Evaluation and Ranking of Market Forecasters

David H. Bailey* Jonathan M. Borwein† Amir Salehipour‡
Marcos López de Prado §

May 31, 2017

Abstract

Many investors rely on market experts and forecasters when making investment decisions, such as when to buy or sell securities. Ranking and grading market forecasters provides investors with metrics on which they may choose forecasters with the best record of accuracy for their particular market exposure. This study develops a novel ranking methodology to rank the market forecaster. In particular, we distinguish forecasts by their specificity, rather than considering all predictions and forecasts equally important, and we also analyze the impact of the number of forecasts made by a particular forecaster. We have applied our methodology on a dataset including 6,627 forecasts made by 68 forecasters.

Key words: Market forecasters ranking; Market forecasters evaluation; Guru ranking; Market forecast

1 Introduction and Background

Many investors rely on market experts and forecasters when making investment decisions, in a sense that the investors follow these forecasts when buying or selling securities. Needless to say, some of these forecasts turn out to be more accurate than others. Ranking and grading market forecasters provides investors with metrics on which they may choose forecasters with the best record of accuracy for their particular market exposure.

Some of these forecasts are optimistic, while others are pessimistic. One example of a relatively optimistic forecast was by Thomas Lee, who on 3 January 2015 predicted that the S&P 500 index would be at 2325 one year hence [**Udland**]. (The S&P 500 ranged between 1867 and 2122 during this period, closing at 2012 on 4 January 2016, well short of the goal.) One example of a relatively pessimistic forecast was made by Chapman University professor Terry Burnham, who in July 2013 forecasted that the Dow Jones Industrial Average (DJIA) would drop to 5,000 before it topped 20,000 [**Burnham2013**]; he repeated this forecast in May 2014 [**Burnham2014**]. (The DJIA exceeded 20,000 on 25 January 2017, having never dropped below 14,700 during the period 1 July 2013 through 25 January 2017.)

There have been several previous analyses of forecaster accuracy, both in academic literature and also in the financial press.

As a single example, recently Nir Kaissar analyzed a set of strategists' predictions from 1999 through November 2016 [**Kaissar**]. He found a relatively high correlation coefficient of 0.76 between the average forecast and the year-end price of the S&P 500 index for the given

*Lawrence Berkeley National Laboratory (retired), 1 Cyclotron Road, Berkeley, CA 94720, USA, and University of California, Davis, Department of Computer Science. E-mail: david@davidhbailey.com.

†CARMA, University of Newcastle NSW 2308, Australia; deceased on 2 August 2016.

‡CARMA, University of Newcastle NSW 2308, Australia. E-mail: a.salehipour@newcastle.edu.au.

§Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA. E-mail: lopezdeprado@lbl.gov.

year. However, Kaissar also found that while the strategists' forecasts were reasonably close most of the time, they were surprisingly unreliable during major inflection points.

For example, Kaissar found that the strategists *overestimated* the S&P 500's year-end price by 26.2 percent on average during the three recession years 2000 through 2002, yet they *underestimated* the index's level by 10.6 percent for the initial recovery year 2003. A similar phenomenon was seen in 2008, when strategists *overestimated* the S&P 500's year-end level by a whopping 64.3 percent in 2008, but then *underestimated* the index by 10.9 percent for the first half of 2009. In other words, as Kaissar lamented, "the forecasts were least useful when they mattered most" [**Kaissar**].

There are numerous challenges to assessing the predictions of forecasters, not the least of which is collecting and assessing these predictions. One promising attempt was in a 2012 study by the CXO Advisory Group of Manassas, Virgina, who ranked 68 forecasters based on their 6,582 forecasts during 1998–2005 for the period of 2005–2012 [**CXO1**]. Although that study did not provide full details of its grading, ranking and metric methodology, it acknowledged some weaknesses: (a) the rankings were all weighted equally, or, in other words, all predictions and forecasts were considered equally significant; and (b) the analysis was not adjusted based on the number of forecasts made by a particular forecaster — some experts made only a handful of predictions, while others made many; weighting these the same may lead to distortions when their forecasting records are compared.

In this study, we propose to investigate in greater detail how market experts and forecasters can be graded and ranked, and then to develop and initially deploy an alternative and comprehensive methodology. We build on the experience of others who have collected lists of forecasters, notably the CXO Advisory Group study [**CXO1**, **CXO2**]. Most of these collections are based on the frequency in which the investors or readers have referenced a particular forecaster. In particular, we will seek answers to the following questions:

- How do we recognize and prioritize predictions and forecasts? For instance, we may find different weights for short- and long-term forecasts, or for importance by a given criteria.
- What metrics and measures are most effective and meaningful?

For this study, we will focus on forecasts made for the S&P 500 index, mainly because this is the basis for the similar studies and hence it provides the same basis for comparison purposes. However, the developed methodology is a general one that is applicable to any index for which comprehensive data and forecasts are available.

2 Methodology

Our methodology has two parts. In the first part, every forecast or comment of every market forecaster is evaluated. This is performed by calculating the return of the S&P 500 index over four periods of time. Typically those four periods are one month, three months, six months, and 12 months. Then the correctness of the forecast, i.e. whether the forecaster has made a true or false forecast, is determined in accordance with the time frame for which the forecast is made, considering the correctness of other forecasts that are supposed to occur before or after the forecast. This part is similar to the methodology used in the study by the CXO Advisory team, and for this part, we directly use their evaluation [**CXO1**, **CXO2**].

In the second part, we treat each individual forecast according to two factors: the time frame of the forecast, and its importance/specificity. This is because not all forecasts are equally important. For example, a forecast referring to the next few weeks should be treated differently than the one referring to the next few months; in particular, long-term forecasts should be treated as more significant than the short-term forecasts. After all, in the short-term anything could happen, as a matter of randomness, but in the long-term underlying trends, if any, tend

to overcome short-term noise. For these reasons, we give more weight to longer-term forecasts, since they imply investing skill with greater confidence. In this regard our study contrasts to the study of CXO Advisory team, which treated every forecast as equally significant.

In this study, we consider four time frames, which are weighted as follows:

- Up to one month: 0.25;
- Up to three months: 0.50;
- Up to nine months: 0.75;
- Beyond nine months (up to two to three years): 1.00.
- If the forecast does not include a time frame, or unless there is an impression stating otherwise, we assign a weight of 0.25.

The parameter $w_t \in \{0.25, 0.50, 0.75, 1.00\}$ denotes the weight associated with these time frame.

Regarding the specificity of a forecast, we assign a weight of either 0.5, for a less specific forecast, or 1.0, for a more specific forecast. For example, a forecast that states "the market will be volatile in the next few days" is not a very specific forecast, because the investor may not be able to make a decision solely based on the forecast. However, the forecast "the market will experience a correction" is more specific, and hence, important. In this example, we assign a weight of 0.5 to forecasts of the first sort, and a weight of 1.0 to forecasts of the second sort. Again, in this regard our study contrasts with the earlier study by the CXO Advisory team, which did not introduce or assign specificity weightings. We use $w_s \in \{0.50, 1.00\}$ to denote specificity of a forecast.

Following definition of w_t and w_s, we may derive a weight for a forecast by multiplying those two weights:

$$w_i^+ = w_t \times w_s \quad \text{if forecast } i \text{ is correct} \tag{1}$$

$$w_i^- = w_t \times w_s \quad \text{if forecast } i \text{ is not correct} \tag{2}$$

Notice that w_i^+ is the combined weight for forecast i when it is true, and w_i^- is when it is false. Then, accuracy of a forecaster may be obtained by Equation (3).

$$\epsilon_j = \frac{\Sigma_{i=1}^{n_j} w_i^+}{\Sigma_{i=1}^{n_j} w_i^+ + \Sigma_{i=1}^{n_j} w_i^-}, \tag{3}$$

where j is the forecaster's index, and n_j is the total number of forecasts made by forecaster j.

Dataset

In this study, we utilize the same dataset that was previously compiled by CXO. This dataset includes 68 separate spreadsheets, each of which refers to the data of one forecaster. The information for each forecaster consists a set of forecast statements (text), the returns of the S&P 500 index and the correctness of forecast as evaluated by CXO [**CXO1**, **CXO2**].

Algorithm

To apply our ranking methodology to the dataset, we have developed a program in the programming language Python 2.7. The program reads every sheet in the dataset, evaluates the texts (forecast statements) by assigning appropriate weightings, performs the calculations, i.e. eqs. (1) to (3), and generates two outputs and saves them as two spreadsheet files. The first spreadsheet file has 68 sheets (same as the input dataset), and in addition to the original data

includes the detailed outcomes of the analyses, with rankings. The second spreadsheet includes the ranking summary for all forecasters, that is, the ranking of all 68 forecasters.

To ensure an appropriate assignment of weights to every forecast, the program has two sets of keywords. The first set includes four subsets of keywords, each of which is associated with one time frame. Each subset includes a set of words and time adverbs that represent a specific time frame. For example, the word "soon" is one keyword, which represents a very short-term time frame. The second set of keywords includes words, adjectives, and adverbs that reflect the importance and specificity of the forecasts. The algorithm analyzes every forecast by reading the associated text strings, applies both sets of keywords to find any match, and then assigns weights accordingly. A default weight of 0.25 and/or 0.5 will be assigned to a forecast if there is no matching with respect to the time frame and/or specificity.

Training the algorithm

It is obvious that the performance of the algorithm heavily depends on those two sets of keywords. For this reason, we consider a set of 14 forecasters (about 20%) as the training dataset. More precisely, we manually analyze and evaluate every forecast in the training set. Then we apply formulas (1) through (3) to calculate the accuracy of the forecasts. Given the accuracy of the forecasters in the training set, we evaluate the performance of our algorithm. To do so, we apply the algorithm to the training dataset, and compare the forecasters' accuracy obtained by the algorithm against the one obtained manually. This comparison allows for tuning the algorithm, because we can update the original sets of keywords by adding new keywords that are not already in the sets.

Testing the algorithm

After tuning the algorithm, we applied it to the remaining 54 forecasters in the dataset, which we call the testing dataset. The results of this stage along with the outcomes of the algorithm on the training dataset (in total analyzing 68 forecasters) may be represented as the evaluation and ranking of market forecasters by our developed methodology. This is discussed in more detail in Section 3.

3 Results

After training our algorithm on the training dataset, we ran it on the entire dataset in order to derive the ranking of each market forecaster. We presented the outcomes and findings in the following sections. Notice that the accuracy of the algorithm over the training dataset has been observed to be 92.16%; in other words, the error of the algorithm on the training dataset is 7.84%.

To calculate the accuracy of the algorithm, we manually derived the accuracy of every forecaster in the training dataset. Then we ran the algorithm, which automatically calculates the accuracy of each forecaster, on the same dataset. Let ϵ_j^* denotes the manually obtained accuracy of forecaster j, and ϵ_j the one obtained by the algorithm. Then, the error of the algorithm in calculating the accuracy of forecaster j is

$$\frac{|\epsilon_j - \epsilon_j^*|}{\epsilon_j^*} \times 100$$

The algorithm's average error over all forecasters in the training dataset can easily be calculated by averaging all errors in the training dataset.

3.1 Forecaster accuracy

Figure 1 shows the accuracy of each of the 68 forecasters analyzed by the algorithm. Because not every forecaster has made an equal number of forecasts, the figure shows the accuracy per forecast, and forecast share. For forecaster j, accuracy per forecast is obtained by dividing its accuracy (which is obtained by the algorithm) by its number of forecasts, i.e. n_j. That is,

$$e_j = \frac{\epsilon_j}{n_j} \tag{4}$$

The forecast share of forecaster j, i.e. s_j can be derived by Equation (5).

$$s_j = \frac{n_j}{\Sigma_j n_j} \times 100 \tag{5}$$

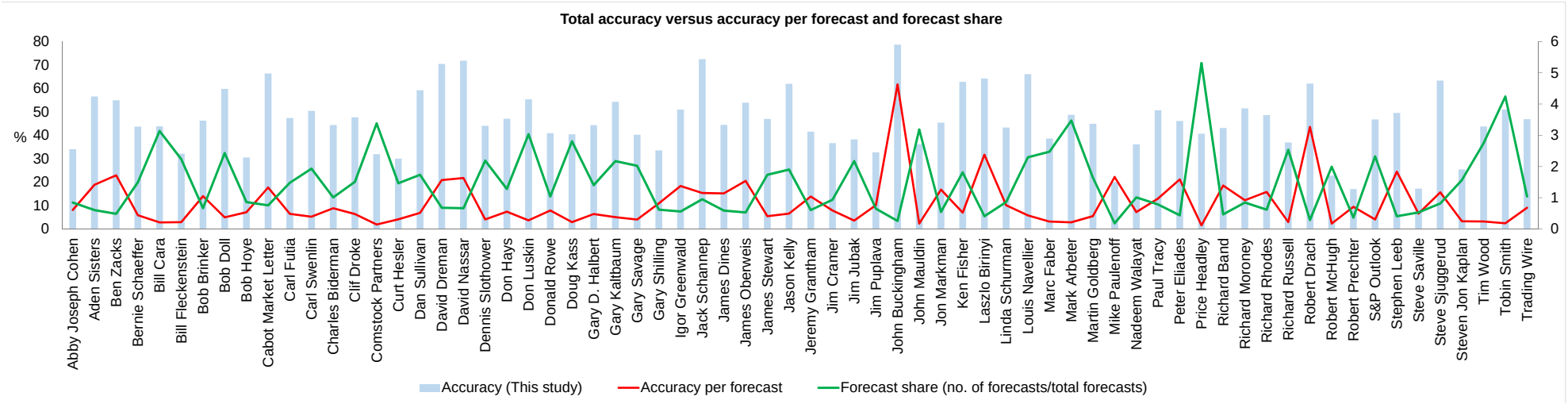

Figure 1: Accuracy of each forecasters (on the left axis) versus accuracy per forecast and forecast share (on the right axis). For forecaster j, the accuracy per forecast is obtained by dividing the accuracy by the number of forecasts (n_j), and forecast share is obtained by dividing the number of forecasts by the total number of forecasts (by all forecasters).

Figure 1 analyzes forecasters' performance along their contribution into the forecasting process. The left axis denotes the values of accuracy, and the right axis denotes the values of accuracy per forecast and forecast share. The reader may admire the statistic e_j (accuracy per forecast) in assessing the performance of forecaster j.

Finally, we compared the accuracy of forecasters obtained by our method against that of published previously in the study of CXO Advisory team (Benchmark). This is graphically depicted in Figure 2.

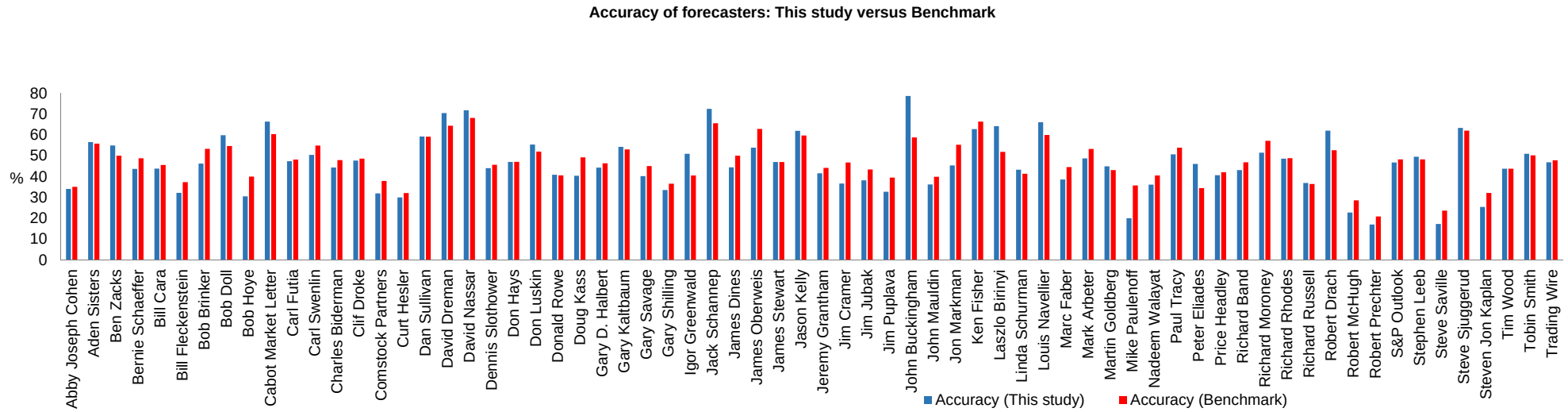

Figure 2: Comparing accuracy of forecasters obtained by our method (this study) against the accuracy obtained by the study of CXO Advisory team (Benchmark). The values of accuracy are in percent.

To have a better grasp of changes in the forecasters accuracy obtained by our method in this study, compared to the earlier study of CXO Advisory team (Benchmark), we define the accuracy gap, which is the difference in values of accuracy between two studies. Let Δ_j denotes the accuracy gap of forecaster j. Equation (6) shows how Δ_j may be derived.

$$\Delta_j = \epsilon_j - \epsilon'_j, \tag{6}$$

where ϵ_j is the value of accuracy for forecaster j, which is obtained by our method, and ϵ'_j is the value of accuracy for forecaster j reported in the study of CXO Advisory team. Gap scores Equation (6) have either positive or negative values. Positive values of gap reflect improvement in the accuracy over the benchmark study, and negative values reflect decreased accuracy. We analyzed the accuracy gap of all forecasters, and illustrated this in Figure 3. Later we report the values of accuracy gap for each forecaster in Table 1. According to the figure, most forecasters have lower accuracy scores with our methodology; in particular, only 36.76% of the forecasters have improved accuracy, and the remaining have lower accuracy. This may be due to the inclusion of additional information of the forecasts' time frames and specificity in our method.

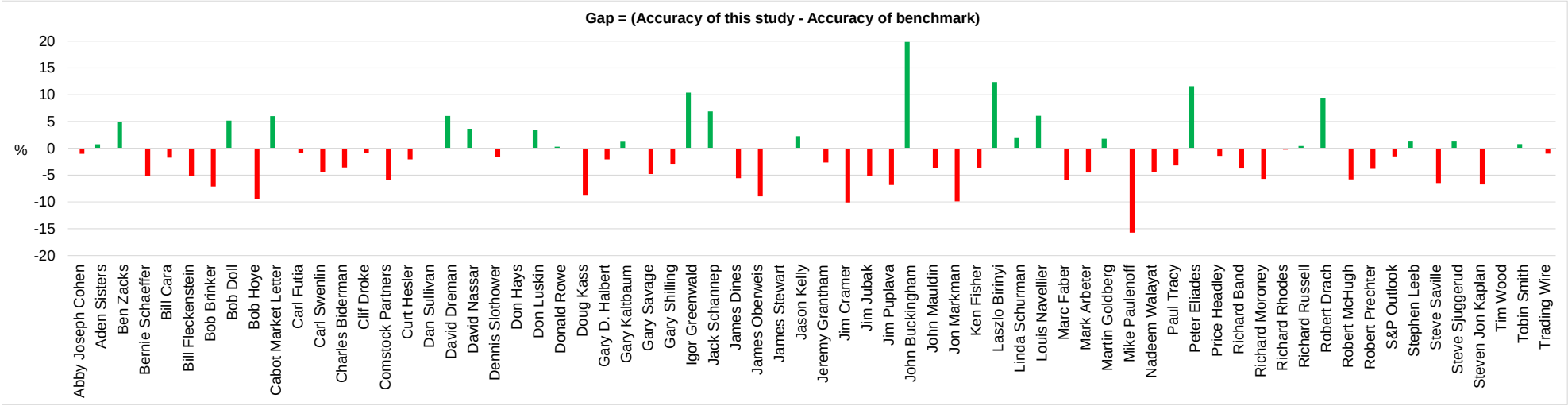

Figure 3: The accuracy gap between our method (this study) and that of the CXO Advisory team (Benchmark) for all forecasters. The accuracy gap Δ_i for forecaster j can be calculated by Equation (6). Positive values of gap reflect higher accuracy, compared with the benchmark, and negative values reflect lower accuracy. As the figure shows, the majority of forecasters have lower accuracy scores.

The histogram of forecasters accuracy may reveal additional information about the behavior of our methodology (this study) versus that of the benchmark. This is illustrated in Figure 4.

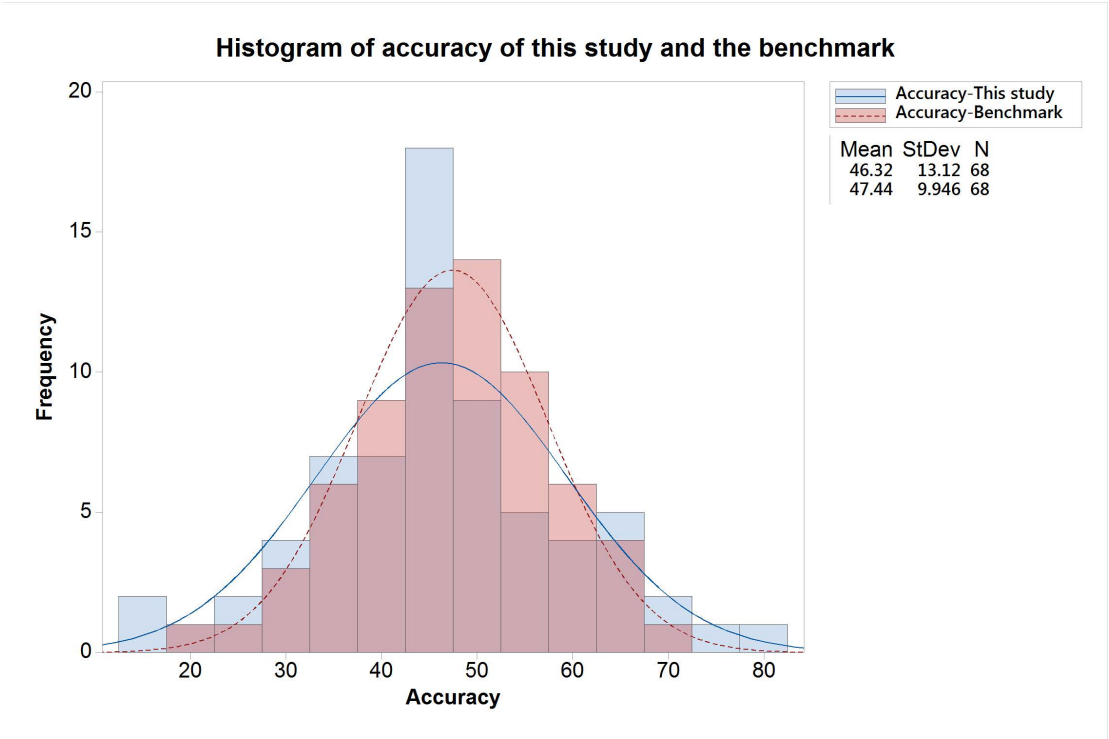

Figure 4: The histograms of forecasters accuracy of this study and the benchmark. According to the histograms, the accuracy obtained by our methodology has a greater standard deviation ("StDev") than the benchmark, which is consistent with the rest of findings reported in this study.

In addition to this, we also analyzed the distribution of forecasters over the accuracy intervals. These were separately calculated for our method (this study) and for the study of CXO Advisory team (Benchmark), and are illustrated in Figure 5. According to the calculated values for accuracy, we considered seven intervals for the values of accuracy, and then calculated the percentage of forecasters that have their accuracy located in an interval. Those seven intervals are:

- $[10, 20)$
- $[20, 30)$
- $[30, 40)$
- $[40, 50)$
- $[50, 60)$
- $[60, 70)$
- $[70, 80)$

There are several points of interest in this data. First, in both studies about 40% of the forecasters have an accuracy score between 40% and 50%. Second, our method identifies two new intervals for accuracy values: a low accuracy interval with ranges for accuracy values

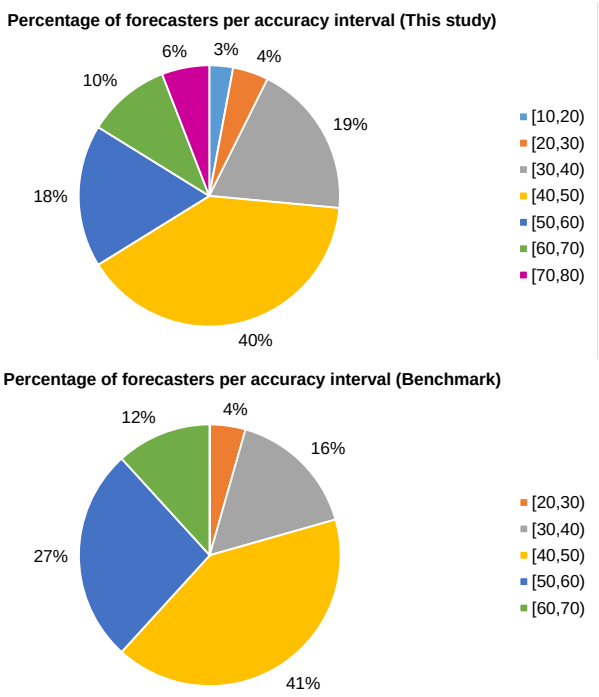

Figure 5: Analyzing the distribution of forecasters over the accuracy intervals. Seven intervals were considered for the values of accuracy, and then percentage of forecasters in every interval was calculated. The figure on the top shows this distribution for our method (this study); the figure on the bottom shows that for the study of CXO Advisory team (Benchmark). In particular, notice that our method grouped the forecasters into seven intervals, while the benchmark study grouped them into five intervals.

between 10% to 20%, in which 3% of the forecasters are located, and a high accuracy interval with ranges for accuracy values between 70% to 80%, in which 6% of the forecasters are located. Third, while the percentage of forecasters in the accuracy interval $[50\%, 60\%)$ has dropped by about 9% (from 27% in the study of CXO Advisory team to 19% in this study), the percentage of the interval $[30\%, 40\%)$ has increased by 3%. This implies that our method assigns fewer forecasters in the accuracy interval of 50% to 60%, and assigns more forecasters to the interval $[30\%, 40\%)$.

3.2 Time frame and specificity analysis

Earlier we discussed the importance of time frame and specificity in forecast statements. It is more difficult to forecast the market's long-term behavior than its short-term behavior, and a specific forecast is more valuable than a non-specific one.

Let us start by investigating time frames distribution of a forecaster. Recall that every forecast may be categorized into one of the four time windows. Hence, for forecaster j, we count the number of forecasts corresponding to each time window, and divide this value by the total number of forecasts of forecaster j. This produces up to four percentage values per forecaster, each for one time window. If we continue this for all forecasters, we obtain the graph of Figure 6.

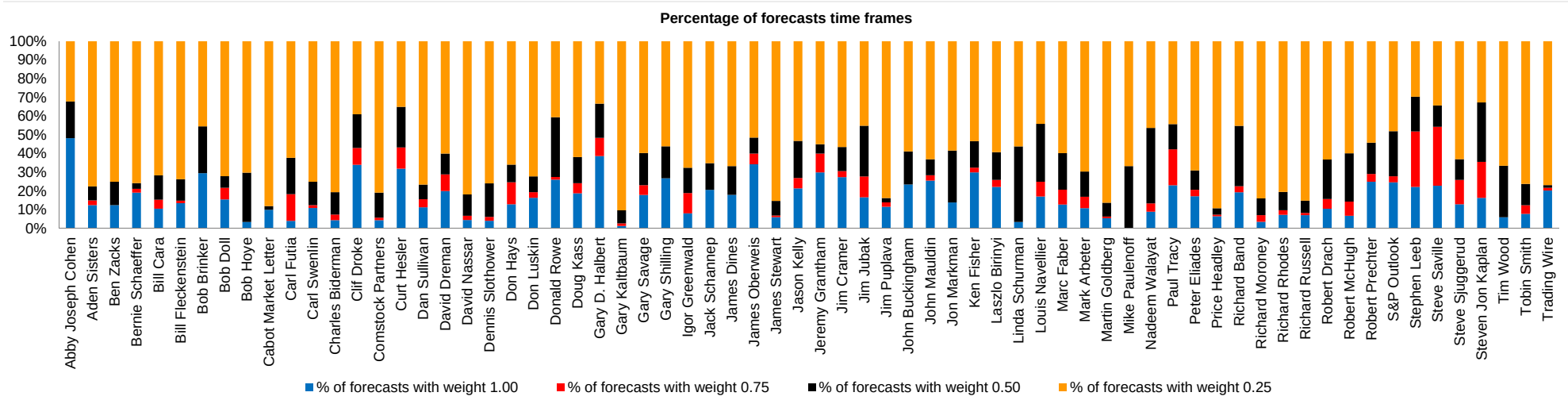

Figure 6: Time frames distribution of forecasts per forecaster. As the graph reveals the majority of forecast statements were made either over a short-term period, i.e. up to a few weeks, or without a specific time frame (associated with a weight of 0.25). Other forecasts were stated covering a long-term period, beyond nine months (associated with a weight of 1.00). Still other forecasts predicted events over a time period between three to nine months (those associated with a weight of 0.75).

A similar analysis can also be performed for those accurate forecasts, that is those turned out to be "correct" forecasts. This is illustrated in Figure 7, which shows the time frames distribution of a forecaster, and only over correct forecasts. In total, only 48% of all forecasts were correct. In this evaluation, we excluded incorrect forecasts, and considered the remaining (both correct or neutral) as correct forecasts.

With respect to the observation that only 48% of all forecasts were correct, it seems that the forecasts were stated at levels not significantly different than chance. Therefore, we performed the Wilcoxon Signed Rank test in order to test whether the occurrences of correct and incorrect forecasts are due to randomness rather than the forecasters' skill. The Wilcoxon Signed Rank is a nonparametric and a distribution-free test for the population median where the test statistic is based on counts of positive and negative values.

We calculate the Wilcoxon Signed Rank test statistic as $d_j = y_j - x_j$, where y_j is the number of correct forecasts of the j-th forecaster, and x_j is the number of incorrect forecasts of the j-th forecaster. We tested $H_0 : M = 0$ versus $H_1 : M \neq 0$, where M is the population median of the test statistic. We assumed a 95% confidence level, and we used the statistical software Minitab version 17.2.1 [**Minitab**] to execute the test.

The test resulted in a p-value of 0.185, which implies that the number of correct forecasts is just as likely as the number of incorrect forecasts. Therefore, it is very difficult to tell if there is any skill present, and it seems that outcomes are due to randomness.

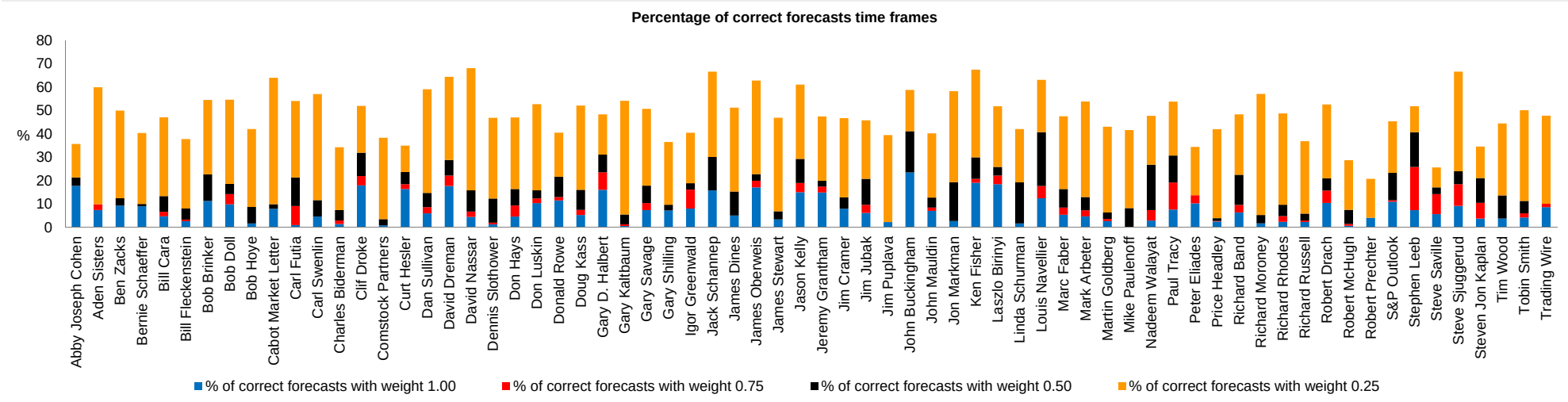

Figure 7: Time frames distribution of correct forecasts per forecaster. A similar behavior to that of Figure 6 is observed here: the majority of correct forecast statements were made over a short-term period (associated with a weight of 0.25) followed by a long-term period (associated with a weight of 1.00).

The time frame distribution of all forecast statements is shown in Figure 8. The graph on the left is over all forecasts, and the graph on the right is over all correct forecasts. Note that the majority of the correct forecasts (around 67.56%) were stated within a short-term period; another 28% of the correct forecasts cover periods between one and three months, and for more than nine months. Only less than 5% of the correct forecasts predicted periods between three to nine months.

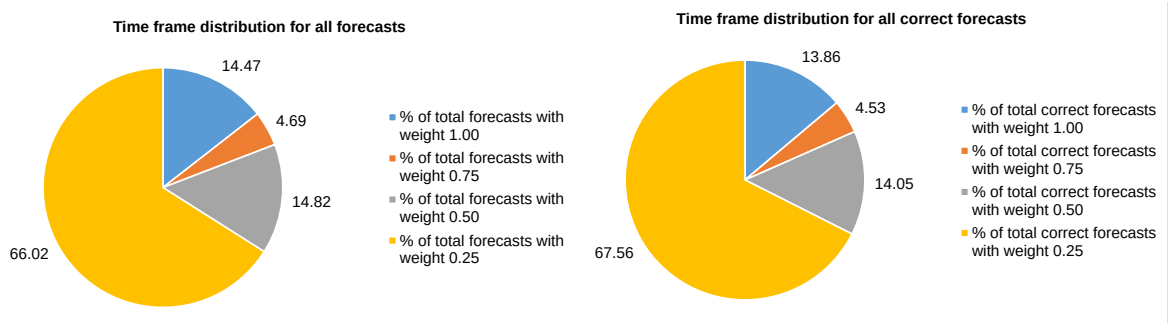

Figure 8: Distribution of the forecasting time frame over all forecasts (figure on the left) and over all correct forecast statements (figure on the right). As the figures show the majority of forecasts are stated over a short-term time frame.

In addition to the time frame distribution, we analyze specificity of the forecast statements. The majority of the forecasts made by forecasters were fairly specific. This is depicted in Figure 9. Approximately 84% of the forecasts are specific, and only a small percentage (around 16%) are vague and non-specific (see Figure 10). Recall that in this study the major criterion of a forecast specificity is whether the investor can solely make a decision by that forecast.

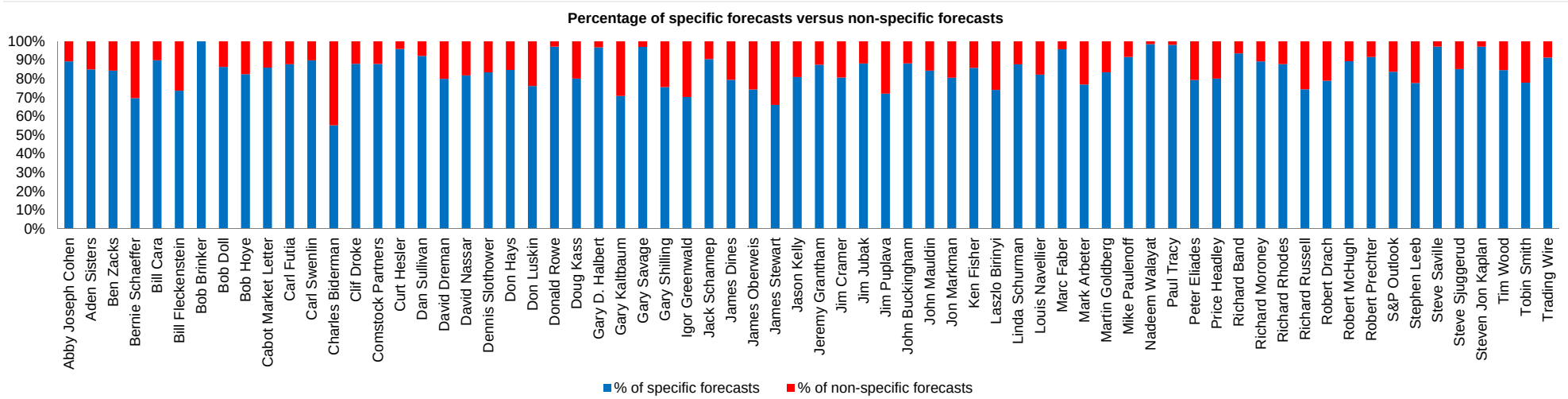

Figure 9: Distribution of specific versus non-specific forecasts per forecaster. The graph reveals that the majority of forecast statements are specific. This observation can almost be concluded for every forecaster.

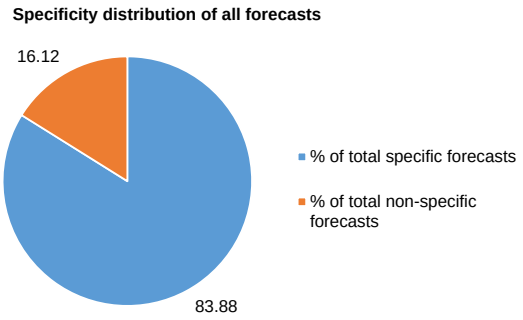

Figure 10: Distribution of the forecasting specificity over all forecast statements. According to the figure the majority of forecasts are specific enough to assist an investor in making decisions.

3.3 Ranking the forecasters

In this section we report the ranking of the market forecasters as resulted by implementing our method. This is fully reported in Table 1. The forecasters in Table 1 were ranked on the basis of their accuracy obtained by our method (this study). For comparison purposes, we reported the accuracy of each forecaster as reported in the study of CXO Advisory team (Benchmark). Also, the values of accuracy gap, which were discussed in Equation (6) are reported here. A positive value of accuracy gap means the forecaster's accuracy is improved over the benchmark, and a negative value means the accuracy has decreased.

Table 1: Comparison of rankings of the 68 forecasters obtained by our method (this study), and that of the CXO Advisory team (Benchmark). The forecasters were sorted by their values of accuracy (rankings) obtained by our method. The values for accuracy are in % (out of 100), and state the accuracy of every forecaster in predicting the market. For the comparison purposes, we also reported the ranking of the benchmark study. The last column of the table reports the values of accuracy gap (see Equation (6)).

Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Bench-mark)	Gap	Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Bench-mark)	Gap
John Buckingham	17	78.69	1	11	19.87	Jon Markman	36	45.37	35	14	-9.89
Jack Schannep	63	72.51	2	3	6.89	Martin Goldberg	109	44.92	36	48	1.80
David Nassar	44	71.84	3	1	3.66	James Dines	39	44.44	37	25	-5.56
David Dreman	45	70.47	4	4	6.03	Charles Biderman	67	44.35	38	34	-3.57
Cabot Market Letter	50	66.39	5	7	6.01	Gary D. Halbert	93	44.32	39	40	-2.07
Louis Navellier	152	66.09	6	8	6.09	Dennis Slothower	145	44.03	40	41	-1.61
Laszlo Birinyi	27	64.21	7	23	12.36	Bill Cara	208	43.84	41	42	-1.74
Steve Sjuggerud	54	63.35	8	6	1.28	Tim Wood	182	43.78	42	46	0.00
Ken Fisher	120	62.80	9	2	-3.59	Bernie Schaeffer	99	43.68	43	29	-5.10
Robert Drach	19	62.07	10	21	9.44	Linda Schurman	57	43.29	44	50	1.91
Jason Kelly	126	61.96	11	9	2.27	Richard Band	31	43.10	45	38	-3.78
Bob Doll	161	59.84	12	16	5.18	Jeremy Grantham	40	41.55	46	45	-2.64
Dan Sullivan	115	59.23	13	10	0.10	Donald Rowe	69	40.89	47	51	0.31
Aden Sisters	40	56.57	14	13	0.76	Price Headley	352	40.65	48	49	-1.40
Don Luskin	201	55.35	15	22	3.39	Doug Kass	186	40.41	49	27	-8.83
Ben Zacks	32	54.95	16	26	4.95	Gary Savage	134	40.24	50	43	-4.79
Gary Kaltbaum	144	54.29	17	20	1.23	Marc Faber	164	38.60	51	44	-5.97
James Oberweis	35	53.90	18	5	-8.96	Jim Jubak	144	38.22	52	47	-5.20
Richard Moroney	56	51.47	19	12	-5.67	Richard Russell	168	36.91	53	60	0.44
Tobin Smith	281	50.96	20	24	0.78	Jim Cramer	62	36.68	54	39	-10.09
Igor Greenwald	37	50.96	21	52	10.42	John Mauldin	211	36.19	55	55	-3.72
Paul Tracy	52	50.66	22	17	-3.19	Nadeem Walayat	67	36.13	56	53	-4.38
Carl Swenlin	128	50.42	23	15	-4.47	Abby Joseph Cohen	56	34.06	57	62	-1.03
Stephen Leeb	27	49.54	24	31	1.26	Gary Shilling	41	33.56	58	59	-3.03
Mark Arbeter	230	48.75	25	19	-4.50	Jim Puplava	43	32.71	59	56	-6.82
Richard Rhodes	41	48.60	26	28	-0.24	Bill Fleckenstein	148	32.17	60	58	-5.16
Clif Droke	100	47.70	27	30	-0.90	Comstock Partners	224	31.93	61	57	-5.96
Carl Futia	98	47.39	28	33	-0.79	Bob Hoye	57	30.53	62	54	-9.47
Don Hays	85	47.04	29	36	-0.02	Curt Hesler	97	30.02	63	65	-2.06
James Stewart	115	46.99	30	37	0.03	Steven Jon Kaplan	104	25.42	64	64	-6.72
Trading Wire	69	46.85	31	35	-0.98	Robert McHugh	132	22.77	65	66	-5.80
S&P Outlook	154	46.76	32	32	-1.52	Mike Paulenoff	12	20.00	66	61	-15.71
Bob Brinker	44	46.24	33	18	-7.09	Steve Saville	35	17.22	67	67	-6.46
Peter Eliades	29	46.07	34	63	11.59	Robert Prechter	24	17.02	68	68	-3.81

We further checked the best and worst forecasters in each of the two studies. In checking the top forecasters in each study, we observe that both share a set of 13 forecasters, so we further analyzed the performance of these 13 forecasters, and shown in Figure 11.

The figure illustrates the percentage of correct forecasts per time frame, and the percentage of correct specific and non-specific forecasts. Also, we included the percentage of total correct forecasts. According to the plot, the number of long-term and specific forecasts that were correctly predicted impact accuracy and ranking the most. For example, "John Buckingham" has a rank of 1 in our study and 11 in the benchmark study, and "David Nassar" has a rank of 3 in our study and 1 in the benchmark study. However, the majority of David's correct forecasts cover periods less than one month, whereas John's correct forecasts mainly cover long-term and middle-term periods. Moreover, John has more correct specific and less correct non-specific forecasts.

On the other hand, if we only consider the number of correct forecast statements in order to evaluate forecasters' performance, David's accuracy would be approximately 70%, while John's would be approximately 60%, thus ranking David before John.

In checking the worst forecasters in each of the two studies, we observe that both share a set of 14 forecasters. We further analyzed the performance of these 14 forecasters, and similar to the analysis we performed for the best forecasters. This is shown in Figure 12. In this regard, the major focus of almost all forecasters has been on short-term forecasts, where the number of correct long-term forecasts constitutes only a small fraction of their total correct forecasts. Moreover, we observe that the percentage of correct specific forecasts has largely been decreased (from the ranges of $[40, 60]$ for the top 10 forecasters to about $[25, 35]$).

Here, three forecasters "Peter Eliades", "Abby Joseph Cohen", and "Curt Hesler" have more than 10% of their long-term forecasts correct, however, they have quite different ranks. After further investigation we realized that all three have very close correctness ratios (around 35%), and "Peter Eliades" has higher ratios for both correct long-term forecasts (to all long-term forecasts), and correct specific (to all specific forecasts). Between "Abby Joseph Cohen" and "Curt Hesler", notice that "Abby Joseph Cohen" has two more advantages: the forecaster has higher percentages of correct long-term and short-term forecasts.

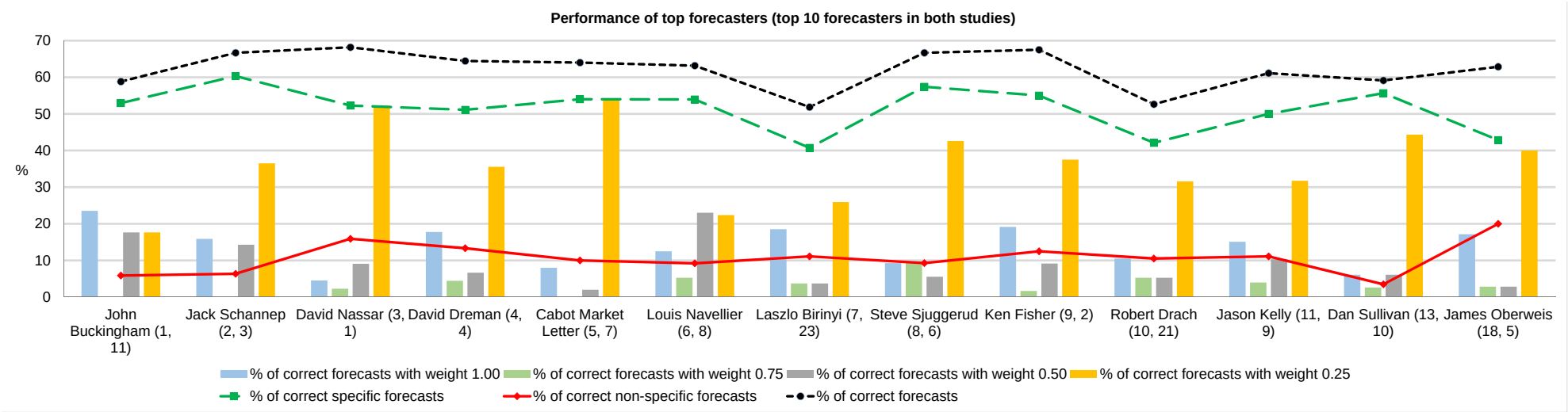

Figure 11: Analyzing performance of top 10 forecasters in each study (in total 13 forecasters were further studied). The graph analyzes the percentage of correct forecasts per time frame, as well the percentage of correct specific and non-specific forecasts. In addition to those, the percentage of total correct forecasts is plotted. According to the plot, the number of long-term and specific forecasts that were correctly predicted impact accuracy and ranking the most. The numbers inside parenthesis next to each forecaster's name (on the horizontal axis) state the forecaster rank obtained by this study, and by the benchmark.

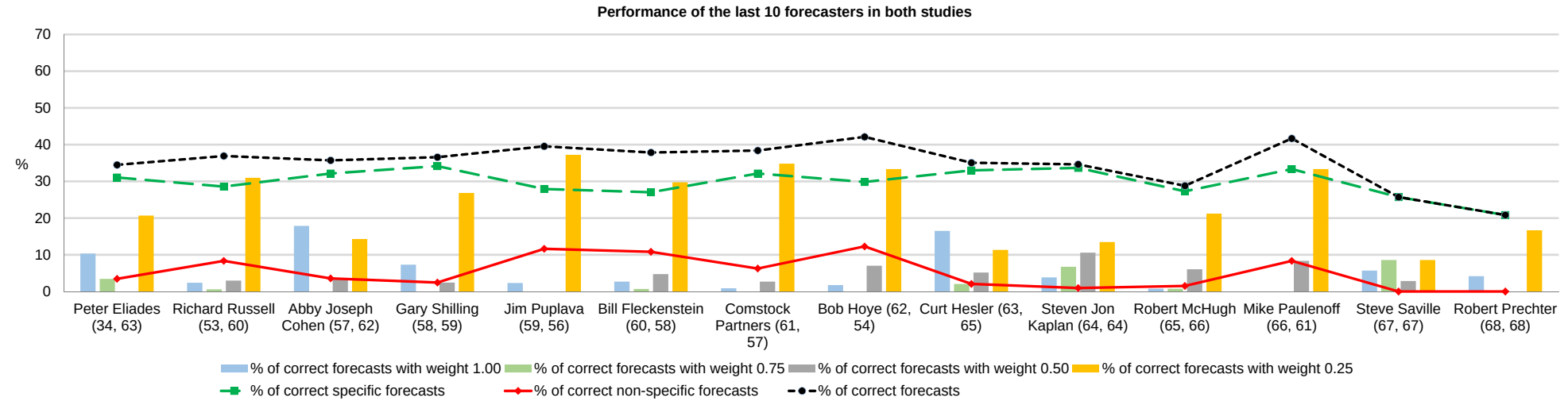

Figure 12: Analyzing performance of the last 10 forecasters in each study (in total 14 forecasters were further studied). The graph analyzes the percentage of correct forecasts per time frame, as well the percentage of correct specific and non-specific forecasts. In addition to those, the percentage of total correct forecasts is plotted. According to the plot, the major focus of almost all 14 forecasters has been on short-term forecasts. Also, we see that the number of both long-term and short-term forecasts impact accuracy and ranking the most. The numbers inside parenthesis next to each forecaster's name (on the horizontal axis) state the forecaster rank obtained by this study, and by the benchmark.

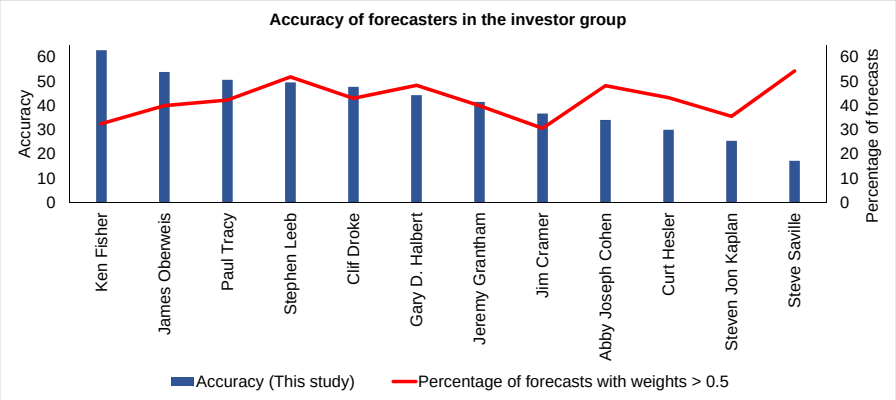

Figure 13: Analyzing performance of the forecasters in the investor group. In total, 12 forecasters were located as investors. While the graph primarily analyzes the accuracy, the percentage of forecasts with weights 0.75 and 1.00 is plotted. According to the plot, the top forecaster of investor group is "Ken Fisher".

3.4 Traders and investors

We may split the forecasters into long-term strategic forecasters ("investors") and short-term tactical forecasters ("traders"). To do so, first we put the forecasts with weights 0.25 and 0.50 into one group ("trading" forecasts, which includes all forecasts with weights less than or equal to 0.50), and those with weights 0.75 and 1.00 into another group ("investing" forecasts, which includes all forecasts with weights greater than 0.50). Then, we analyze the percentage of each group per forecaster. The next step includes deciding which forecaster must be put into trader and which one into investor. For this purpose we define a threshold. If the percentage of investing forecasts of a forecaster is above this threshold, we put the forecaster into "investor" group; otherwise we put the forecaster into "trader" group. In this study, the threshold is 30%. We observed that this value for threshold defines a good trade-off between trading and investing forecasts. Also, we observed that no forecaster has 50% or more of his forecasts with weights greater than 0.50.

With this threshold, we observed that only 12 of forecasters (out of 68) can be considered investors. In other words, 17.65% of forecasters are investors while 82.35% are traders (56 forecasters). In addition to this, we further analyzed the investors and traders by plotting their accuracy. This is illustrated in Figures 13 and 14. According to the figures, "Ken Fisher" is the top forecaster in the investor group, and "John Buckingham" is the top forecaster in the trader group. The figures also plot the percentage of forecasts grouped as "investing" and "trading".

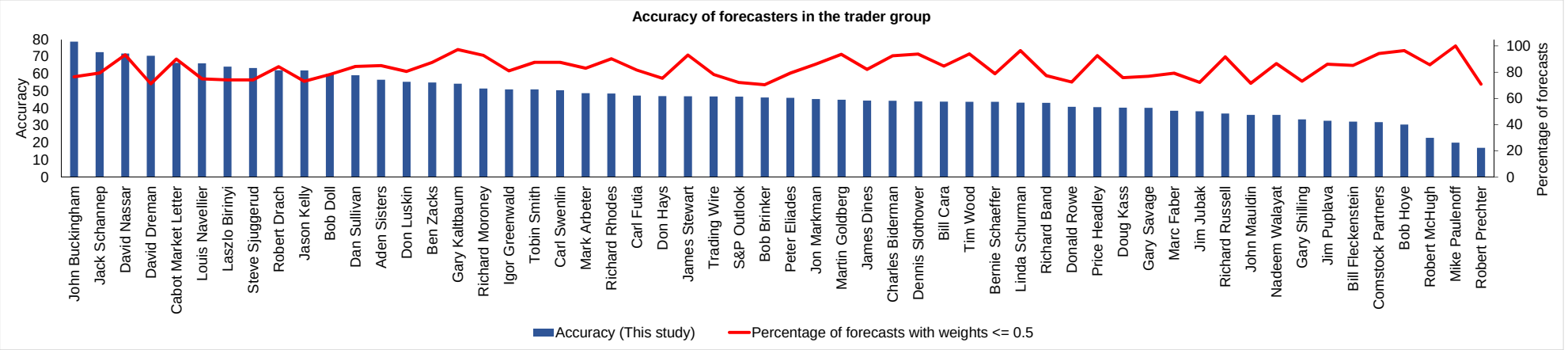

Figure 14: Analyzing performance of the forecasters in the trader group. In total, 54 forecasters were located as traders. According to the plot, the top forecaster of trader group is "John Buckingham".

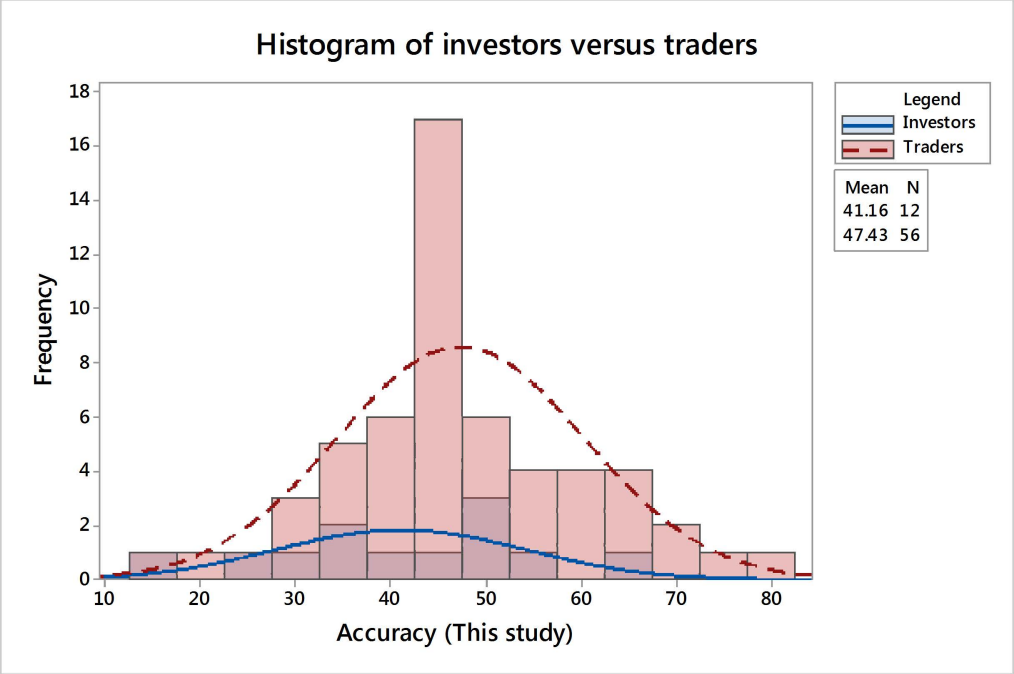

Figure 15: Histogram of accuracy of investors and traders. Notice that the forecasters in the investor group have lower accuracy than those in the trader group.

Finally, Figure 15 illustrates the performance of forecasters in both investor and trader groups. The reader may realize that the forecasters in the investor group have lower accuracy than those in the trader group. We also observed that only 47.38% of the forecasts made by the investors are correct, and 48.09% of the forecasts made by the traders are correct. The latter is perfectly in line with the previous observation that the majority of forecasters perform at levels not significantly different than chance.

4 Conclusion

Market forecasts are widely read in the investment community. Some of these forecasts turn out to be uncannily accurate, while others lead to significant losses. To better understand the extent to which various forecasters have forecasting skill, we have developed a ranking methodology to rank and grade market forecasters. This study builds upon a previous study by the CXO Advisory Group in several directions. In particular, we distinguish forecasts by their specificity, rather than considering all predictions and forecasts equally important, and we also analyze the impact of the number of forecasts made by a particular forecaster.

Across all forecasts, the accuracy is around 48%. Also, the distribution of forecasting accuracy is very similar to the proverbial bell curve implying that the outcomes are due to randomness. This is further acknowledged by the outcomes of the Wilcoxon Signed Rank test. We observed that two-thirds of forecasts predict short-term returns and as far as only a month, and the remaining one-third predict periods over one month. Following the more random nature of short-term returns, this is another argument supporting our findings of random performance of forecasters, and that existence of little skill in doing so. Finally, the highest accuracy value is 78.69%, and while only about 6% of forecasters have their accuracy values between 70% and 79%, the majority of forecasters (two-thirds) have an accuracy level below 50%.

In brief, our findings and results show that some forecasters have done very well, even more so than reflected in earlier studies, but the majority perform at levels not significantly different

than chance, which makes it very difficult to tell if there is any skill present.

Acknowledgment

Our co-author Jonathan M. Borwein sadly passed away in August 2016, while the team was working on this research project. Jon's love of life, pursuit of quality, and extensive knowledge on a wide range of topics have inspired us to extend our reach.

Also, we would like to thank the CXO Advisory team, in particular Steve LeCompte, for his support on providing us with the dataset, and for several discussions during the course of this research.

Finally, we thank the reviewers for their constructive comments on the first draft of the study.

References

[1] Terry Burnham, "Ben Bernanke as Easter bunny: Why the Fed can't prevent the coming crash," *PBS NewsHour*, 11 July 2013, available at `http://www.pbs.org/newshour/making-sense/ben-bernanke-as-easter-bunny-why-the-fed-cant-prevent-the-coming-crash/`.

[2] Terry Burnham, "Why one economist isn't running with the bulls: Dow 5,000 remains closer than you think," *PBS NewsHour*, 21 May 2014, available at `http://www.pbs.org/newshour/making-sense/one-economist-isnt-running-bulls-dow-5000-remains-closer-think/`.

[3] Nir Kaissar, "S&P 500 forecasts: Crystal ball or magic 8?," *Bloomberg News*, 23 December 2016, available at `https://www.bloomberg.com/gadfly/articles/2016-12-23/s-p-500-forecasts-mostly-hit-mark-until-they-matter-most`.

[4] Steve LeCompte, editor, "Guru grades," CXO Advisory Group, 2013, available at `http://www.cxoadvisory.com/gurus/`.

[5] Steve LeCompte, editor, "The most intriguing gurus?," CXO Advisory Group, 2009, available at `http://www.cxoadvisory.com/4025/investing-expertise/the-most-intriguing-gurus/`.

[6] Miles Udland, "Here's what 13 top Wall Street pros are predicting for stocks in 2015," *Business Insider*, 3 January 2015, available at `http://www.businessinsider.com/wall-street-2015-sp-500-forecasts-2015-1`.

[7] Minitab Inc, "Minitab 17 Statistical Software," *www.minitab.com*, 2015, `www.minitab.com`.