Evaluation and Ranking of Market Forecasters

Amir Salehipour

ARC DECRA Fellow, University of Technology Sydney
Joint work with David H. Bailey, Jonathan M. Borwein, and Marcos López de Prado

September 29, 2017
1. Background, motivation, and outcomes

2. Methodology
   - Evaluation algorithm
   - Training and testing the algorithm

3. Results
   - Forecaster accuracy
   - Traders and investors

4. Conclusion
Motivation

1. Many investors rely on market experts and forecasters when making investment decisions.

2. Ranking and grading market forecasters provides investors with metrics on which they may choose forecasters with the best record of accuracy.

Aim of this study

This study develops a novel ranking methodology to rank the market forecaster; in particular,

1. we distinguish forecasts by their time frame, and specificity, rather than considering all forecasts equally important, and

2. we analyze the impact of the number of forecasts made by a particular forecaster.

Outcomes of this study

1. Across a dataset including 6,627 forecasts made by 68 forecasters, the accuracy is around 48% implying the majority of forecasters perform at levels not significantly different than chance.
Forecasts are not reliable enough!

“The forecasts were least useful when they mattered most” [Kaissar].

When analyzing a set of strategists’ predictions from 1999 to 2016, Kaissar found that forecasts were surprisingly unreliable during major inflection points [Kaissar]:

1. The strategists **overestimated** the S&P 500’s year-end price by 26.2% on average during the three recession years 2000 to 2002.

2. They **underestimated** the index’s level by 10.6% for the initial recovery year 2003.

3. They **overestimated** the S&P 500’s year-end level by a whopping 64.3% in 2008, but then **underestimated** the index by 10.9% for the first half of 2009.
In 2012, the CXO Advisory Group ranked 68 forecasters based on their 6,582 forecasts (forecasts made for the S&P 500 index) [CXO1].

That study acknowledged some weaknesses:

- All predictions and forecasts were considered equally significant.

- 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.
Our ranking

We extended and advanced the previous ranking (therefore, we used the same dataset).

In particular, the major contributions of our study are:

- Investigating in greater detail how market forecasters can be graded and ranked;

- developing an alternative and comprehensive ranking methodology;

- recognizing and prioritizing forecasts by considering different weights for short- and long-term forecasts, or for important/specific forecasts; and,

- investigating and analyzing the most effective and meaningful metrics and measures.
Our ranking methodology

Part 1:
- Every forecast statement is evaluated by calculating the return of the S&P 500 index over four periods of one month, three months, nine months, and 12 months.
- The correctness of the forecast is determined in accordance with the time frames.

Part 2
- Each individual forecast is treated according to two factors of time frame and importance/specificity (not all forecasts are equally important).
- Long-term forecasts are treated as more significant than the short-term forecasts (because in the long-term underlying trends, if any, tend to overcome short-term noise; $w_t \in \{0.25, 0.50, 0.75, 1.00\}$).
- Specific forecasts are treated more important than non-specific ones ($w_s \in \{0.50, 1.00\}$).
Forecaster’s score

Combined weight for a forecast:

\[ w_i^+ = w_t \times w_s \quad \text{if forecast } i \text{ is correct} \]

\[ w_i^- = w_t \times w_s \quad \text{if forecast } i \text{ is not correct} \]

Where \( w_i^+ \) implies when the forecast is true, and \( w_i^- \) when it is false.

Then, score (accuracy) of a forecaster may be obtained as:

\[ \epsilon_j = \frac{\sum_{i=1}^{n_j} w_i^+}{\sum_{i=1}^{n_j} w_i^+ + \sum_{i=1}^{n_j} w_i^-} \quad (1) \]

where \( j \) is the forecaster’s index, and \( n_j \) is the total number of forecasts made by forecaster \( j \).
Evaluation algorithm

- We developed an algorithm, and coded that in the programming language Python 2.7.

- The algorithm reads every record in the dataset, processes the forecast statements by assigning weights according to the six sets of pre-defined keywords, and derives the forecasters score.

- Keywords: $K_t = \{\{K_{t_1}\}, \{K_{t_2}\}, \{K_{t_3}\}, \{K_{t_4}\}\}$, and $K_s = \{\{K_{s_1}\}, \{K_{s_2}\}\}$, where $K_t$ includes subsets of keywords associated with time frames, and $K_s$ includes subsets of keywords associated with the importance of the forecasts.

- The algorithm analyzes every forecast by applying both sets of keywords to find any match, and then assigns weights accordingly.
Training and testing

Training the algorithm:

- The performance of the algorithm depends on the keywords; we consider a set of 14 forecasters (about 20%) as the training dataset.

- We manually analyze and evaluate every forecast in the training set, and calculate the score of the forecasts.

- Then we apply the algorithm to the training dataset, and compare the forecasters’ accuracy obtained by the algorithm against the one obtained manually. We update the sets of keywords accordingly.

- After several attempts, we obtained an accuracy of 92.16% for the algorithm’s performance.

Testing the algorithm:

- We apply the algorithm to the remaining 54 forecasters, which we call the testing dataset.
Because not every forecaster has made an equal number of forecasts, we analyzed the accuracy per forecast:

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

(2)

Where, $\epsilon_j$ is obtained by the algorithm, and $n_j$ number of forecasts made by forecaster $j$.

And forecast share of forecaster $j$:

$$s_j = \frac{n_j}{\sum_j n_j} \times 100$$

(3)
Accuracy gap grasps the changes in the forecasters' accuracy between two studies:

- Positive values reflect improvement in the accuracy over the previous study, and negative values reflect decreased accuracy.

In particular, **63.24% of forecasters have lower accuracy.**
Results – Forecasters time frame and specificity

Percentage of forecasts time frames

- % of forecasts with weight 1.00
- % of forecasts with weight 0.75
- % of forecasts with weight 0.50
- % of forecasts with weight 0.25

Percentage of specific forecasts versus non-specific forecasts

- % of specific forecasts
- % of non-specific forecasts
Results – Traders and investors

Long-term forecasters or “investors”, who have at least 30% of the forecasts with weights 0.75 and 1.00, and short-term forecasters or “traders”, with the majority of the forecasts with weights 0.25 and 0.50.

We observed no forecaster has 50% or more of his forecasts with weights greater than 0.50.
Results – Traders and investors

Investors

Accuracy of forecasters in the investor group

Traders

Accuracy of forecasters in the trader group
Results – Summary

Major finding: the majority of forecasters perform at levels not significantly different than chance, which makes it very difficult to tell if there is any skill present.

- Across all forecasts, the accuracy is around 48%.
- Two-thirds of forecasts predict as far as only a month.
- Only one-third of forecasts predict periods over one month.
- Two-thirds of forecasters have an accuracy level below 50%.
- Only about 6% of forecasters have their accuracy values between 70% and 79%; the highest accuracy value is still below 80%.
Publication and awards


- Awarded “Silver Bullet” award (20 May 2017).

- Among the most viewed papers on the SSRN (www.ssrn.com) with a total 2,762 views within six months (28 September 2017).
Bibliography


