3: Nearest Neighbor Parameter Tuning

	M = 1		M = 2		M = 3		M = 4		M = 5	
K	MAFE	Diff	MAFE	Diff	MAFE	Diff	MAFE	Diff	MAFE	Diff
20	7.223	-0.220***	7.124	-0.195***	7.113	-0.203***	7.202	-0.149***	7.377	-0.124***
30	7.137	-0.086***	7.063	-0.061***	7.066	-0.047***	7.170	-0.032***	7.362	
40	7.109	-0.028***	7.029	-0.034***	7.047	-0.019***	7.168		7.375	
50	7.087	-0.023***	7.011	-0.018***	7.034	-0.013***	7.170		7.383	
60	7.071	-0.016***	7.007	-0.004	7.028	-0.006**	7.174		7.390	
70	7.062	-0.009***	6.998	-0.009**	7.026		7.180		7.403	
80	7.057	-0.004***	6.993	-0.006***	7.024		7.188		7.415	
90	7.054	-0.003**	6.989		7.025		7.198		7.429	
100	7.051	-0.003	6.987		7.026		7.209		7.441	
110	7.046	-0.005***	6.985		7.028		7.218		7.452	
120	7.044	-0.003**	6.985		7.030		7.228		7.463	
130	7.040	-0.004***	6.984		7.031		7.238		7.472	
140	7.039	-0.001	6.984		7.033		7.249		7.483	
150	7.038	-0.001	6.983		7.037		7.258		7.493	
160	7.035	-0.003***	6.982		7.039		7.267		7.503	
170	7.034		6.984		7.044		7.277		7.513	
180	7.033		6.984		7.048		7.287		7.521	
190	7.032		6.983		7.051		7.294		7.530	
200	7.031		6.983		7.054		7.304		7.539	

Table 3 tabulates the mean absolute t+1 forecast error (MAFE) from the k-NN model for various combinations of the model tuning parameters K and M. K is the number of nearest neighbors (peers) used to generate the forecast. M is the number of years of financial data used in the matching process. The analysis is performed on a constant sample of observations with data available for all values of M. Each column examines the effect of increasing K, in increments of 10, for a given value of M. Beside each tabulated value of MAFE, we tabulate the difference in MAFE relative to the MAFE using 10 fewer peers. For each value of N, we stop tabulating differences in MAFE at $k*$, the value of K where additional 10 peer increases in K no longer generate significant improvements in MAFE at the 5 percent level. Statistical significance is determined based on t-statistics clustered by firm and calendar year. ***, **, and * denote statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively. See Table A.1 for formal variable and error metric definitions.

Table 4: Alternative Specifications

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error for different deflators					
k-NN1$_{MVE}$	166,824	6.965	2.412	7.631	1.973
k-NN1$_{BVE}$	166,824	0.330***	0.063***	5.190	0.135***
k-NN1$_{TA}$	166,824	0.452***	0.051***	6.675	0.234***
k-NN1$_{Sale}$	166,824	0.507***	0.065***	6.034*	0.284***
(b) t+1 forecast error for different matching variables					
k-NN1	125,484	6.893	2.529	6.716	1.797
k-NN2	125,484	0.076***	0.055***	0.141***	0.067***
k-NN3	125,484	0.232***	0.190***	0.210**	0.151***
(c) t+1 forecast error for different stratification clusters					
k-NN1$_{MVE}$	166,824	6.965	2.412	7.631	1.973
k-NN$_{FF12}$	166,824	0.074***	0.018	0.405**	0.132***
k-NN$_{Size}$	166,824	0.103***	0.033**	0.549***	0.159***

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. Panel (a) compares the accuracy of the k-NN model using alternate deflators (equity book value, total assets, or sales revenue). Panel (b) compares the accuracy of the k-NN model with that of an expanded model that adds ACC (k-NN2) or the full set of HVZ variables (k-NN3) as matching variables. Panel (c) examines the effect of estimating the k-NN model within strata based on FF12 industry classification or MVE deciles. 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: Forecast Comparison

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error					
k-NN	132,039	6.876	2.557	6.413	1.769
RW - k-NN	132,039	0.727***	0.111***	4.303	0.423***
HVZ - k-NN	132,039	2.553***	1.451***	3.617***	1.161***
(b) t+2 forecast error					
k-NN	121,097	8.867	3.936	5.448	2.546
RW - k-NN	121,097	1.210***	0.117***	7.628**	1.005***
HVZ - k-NN	121,097	1.277***	0.591***	4.697**	0.815***
(c) t+3 forecast error					
k-NN	110,908	10.531	4.872	7.483	3.461
RW - k-NN	110,908	1.284***	0.144***	5.015***	1.340***
HVZ - k-NN	110,908	1.422***	0.804***	2.476*	1.020***

Table 5 tabulates forecast error metrics from the nearest neighbor matching (k-NN) model for various forecast horizons. 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 compared with that from the random walk (RW) and regression-based (HVZ) 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: Forecast Comparison of k-NN and BCG

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error					
k-NN	166,269	6.950	2.409	7.614	1.958
BCG - k-NN	166,269	0.923***	0.210***	5.728**	0.611***
BCG_{MVE} - k-NN	166,269	0.702***	0.212***	3.368**	0.361***
(b) t+2 forecast error					
k-NN	151,531	8.832	3.758	6.842	2.624
BCG - k-NN	151,531	1.286***	0.331***	8.729*	0.970***
BCG_{MVE} - k-NN	151,531	1.434***	0.390***	12.978**	0.991***
(c) t+3 forecast error					
k-NN	137,817	10.548	4.722	26.134	3.451
BCG - k-NN	137,817	1.400***	0.461***	11.924	1.142***
BCG_{MVE} - k-NN	137,817	1.785***	0.537***	15.907*	1.589***

Table 6 tabulates forecast error metrics from the nearest neighbor matching (k-NN) model for various forecast horizons. 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 compared with that from the model suggested by Blouin, Core and Guay (2010). BCG matches observations using asset deflated earnings and assets as proxy for size; BCG_{MVE} matches observations using equity market value deflated earnings and equity market value as proxy for size. 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: Comparison of k-NN Street Earnings Forecasts with Analysts' Forecasts

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error					
k-NN	96,345	7.672	1.584	2350.366	1.389
ANALYST	96,345	-0.955	-0.322***	-56.176	0.031
RW	96,345	0.356	0.113***	-40.871	0.532***
(b) t+2 forecast error					
k-NN	74,679	6.419	2.336	87.634	1.047
ANALYST	74,679	-0.661	-0.130***	-81.512	0.109***
RW	74,679	0.435	0.210***	-60.223	0.626***

Table 7 tabulates forecast error metrics from nearest neighbor (k-NN) forecasts of I/B/E/S street earnings per share. We use I/B/E/S actual EPS_t and EPS_{t-1}, both scaled by fiscal year-end price per share from Compustat (data item *prcc_f*). 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 street earnings model compared with that from the mean I/B/E/S consensus forecast (ANALYST) and a street earnings random walk (RW). 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 8: Analysis of k-NN Forecast Performance Across Sample Partitions

Panel A: Split by analyst coverage

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error with analyst coverage					
k-NN	86,272	5.482	2.033	5.091	1.145
RW	86,272	0.640***	0.142***	5.027	0.252***
HVZ	86,272	2.055***	1.334***	1.744***	0.688***
(b) t+1 forecast error without analyst coverage					
k-NN	45,767	9.505	3.931	8.906	3.114
RW	45,767	0.892***	0.077*	2.937***	0.764***
HVZ	45,767	3.493***	1.739***	7.149***	2.207***

Panel B: Forecast errors split by ten-year sub-samples

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error, 1979 – 1988					
k-NN	28,749	6.375	2.758	2.305	1.455
RW	28,749	0.220	0.156***	0.548	0.020
HVZ	28,749	3.257***	1.505***	4.159**	1.579***
(b) t+1 forecast error, 1989 – 1998					
k-NN	37,485	5.799	2.463	1.753	1.139
RW	37,485	0.339***	0.102***	0.266***	0.148***
HVZ	37,485	1.567***	1.003***	0.992***	0.475***
(c) t+1 forecast error, 1999 – 2008					
k-NN	37,582	7.675	2.635	11.228	2.199
RW	37,582	1.016***	0.071**	2.995**	0.818***
HVZ	37,582	3.096***	1.683***	6.215***	1.796***
(d) t+1 forecast error, 2009 – 2017					
k-NN	28,223	7.755	2.365	10.377	2.763
RW	28,223	1.374***	0.107***	15.230	1.014***
HVZ	28,223	2.423***	1.732***	3.094***	1.331***

Table 8 tabulates forecast error metrics from the nearest neighbor (k-NN) model within sub-samples. 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) models. Panel A examines forecast errors for firm-years with vs. without analyst coverage. Panel B examines forecast errors within ten-year sub-samples, where forecasts are estimated in June of each calendar year t. 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 9: Analysis of k-NN Forecast Performance Across Sample Partitions

Panel A: Average monthly hedge returns for growth portfolio sorts

	k-NN	HVZ	Mean Differences
N	0.0098***	0.0126***	
P	0.0117***	0.0105***	
P-N	0.0019***	-0.0022***	0.0041***
Q1 (Bottom)	0.0100***	0.0139***	
Q2	0.0105***	0.0121***	
Q3	0.0115***	0.0106***	
Q4	0.0115***	0.0102***	
Q5 (Top)	0.0134***	0.0100***	
Q5–Q1	0.0034***	-0.0039***	0.0073***

Panel B: Average monthly hedge returns for growth portfolio sorts by analyst coverage

	k-NN	HVZ	Mean Differences	k-NN	HVZ	Mean Differences
	With analyst following			**Without analyst following**		
N	0.0106***	0.0125***		0.0089***	0.0125***	
P	0.0121***	0.0111***		0.0110***	0.0092***	
P - N	0.0015	-0.0014**	0.0029**	0.0021**	-0.0033***	0.0053***
Q1 (Bottom)	0.0111***	0.0140***		0.0085***	0.0137***	
Q2	0.0108***	0.0124***		0.0094***	0.0107***	
Q3	0.0117***	0.0109***		0.0111***	0.0098***	
Q4	0.0117***	0.0111***		0.0106***	0.0094***	
Q5 (Top)	0.0139***	0.0108***		0.0127***	0.0086***	
Q5–Q1	0.0028***	-0.0032***	0.0059***	0.0042***	-0.0052***	0.0094***

Table 9, Panel A shows average monthly returns and hedge returns for a set of two simple strategies. For all strategies, each year, we form portfolios at the end of June and consider monthly returns over the next 12-months for each firm. For the first strategy, P - N, we sort firms into the (P)ositive portfolio if the model (k-NN or HVZ) forecasts positive earnings growth for next year ($FEBSI_{t+1} - EBSI_t \geq 0$) and into the (N)egative portfolio otherwise. We compute monthly portfolio returns and compute monthly hedge returns as the monthly P portfolio return minus the N portfolio return. The second, Q5 – Q1, strategy sorts firms into quintile portfolios by forecast earnings growth, measured as $(FEBSI_{t+1} - EBSI_t)/MVE_t$. Monthly hedge returns are computed as the monthly Q5 portfolio return minus the monthly Q1 portfolio return. Panel A depicts the average of the monthly hedge returns for both strategies and by forecast model. Panel B expands on the analysis of Panel A by computing average monthly portfolio returns and hedge returns separately for the sub-sample of firm-years with analyst coverage and for the remainder of firm-years without analyst coverage.

APPENDIX A: VARIABLE DEFINITIONS

Table A.1: Variable Definitions

Variable	Definition	Construction
Panel A: Financial variables and fundamental features		
$EBSI_{i,t}$	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	$1(DIV_{i,t} > 0)$
$LOSS_{i,t}$	Indicator variable equal to 1 for firms with negative $SEBSI_{i,t}$ and 0 otherwise	$1(SEBSI_{i,t} < 0)$
Panel B: Forecast evaluation metrics		
MAFE	Mean absolute forecast error (% 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 (% 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% signed forecast errors	

Lowercase variables in the construction column refer to Compustat identifiers.

APPENDIX B: FORECAST ACCURACY OF ALTERNATIVE REGRESSION MODELS

In this appendix we compare the accuracy of k-NN forecasts and forecasts based on the HVZ regression model with other regression models used in the literature. The HVZ model is shown below:

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

In the above equation, $SEBSI_{i,t+h}$ 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; and, $ACC_{i,t}$ denotes firm $i's$ scaled accruals for year t. (When calculating the denominator of $ACC_{i,t}$, we use the same definition of accruals as HVZ.) With the exception of the indicator variables $DD_{i,t}$ and $LOSS_{i,t}$, the variables in equation [B.1] are scaled by firm $i's$ equity market value at the end of year t. We elaborate on how we compute all of our variables in Section 3.2 and Table A.1.

Li & Mohanram (2014) compare the HVZ model with two models that they refer to as the earnings persistence (i.e., EP) model and the residual income (i.e., RI) model:

$$SEBSI_{i,t+h} = \alpha_0 + \alpha_1 \times SEBSI_{i,t} + \alpha_2 \times LOSS_{i,t} + \alpha_3 \times LOSS_{i,t} \times SEBSI_{i,t} + \epsilon_{i,t}. \quad \text{[EP]}$$

$$SEBSI_{i,t+h} = \alpha_0 + \alpha_1 \times ACC_{i,t} + \alpha_2 \times BV_{i,t} + \alpha_3 \times SEBSI_{i,t} + \alpha_4 \times LOSS_{i,t} + \alpha_5 \times LOSS_{i,t} \times SEBSI_{i,t} + \epsilon_{i,t}. \quad \text{[RI]}$$

In the above equation, $BV_{i,t}$ denotes the ratio of firm $i's$ equity book value at the end of year t to its equity market value at the end of year t.

In light of the central role that the loss interaction term (i.e., $LOSS_{i,t} \times SEBSI_{i,t}$) plays in the Li and Mohanram (2014) models, we consider the following modifications to each of the regressions shown above: (1) excluding both the loss indicator and the loss interaction term (denoted model XX_{LNIN}, e.g., HVZ_{LNIN}); (2) including the loss indicator but not the interaction term (denoted model XX_{LYIN}); and, (3) including both the loss indicator and the interaction term (denoted model XX_{LYIY}).

The forecast accuracy of each of the models compared with HVZ_{LYIN}, which is the model we evaluate in the main text and main tables of the paper, is summarized in Table B.1. We estimate the regression-based models with OLS and median regressions.[1] We also add k-NN and RW to the model comparison. We report the results for the one-year-ahead forecasts; the inferences are the same when we evaluate forecast of two- and three-year-ahead $EBSI$.

The models estimated by median regressions do not significantly outperform the HVZ_{LYIN} model. Only the HVZ_{LYIY} model has a negative and significant error difference (MAFE of -0.002 significant at the 10 percent level). The models estimated by ordinary least squares also do not conclusively outperform the HVZ_{LYIN} model. Only the squared error metrics (MSE and TMSE) have significant and negative error differences (for example, EP_{LYIY}_OLS has a TMSE of -0.048 and EP_{LYIY}_OLS has a TMSE of -0.050). We also note that the MAFE and MDAFE are much higher for forecasts based on OLS regressions.

We note that all negative error differences are smaller than those of the k-NN model, which is included for reference. Nonetheless, we formally test whether any of the regression-based models outperforms the k-NN model. In Table B.2, we report the differences between the regression-based

[1] In this Appendix, we winsorize all variables in the estimation models at the extreme percentiles each year. This mimics the methodology of HVZ and of Li and Mohanram (2014). The inferences are unchanged when we do not winsorize the variables in the estimation sample, but, the forecasts obtained from OLS regression-based models include extreme values.

forecast errors and the k-NN-based forecast errors. These results show that the k-NN model outperforms all regression-based models.

Table B.1: Forecast Comparison of Different Regression Models relative to HVZ

Model	N	MAFE	MDAFE	MSE	TMSE
Year t+1 forecast error					
HVZ_{LYIN}	132,039	7.098	2.664	7.829	1.831
HVZ_{LNIN}	132,039	0.110***	0.086***	0.331**	0.044***
HVZ_{LYIY}	132,039	-0.002*	-0.003	-0.004	0.000
EP_{LNIN}	132,039	0.162***	0.110***	0.491*	0.041**
EP_{LYIN}	132,039	0.032**	0.026	0.011	-0.013
EP_{LYIY}	132,039	0.022	0.021	0.003	-0.014
RI_{LNIN}	132,039	0.135***	0.062***	0.459*	0.060***
RI_{LYIN}	132,039	-0.002	-0.019	-0.025	0.002
RI_{LYIY}	132,039	-0.005	-0.021	-0.027	0.001
HVZ_{LNIN}_OLS	132,039	0.610***	1.018***	-0.273**	-0.013
HVZ_{LYIN}_OLS	132,039	0.323***	0.550***	-0.549**	-0.061**
HVZ_{LYIY}_OLS	132,039	0.322***	0.550***	-0.553**	-0.061**
EP_{LNIN}_OLS	132,039	0.757***	1.345***	-0.408	0.010
EP_{LYIN}_OLS	132,039	0.279***	0.554***	-0.752	-0.048**
EP_{LYIY}_OLS	132,039	0.272***	0.547***	-0.756	-0.050**
RI_{LNIN}_OLS	132,039	0.675***	1.105***	-0.437	0.005
RI_{LYIN}_OLS	132,039	0.253***	0.474***	-0.782	-0.060**
RI_{LYIY}_OLS	132,039	0.252***	0.474***	-0.784	-0.060**
k-NN	132,039	-0.204***	-0.105***	-1.353	-0.070***
RW	132,039	0.505***	0.004	2.887**	0.361***

Table B.1 shows forecast error metrics for various forecast horizons for different specifications of the regression-based model, the k-NN model, and the RW model. See Table A.1 for definitions of the forecast evaluation metrics tabulated in each column. 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 B.2: Forecast Comparison of Different Regression Models relative to k-NN

Model	N	MAFE	MDAFE	MSE	TMSE
t+1 forecast error					
k-NN	132,039	6.894	2.558	6.476	1.761
HVZ_{LNIN}	132,039	0.315***	0.192***	1.684	0.114***
HVZ_{LYIN}	132,039	0.204***	0.105***	1.353	0.070***
HVZ_{LYIY}	132,039	0.202***	0.103***	1.349	0.070***
EP_{LNIN}	132,039	0.367***	0.215***	1.843	0.111***
EP_{LYIN}	132,039	0.236***	0.132***	1.364	0.057**
EP_{LYIY}	132,039	0.227***	0.126***	1.356	0.056***
RI_{LNIN}	132,039	0.340***	0.168***	1.812	0.130***
RI_{LYIN}	132,039	0.202***	0.086***	1.328	0.072***
RI_{LYIY}	132,039	0.200***	0.085***	1.326	0.071***
HVZ_{LNIN}_OLS	132,039	0.815***	1.123***	1.080	0.057**
HVZ_{LYIN}_OLS	132,039	0.528***	0.655***	0.804	0.009
HVZ_{LYIY}_OLS	132,039	0.527***	0.656***	0.800	0.009
EP_{LNIN}_OLS	132,039	0.962***	1.450***	0.945	0.080***
EP_{LYIN}_OLS	132,039	0.483***	0.660***	0.601	0.022
EP_{LYIY}_OLS	132,039	0.477***	0.653***	0.596	0.020
RI_{LNIN}_OLS	132,039	0.879***	1.211***	0.916	0.075***
RI_{LYIN}_OLS	132,039	0.457***	0.579***	0.571	0.010
RI_{LYIY}_OLS	132,039	0.456***	0.579***	0.569	0.010
RW	132,039	0.710***	0.109***	4.24	0.431***

Table B.2 tabulates forecast error metrics for various forecast horizons for different specifications of the regression-based model, the k-NN model, and the RW model. See Table A.1 for definitions of the forecast evaluation metrics tabulated in each column. 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.

APPENDIX C: PREDICTING THE BEST MODEL EX-ANTE

The relative accuracy of different forecasting models may vary in *predictable* ways across firms and time. If so, different models will perform better for different firm-years *and* it may be possible (at least on average) to select the best model for each firm-year *ex ante*. In this appendix, we use a random forest classifier to examine this possibility.

We use a random forest algorithm to classify firm-years out of sample. For each firm-year and forecasting model (i.e., the k-NN, the random walk and the HVZ model), the random forest estimates the probability that the model will be the most accurate; and then, we select the forecast generated by the model with the highest estimated probability. The probabilities are a function of 21 observable firm-year features that fall into two groups: (1) fundamental features, which consist of 14 variables that reflect current performance (e.g., $SEBSI_{i,t}$) and firm-specific characteristics such as size, etc. and (2) forecast features, which consist of seven variables that reflect the relative properties of the forecasts (i.e., their dispersion and relative magnitudes).[2]

We find that the random forest classifier correctly classifies 51 percent of the observations out-of-sample. This is impressive given that the no-information rate is 39 percent. However, when compared to the k-NN forecasts, we find that the forecasts chosen by the classifier have a higher MAFE, MDAFE and MSE. Our conclusion is that even using sophisticated approaches such as the random forest, which is among the most popular machine learning approaches, does not improve upon our simple k-NN model.[3]

[2] Random forests are a type of supervised machine learning. Given a set of training data that contains observed classes and observed features, they learn the probabilities (given the features) that a firm-year will fall into each of the different classes. In our setting, a class is a set of firm-years for which a particular forecasting model is the most accurate.

[3] We did not expect this result. We began this research project with the idea that we would compare the three models, and then use the random forest to select the best model for each firm-year. *Ex post*, we realize that the simplicity of the k-NN model makes it appropriate for forecasting earnings. Earnings forecasting is a setting with a large number of highly correlated financial variables, which reduces the effective sample size available for more data-intensive methods that try to exploit those variables (Hastie et al., 2009; van der Ploeg, Austin, Steyerberg, 2014).

Of course, this "no result" must be based on a thorough analyses of random forest classification. Ergo, this appendix.

The Random Forest Model

Decision Tree Learning

Our goal is to select for each firm-year the forecasting model (i.e., the k-NN model, the random walk or the HVZ model) that is the most accurate for that firm-year. This is a classification problem in which class C_{FM} is the set of firm-years for which forecasting model FM is the most accurate. We refer to members of that class as "model FM firm years;" so, for example, when we use the expression "k-NN firm-year" we are referring to a firm-year for which the k-NN model generates the most accurate forecast.

A straightforward way of dealing with this classification problem is to use decision tree learning, which is a recursive trial and error process. In the first step of the process the entire set of training data is split into two subsets. This is done by separately evaluating each of the observable features and identifying the feature and splitting rule that maximizes the purity of the resulting subsets. Perfect purity is achieved if each resulting subset consists of firm-years that are all from the same class (e.g., k-NN firm-year forecasts with k-NN firm-year forecasts, random walk firm-year forecasts with random walk firm-year forecasts and HVZ firm-year forecasts with HVZ firm-year forecasts, etc.). On the other hand, perfect *im*purity is achieved if each resulting subset consists of an equal number of observations from each of the different classes (e.g., if there are N observations in the dataset, the subsets will each consist of $\frac{N}{3}$ k-NN firm-year forecasts, $\frac{N}{3}$ random walk firm-year forecasts and $\frac{N}{3}$ HVZ firm-year forecasts).

After completing step one, we repeat the splitting process described above for each of the resulting subsets, and then for the resulting subsets of the subsets, etc. until either: (1) perfect purity

is achieved, which is rare, or (2) there are no further improvements in purity. Upon completion of the process, we are left with a "tree" that starts with the entire dataset, and then branches out to subsets, which branch out to further subsets, etc. until the purest subsets, which are referred to as "leafs," are reached. The fraction of observations within a particular leaf that belong to class FM is an estimate of the probability that the observations in that leaf are a member of class FM. For example, if a leaf contains 900 k-NN firm-year forecasts, 75 random walk firm-year forecasts and 25 HVZ firm-year forecasts, the probabilities that a firm-year forecast in the leaf is a k-NN forecast, a random walk forecast and an HVZ forecast are 0.90, 0.075 and 0.025, respectively.

Finally, we use the tree to assign out-of-sample probabilities to each subject firm-year. Specifically, we use the observable features of each subject firm-year to determine the leaf that it belongs to. Then, for this subject firm-year, we assign the probabilities per the leaf to the forecasting models. For instance, if the subject firm-year belongs to the leaf described at the end of the previous paragraph, we assign probabilities of 0.90, 0.075 and 0.025 to the k-NN model, random walk model and the HVZ model, respectively

The above approach to classification is quite popular and useful. For example, on page 352 of the second edition of *The Elements of Statistical Learning*, Hastie, Tibshirani and Friedman (2009) provide the following summary of decision trees:

> Of all the well-known learning methods, decision trees come closest to meeting the requirements for serving as an off-the-shelf procedure for data mining. They are relatively fast to construct and they produce interpretable models (if the trees are small). … [T]hey naturally incorporate mixtures of numeric and categorical predictor variables and missing values. They are invariant under (strictly monotone) transformations of the individual predictors. As a result, scaling and/or more general transformations are not an issue, and they are immune to the effects of predictor outliers. They perform internal feature selection as an integral part of the procedure. They are thereby resistant, if not completely immune, to the inclusion of many irrelevant predictor variables. These properties of decision trees are largely the reason that they have emerged as the most popular learning method for data mining.

However, as Hastie, Tibshirani and Friedman (2009) hasten to point out, there is a crucial caveat to the above: Decision trees tend to be inaccurate. As discussed in Gareth, Witten, Hastie and Tibshirani (2015), small changes in the training data can lead to large changes in the structure of the tree and the predictions generated by it. And, as discussed in Bramer (2007), decision trees tend to overfit the data. Consequently, although probabilities obtained from decision trees have low bias, they also have high variance. Consequently, they are inaccurate.

Combining Decision Trees into a Random Forest

We use a common approach for dealing with the problem described above: Random forests (e.g., Hastie, Tibshirani and Friedman, 2009). Specifically, we form a collection of B separate decision trees. We then use each tree to assign probabilities to the observations in a *randomly* selected (with replacement) sub-sample of the training data. (This is referred to as bootstrap aggregation or "bagging.") Moreover, when splitting the sub-sample and each subsequent subset of it, we consider a different set of *randomly* selected features. (This is referred to as "feature bagging.") Hence, we form a *random forest* of B de-correlated trees.

As discussed at the end of the previous sub-section, each of the trees in our random forest generates noisy estimates of the probabilities. However, because the trees are de-correlated (or random), this noise is *idiosyncratic*. This implies that the average of the probabilities obtained from the B trees are less noisy and more accurate (often much more accurate) than the probabilities implied by any single tree. Consequently, for each subject firm-year, we follow a three-step classification process. In the first step, we determine the probability of each class that is implied by each of the B trees in our random forest. Second, for each class, we calculate the average of the B probabilities, and then we assign this average probability to the subject firm-year. Finally, we

assign the subject firm-year to the class with the highest average probability. In the tables we refer to the random forest as model RF.

Accuracy of the Random Forest Algorithm

To train the random forest classifier we evaluated a large selection of potential features. There are multiple ways of generating features from the same data and it is beneficial for the performance of a random forest model to not unduly increase the feature space with too many highly correlated features. For example, including a loss indicator or an interaction term is unnecessary because non-linearities are embedded into each tree as that tree is "grown." From an initial, large pool of potential variables, we end up using 21 features that fall into two groups. The first group contains 14 variables that reflect firm-level fundamentals such as reported accounting numbers (e.g., total assets, $LgTA_{i,t}$), financial performance (e.g., $SEBSI_{i,t}$) or capital structure (e.g., leverage, $LEV_{i,t}$). The second group consists of seven variables that reflect properties of the forecasts generated by the three models. For example, $FSTD_{i,t}$ is the standard deviation of the forecasts of $SEBSI_{i,t+1}$ generated by the three different models. We refer to the variables in the first group as "fundamental features" and the variables in the second group as "forecast features." The definitions of all 21 features are provided in Table C.1. We elaborate on the features, their importance and relations with the outcome variables in the next sub-section.

In Panel A of Table C.2, we show a confusion matrix and related statistics. The rows of the matrix correspond to out-of-sample classifications per the random forest and the columns correspond to the actual (or "true") classifications. For example, the random walk generates the best forecast for 41,081 of the firm-years in the sample. However, the random forest classifies only 39,677 of the firm-years as random walk "RW" firms. It also selects the HVZ model too infrequently: HVZ is the best model for 32,199 of the firm-years but it is selected by the random

forest for only 21,208 of the firm-years. Taken together, these results imply that the random forest selects the k-NN model too often, which is the case: k-NN is selected for 59,599 of the firm-years but it is the correct model for only 47,204 of the firm-years.

The results regarding recall, precision and accuracy provide further evidence on the usefulness of the random forest. "Recall" reflects the fraction of observations that belong to a class *and* are correctly classified. (It is also referred to as the true positive rate.) The random forest has good recall with regards to the k-NN model (67.8 percent), moderate recall with regards to the random walk (49.1 percent) but low recall with regards to the HVZ model (29.5 percent). Regarding precision, which reflects the fraction of classifications that are correct, the random forest continues to do well with regards to k-NN (53.7 percent) and RW (50.8 percent) but it is less precise when it selects the HVZ model (44.8 percent). Finally, the overall accuracy of the random forest is 51.2 percent – i.e., it correctly classifies half of the firm-years in the sample, which is good given that the no-information rate is 39.2 percent.

The results shown above imply that the random forest is not a straw man. Consequently, the obvious question is: When compared to the random forest, how does the k-NN model fare? With this question in mind, in Panel B of Table C.2, we compare the k-NN model to the random forest; and, we find that although the random forest selects the correct model for 51.2 percent of the subject firm-years, it is significantly worse in terms of MAFE, MDAFE and MSE than the k-NN model. For example, the MAFE for the RF model is slightly, although significantly higher than the MAFE for the k-NN model (0.12 percent) while the MAFE for the RW and HVZ models are much greater than that for the k-NN model (0.754 and 2.566 percent, respectively).

One concern about the random forest is that it might be unnecessarily complex. It uses bagging and feature bagging to build a set of de-correlated trees, it then obtains the probabilities implied

by the trees and uses the average of these probabilities to select a forecasting model for each firm-year. This seems complicated and computationally expensive and begs the question – might there be a simpler way of combining forecasts? For instance, rather than using averages of probabilities, why not simply use the average of the forecasts? This approach is much simpler; and, as shown in Makridakis and Hibon (1979), the average forecast is often more accurate than forecasts generated by complicated models.

We consider two new models, which we refer to as the AVG model and the MED model, which equal the average and the median of the three forecasts, respectively. We compare these two forecasts to the forecasts generated by the k-NN model, the random forest and a fifth model, which we refer to as the weighted random forest or model WRF. To compute the forecast for model WRF, we first obtain the probabilities assigned by the random forest to each of the forecasts generated by the three models. We then use these probabilities as weights and we calculate the weighted average of the forecasts implied by the three models. We compare the forecasts from these five models in Table C.3.

The results in Table C.3 lead to two conclusions: (1) the random forest is superior to the average, median or weighted random forest and (2) more importantly, the k-NN model is superior to all of the models. This second point is important because it implies that the k-NN model is quite robust. It is better than the random walk; regression-based models; the matching model proposed by BCG; sophisticated machine-learning algorithms such as the random forest; and, "averaging" approaches that use the average, median or weighted average of the forecasts.

Feature Importance

In this section we address the question: when determining class membership, which features matter most? We consider 21 features that fall into two groups: (1) fundamental features and (2) forecast features. We provide definitions for all the features in Table C.1.

To evaluate the relative importance of the features, we follow a six-step process. First, for each individual tree b in the random forest of B trees we identify the set of firm-years that are *not* in the subset of training data that we use to grow tree b. We refer to this subset as the out-of-bag sample for b (i.e., $OOB4b$) and it contains N_{OOB4b} firm-years. Second, for each firm-year in $OOB4b$, we use tree b to determine the predicted probability for each class. These are referred to as the out-of-bag probabilities. Third, we determine the out-of-bag errors by subtracting each out-of-bag probability from either: (1) one if the corresponding class is the true class or (2) zero if the corresponding class is not the true class. Hence, for each tree, we have N_{OOB4b} out-of-bag errors for each class, which, given there are three classes, implies that we have a total of $3 \times (\sum_{b \in B} N_{OOB4b})$ out-of-bag errors. Fourth, we calculate the mean of the squared out-of-bag errors, which is the Brier (1950) score. We refer to this as the "true" Brier score. Fifth, for each of the 21 features, we repeat steps one through four. However, when computing the predicted probabilities, we randomly permute (or shuffle) the values of the feature so that, for each firm-year in $OOB4b$, the feature is assigned a value from another randomly selected firm-year in $OOB4b$. Consequently, the resulting "permuted" Brier score equals the decrease in accuracy attributable to replacing the feature with a random noise variable that is drawn from the same distribution as the feature itself. Finally, in step six, we calculate the permutation importance of each feature, which equals the difference between the permuted Brier score and the true Brier score.

We document the permutation importance of the 21 features in Figure C.1. A key result stands out: Forecast features are much more important than fundamental features. Specifically, the five most important features are all forecast features. This is striking given that there are only seven forecast features and that the fundamental features outnumber the forecast features by a ratio of two to one. Consequently, when selecting from a set of models, the relative values of the forecasts generated by the models play a much more important role in determining the selected model than either: (1) the model inputs (i.e., the regression predictors and the features used for matching) or (2) firm-level characteristics such as size, capital structure, etc.

Figures and Tables

Figure C.1: Variable Importance

Figure C.1 shows the average permutation importance, in percent, averaged over the ten rolling estimation samples in each subperiod. The permutation importance of each feature, which equals the difference between the permuted Brier score and the true Brier score (Brier, 1950. See Section) It is computed as following a six-step process. First, for each individual tree (b) in the random forest of B trees we identify the set of firm-years that are *not* in the subset of training data that we use to "grow" tree b. We refer to this subset as the out-of-bag sample for b (i.e., (OOB4b)) and it contains N OOB4b firm-years. Second, for each firm-year in OOB4b, we use tree b to determine the predicted probability for each class. These are referred to as the out-of-bag probabilities. Third, we determine the out-of-bag errors by subtracting each out-of-bag probability from either: (1) one if the corresponding class is the true class or (2) zero if the corresponding class is not the true class. Hence, for each tree we have N_OOB4b out-of-bag errors for each class. Fourth, we calculate the mean of the squared out-of-bag errors, which is the Brier (1950) score. We refer to this as the ``true" Brier score. Fifth, for each of the 19 features, we repeat steps one through four. However, when computing the predicted probabilities, we randomly permute (or shuffle) the values of the feature so that, for each firm-year in OOB4b, the feature is assigned a value from another randomly selected firm-year in OOB4b. Consequently, the resulting ``permuted" Brier score equals the decrease in accuracy attributable to replacing the feature with a random noise variable that is drawn from the same distribution as the feature itself. Finally, in step six, we calculate the permutation importance of each feature, which equals the difference between the permuted Brier score and the true Brier score.

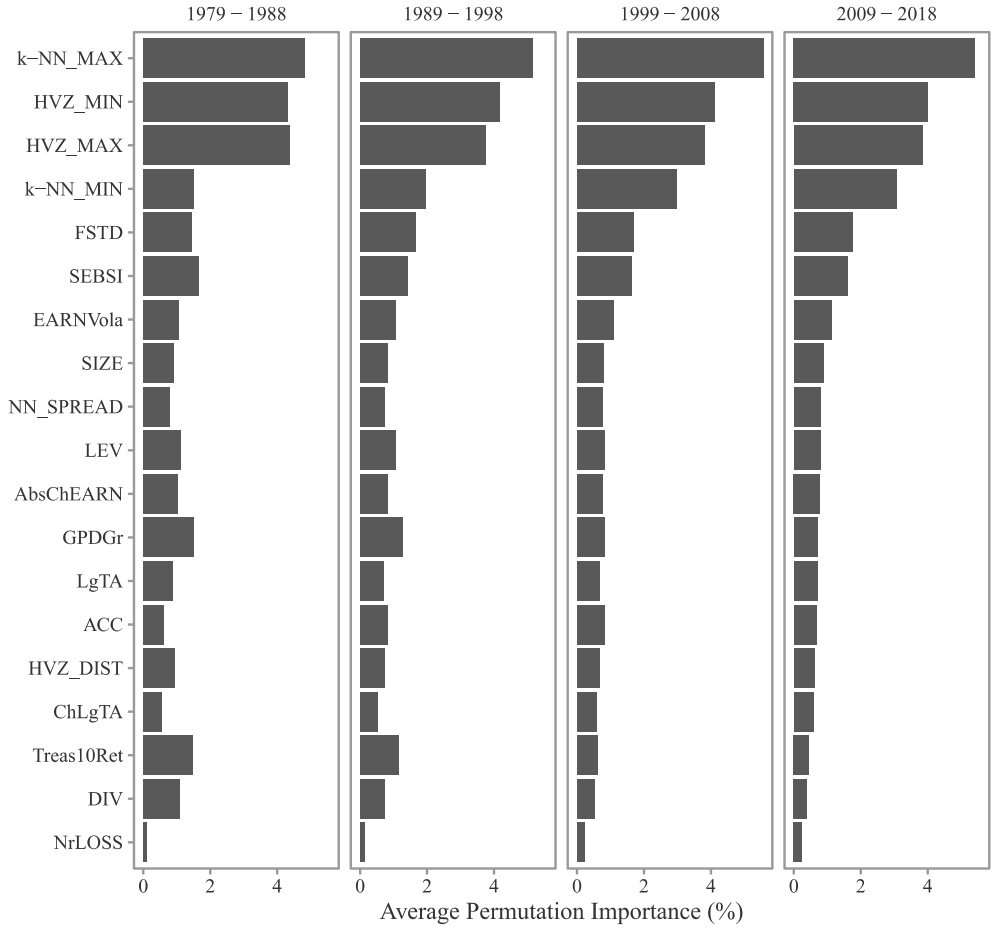

Table C.1: Variable Definitions

Variable	Definition	Construction
Panel A: Financial variables and fundamental features		
$AbsChEARN_{i,t}$	Absolute change in current earnings	$abs(SEBSI_{i,t} - SEBSI_{i,t-1})$
$ACC_{i,t}$	Accruals for firm i at time t scaled by $MVE_{i,t}$	$(\Delta(act_{i,t} - che_{i,t}) - \Delta(lct_{i,t} - dlc_{i,t} - txp_{i,t}) - dp_{i,t}) / MVE_{i,t}$
$ChLgTA_{i,t}$	Change in LgTA	$log(at_{i,t}) - log(at_{i,t-1})$
$DIV_{i,t}$	Dividends for firm i at time t scaled by $MVE_{i,t}$	$dvc_{i,t} / MVE_{i,t}$
$EARN\_VOLA_{i,t}$	Standard deviation of $SEBSI$ over past three years	$sd(SEBSI_{i,t}, SEBSI_{i,t-1}, SEBSI_{i,t-2})$
$EBSI_{i,t}$	Earnings before special items for firm i at time t	$ib_{i,t} - spi_{i,t}$
$FSEBSI_{i,t+h}$	Forecast of $EBSI_{i,t+h}$ scaled by $MVE_{i,t}$	
$GDPGr_{i,t}$	Growth in current gross domestic product (Federal Reserve Bank data)	$GDP_{i,t} / GDP_{i,t-1} - 1$
$MVE_{i,t}$	Equity market value for firm i at the end of fiscal year t	$prcc\_f_{i,t} * csho_{i,t}$
$NrLOSS_{i,t}$	Count of the number of loss-making fiscal periods within the last three years	
$LEV_{i,t}$	Total assets for firm i at time t scaled by $MVE_{i,t}$	$at_{i,t} / MVE_{i,t}$
$LgTA_{i,t}$	Logarithm of total assets	$log(at_{i,t})$
$SEBSI_{i,t}$	$EBSI_{i,t}$ scaled by $MVE_{i,t}$	$(ib_{i,t} - spi_{i,t}) / MVE_{i,t}$
$SIZE_{,t}$	Logarithm of $MVE_{i,t}$	$log(MVE_{i,t})$
$Treas10Ret_{i,t}$	Current period return on a 10-year treasury bond (Federal Reserve Bank data)	
Panel B: Forecast features		
$FSTD_{i,t}$	Standard deviation of forecasts of the three models, k-NN, HVZ, and RW	$sd(FSEBSI^{KNN}_{i,t+1}, FSEBSI^{HVZ}_{i,t+1}, FEARN^{RW}_{I,t+1})$
$HVZ\_DIST_{i,t}$	Euclidean distance between the values of the HVZ variables for i in t and the average values of the HVZ variables in the estimation window. Variables are standardized before computing the distance	
$HVZ\_MAX_{i,t}$	Indicator variable equal to 1 if the k-NN forecast is higher than the HVZ and RW forecast and 0 otherwise	$I(FSEBSI^{HVZ}_{i,t+1} >= FSEBSI^{k\text{-}NN}_{i,t+1} \& FSEBSI^{HVZ}_{i,t+1} >= FSEBSI^{RW}_{I,t+1})$
$HVZ\_MIN_{i,t}$	Indicator variable equal to 1 if the k-NN forecast is higher than the HVZ and RW forecast and 0 otherwise	$I(FSEBSI^{HVZ}_{i,t+1} <= FSEBSI^{k\text{-}NN}_{i,t+1} \& FSEBSI^{HVZ}_{i,t+1} <= FSEBSI^{RW}_{I,t+1})$
$k\text{-}NN\_MAX_{i,t}$	Indicator variable equal to 1 if the k-NN forecast is higher than the HVZ and RW forecast and 0 otherwise	$I(FSEBSI^{k\text{-}NN}_{i,t+1} >= FSEBSI^{HVZ}_{i,t+1} \& FSEBSI^{k\text{-}NN}_{i,t+1} >= FSEBSI^{RW}_{I,t+1})$
$k\text{-}NN\_MIN_{i,t}$	Indicator variable equal to 1 if the k-NN forecast is higher than the HVZ and RW forecast and 0 otherwise	$I(FSEBSI^{k\text{-}NN}_{i,t+1} <= FSEBSI^{HVZ}_{i,t+1} \& FSEBSI^{k\text{-}NN}_{i,t+1} <= FSEBSI^{RW}_{I,t+1})$
$k\text{-}NN\_SPREAD_{i,t}$	Standard deviation of next year's earnings ($EBSI_{i,h+1}$) of the K=80 peers used to compute the k-NN forecast	$sd(SEBSI_{i,h+1} * MVE_{i,t}) / MVE_{i,t}$

Lowercase variables in the construction column refer to Compustat identifiers.

Table C.2: Predicting the best model to use by observation

Panel A: Confusion matrix

Predicted	k-NN	HVZ	RW	Total Predicted	Precision	Recall
k-NN	32,018	13,069	14,512	59,599	0.537	0.678
HVZ	5,305	9,501	6,402	21,208	0.448	0.295
RW	9,881	9,629	20,167	39,677	0.508	0.491
Total Reference	47,204	32,199	41,081	120,484		
Accuracy	**0.512**					
No-information rate	**0.392**					
Accuracy P-value	**0.000**					

Panel B: Overall forecast error comparison

Model	N	MAFE	MDAFE	MSE	TMSE
k-NN	120,484	6.923	2.545	6.666	1.805
RW	120,484	0.754***	0.085***	4.688	0.444***
HVZ	120,484	2.566***	1.461***	3.786***	1.185***
RF	120,484	0.122**	0.041**	0.458*	0.012

Table C.2 shows predicted outcomes as per the random forest and realized outcomes for each classified model. The rows of the matrix correspond to out-of-sample classifications as per the random forest. The columns correspond to the realized ("true") outcomes (classifications). Recall reflects the fraction of observations that belong to a class and are correctly classified. Precision reflects the fraction of predicted classifications that are correct.

Table C.3: Predicting the best model to use by observation

Model	N	MAFE	MDAFE	MSE	TMSE
k-NN	120,484	6.923	2.545	6.666	1.805
RF	120,484	0.122**	0.041**	0.458*	0.012
WRF	120,484	0.155***	0.097***	0.239	-0.006
AVG	120,484	0.549***	0.312***	0.956	0.183***
MED	120,484	0.284***	0.031	1.094**	0.133***

Table C.3 compares the mean (MEAN) and median (MED) forecast of our three forecast models to the random forest (RF) and a weighted random forest (WRF). WRF is weighting each model (k-NN, HVZ, and RW) by the random forest predicted probability of that model being the lowest forecast error model. PCT_BEST is the percentage of times that forecast is the most accurate forecast. MAFE is the mean absolute forecast error (% of MVE). MDAFE is the median absolute forecast error (% of MVE). MSE is the mean of squared forecast error. TMSE is the mean of squared forecast error after truncating the top and bottom 0.1% signed forecast errors