FORESIGHT

The International Journal of Applied Forecasting

ARTICLES *by* FORECASTERS *for* FORECASTERS: Q4:2024

Limiting Extreme Behavior in Forecasting Competitions

Join the *Foresight* readership by becoming a member of the International Institute of Forecasters
forecasters.org/foresight/

Limiting Extreme Behavior in Forecasting Competitions

MATTHEW J. SCHNEIDER, JETHRO BROWELL, AND RUFUS RANKIN

PREVIEW *Although the objectives of forecasting competitions vary, the primary goal of competitors is to win. As a result, competitors may take large risks in the hope of earning a top prize. But this is in sharp contrast to most business settings, where extreme behavior is not tolerated. To address this conflict, authors Schneider, Browell, and Rankin provide ideas on restructuring incentives to limit extreme behavior. They propose limiting downside risks and bad forecasts, rewarding long-term performance and statistical significance, and linking forecasts to value-add. In addition, they find that the recent HEFTcom and M6 competitions connected forecast skill to performance in decision making, but could be improved to limit extreme behavior.*

Forecasting competitions are excellent tools for empirically evaluating the performance of forecasting methods. However, the objectives of many competitions have often been unclear (Hyndman 2019). What is clear, though, is that the objective of each competitor is to win the competition.

The goal of winning a competition incentivizes shorter-term, extreme behavior (Witkowski and colleagues, 2023). Forecasters often must do something unique and risky to win, with little consequence for losing. However, in many business settings, the costs of losing far exceed the benefits of winning. This incentive imbalance raises concerns about the long-term stability of winning forecasting methods in practice.

During the recent Hybrid Energy and Forecasting Trading (HEFT) and M6 competitions, competitors used forecasts to trade fictional money. Competitors received prizes if they ranked at the top in revenue from shadow trading in HEFTcom or had a top return/risk ratio or ranking of assets in M6. However, in practice, most asset managers only deliver a small yet positive tracking error relative to their benchmark. The downsides for underperformance are far greater and can include reputational risks, business disruptions, lawsuits, investigations, and monetary losses. As a result, large companies often focus on what not to do, and the failures are usually not revealed to the public.

On the other hand, forecasting competitions reduce survivorship biases by transparently sharing data, competitors, metrics, and results. To inform practice, we believe it is equally important to understand both the winners and the losers. We provide some ideas on what it might take for forecasting competitions to better inform practice.

RECOMMENDATIONS TO IMPROVE FORECASTING COMPETITIONS

Makridakis and colleagues (2022) argue that forecasting competitions are the equivalent of laboratory experimentation used in the physical and life sciences. As such, "[t]hey provide useful, objective information to improve the theory and practice of forecasting, advancing the field, expanding its usage, and enhancing its value to decision and policymakers" (p. 96). Their article describes 10 design attributes to consider when organizing future forecasting competitions. We now augment this list with additional considerations based on our analysis of the HEFTcom and M6 results.

Limit Extreme Behavior

Currently, competition incentives may encourage strategic competitors to exaggerate their forecasts to differentiate themselves (Lichtendahl and Winkler, 2007), risking poor performance to increase their expected rank relative to others. We think each competitor should have some skin in the game to take a measured risk. For example, over 100 competitors competed for $300,000 in prize money in the M6 competition. Most competitors lost money, resulting in a net loss overall. Such a result in business among all trading teams would not be tolerable.

Instead, perhaps a different result would have occurred if 100 total competitors were each given $1,500 to manage and allowed to keep what they did not lose (provided they competed in good faith). The remaining $150,000 plus losses from other competitors could have been allocated as additional prize money for those with improved performance. Prize money distributions could also be determined ex-post using scoring rules or statistical significance thresholds. This prize money would also incentivize more forecasters from developing countries to participate and support the Forecasting for Social Good (***f4sg.org***) efforts.

Lengthen Incentives

A "post-competition" could also last a few years to incentivize longer-term stability. Organizers and competitors could choose to release automated product-ready code at the end of the competition, which would self-run for a few years. Some of the prize money could be allocated for those forecasting methods that do not experience large losses in the future. Such rewards could incentivize competitors to develop risk management heuristics or forecasting methods that adapt quicker (e.g., adaptive smoothing) in changing environments.

The ability to attract competitors from various backgrounds and career stages is related to having skin in the game. Students, companies, researchers, and practitioners could be even more evenly distributed. However, this is a challenging task because more experienced professionals are less likely to make time to compete regularly. For example, the best financial traders make more in a day than the competition prizes. Perhaps such professionals would be willing to act in more of an advisory role for more junior teams.

Key Points

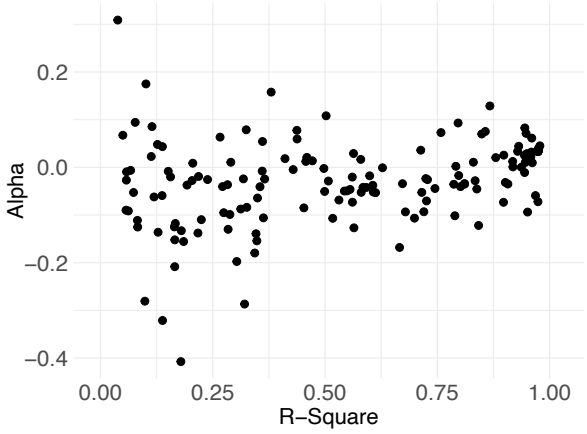
- Forecasting competitions incentivize shorter-term, extreme behavior.
- Unlike businesses, forecasters do not suffer significant consequences for losing.
- With hundreds or thousands of competitors, it is hard to distinguish between skill and randomness. Competition organizers could monitor competitors' performance above low-cost benchmark methods with statistical significance.
- Forecasting competitions should aim to reward forecast value. Sponsors could align monetary incentives in competitions with fiduciary responsibilities in business to limit extreme behavior.

Evaluate Randomness

In the early days of the M competitions, there were only a few competitors with analytically tractable statistical methods. Conclusions were easier to make because the data-generating process and statistical properties of each forecasting method were known. However, as time passed, the barriers to entry for forecasting competitions decreased. In 2020, the M5 competition was released on Kaggle with thousands of competitors, many achieving different results by customizing hyperparameters of similar ML methods. As a result, it has become harder to distinguish between skill and random chance.

It is important to evaluate a competitor's performance conditionally based on the number of competitors and low-cost

Figure 1. Scatter plot of the M6 competitors' alphas against R^2 values.



benchmarks. MASE has been traditionally used to assess whether a forecasting method improves upon a simple benchmark. However, MASE is measured on a single benchmark (the mean absolute error of the in-sample, one-step-ahead naive forecast) and based on forecast accuracy directly. Instead, the randomness of competitors could be assessed by evaluating value-add, conditional on multiple benchmark methods.

In finance, *alpha* (α) is commonly used to measure the value-add above benchmarks by assuming that an investor can freely invest in a linear combination of benchmarks. α is the intercept in a regression where the dependent variable is the focal competitor's returns (minus the risk-free rate), and the independent variables are the benchmark returns (minus the risk-free rate). Since the returns minus the risk-free rate are generally centered around 0%, a competitor has good and independent performance if α is both positive and statistically significant, respectively.

$$Returns_{competitor} \sim \alpha + \beta_1 Returns_{benchmark1} + \beta_2 Returns_{benchmark2} + \varepsilon$$

In the M6 competition, competitors with lower values of R^2 exhibited more extreme performance. Twelve competitors (11 had negative alphas) had a statistically significant alpha when using seven benchmarks, and only five competitors (four had negative alphas) when using the S&P 500 as the only benchmark (Schneider and colleagues, 2024). Although widely used in practice, these tests are overly simple. They assume independent and identically distributed returns and no autocorrelation, which are unlikely to be true. The competitors that did not have extreme alpha (or Sharpe Ratio) values all had R^2 values above 0.50, indicating that winning (or losing) was not likely when over 50% of the variance in a competitor's returns was explained by benchmarks.

Figure 1 is a scatter plot of the M6 competitors' alphas against R^2 values when using seven benchmarks. The figure shows that the value-add is more moderate when the competitors' returns are more similar to benchmark returns (high values of R^2). Conclusions are similar with just one benchmark (S&P 500).

Although the above equation mainly applies to financial returns, it could also be applied to forecasts to determine the similarity of a competitor's forecasts to freely available benchmarks. For example, returns could be replaced with forecasts (the mean is no longer 0). If the R^2 is close to 100%, then the competitor's forecasts may not be unique because they are explainable by a combination of benchmark forecasts.

Alternatively, organizers could calculate a skill score, where the numerator is the accuracy of a benchmark method minus the accuracy of a competitor, and

the denominator is the accuracy of a benchmark method minus the accuracy of a perfect forecasting method.

$$Skill\ Score = \frac{Accuracy_{benchmark} - Accuracy_{competitor}}{Accuracy_{benchmark} - Accuracy_{perfect}}$$

As a competitor's accuracy improves, the skill score increases. Bootstrapping can be used to introduce randomness into skill scores. For example, if a competitor made 100 forecasts, then we can randomly sample 100 forecasts (with replacement) to get an empirical distribution of skill scores. After doing this many times for each competitor, a range of skill scores would emerge. **Figure 2** shows that the first five teams have higher skill scores than the last five teams, after introducing the randomness.

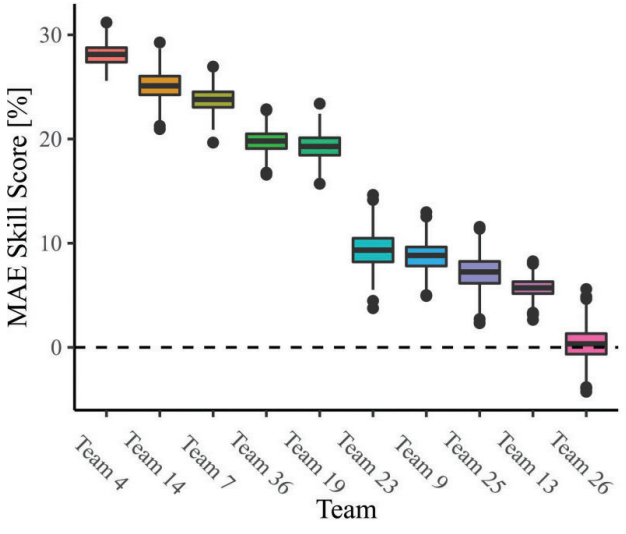
More traditionally, a Diebold-Marino (DM) test could be applied to determine if a competitor's forecasts are more accurate than a benchmark's forecasts. The DM test rejects the null hypothesis that forecast accuracy is the same when the competitor's forecasts have consistently lower errors, averaged across each time period.

Reduce Multiple Testing Biases

Statistical significance (p-values) can be adjusted to test multiple competitors simultaneously using corrections such as the Benjamini-Hochberg (BH) method. The BH method makes it more difficult for a competitor's performance to be statistically significant as the number of competitors increases, as we illustrate here:

> Consider the case of two competitions, one with four competitors and one with 10 competitors. Let's say the top-performing method in each competition had the lowest p-value of 0.01 – i.e., the probability of seeing a result at least that extreme under random chance was 1%. The BH method adjusts this p-value by multiplying it by the number of competitors divided by the rank of a competitor's p-value. This results in an adjusted p-value of 1%×4/1=4% for the competition with four competitors and 1%×10/1=10% for the competition with 10 competitors. When the significance cutoff is 5%, the top-performing method has a significant result only in the competition with four competitors.

Figure 2. Boxplots of competitors' block bootstrapped skill scores when the accuracy metric is Mean Absolute Error (MAE) (Source: Farrokhabadi and colleagues, 2022).



The BH method assumes independence; more conservative alternatives, such as the Holm-Bonferroni (HB) method, use a sequential procedure to decide which competitors have statistically significant results. However, these procedures do not address the problem that two competitors may be using identical methods while both being statistically significant.

How should we evaluate randomness if one competitor using LightGBM wins a competition, but LightGBM was also used by many other competitors with performance across the spectrum? This is what occurred in the M5 competition. The top five winners disclosed their method, with most using LightGBM. But very few of the losers disclosed their method. Teams with names *lightworkgbm* (2,268th place), *LGBMaster* (3,024th place), and *LightGBM* abuser (3,960th place) suggest LightGBM was not always highly successful (***kaggle.com/competitions/m5-forecasting-accuracy/leaderboard***).

In the M5 case, should we downgrade the winners, or compliment them for their improved (human) skill in tuning hyperparameters? A more nuanced understanding of these differences would be helpful in the future. Winners of forecasting competitions often write articles describing how they won, but we should also hear from the losers and the middle of the pack who used methods similar to those of the winners. For example, HEFTcom required all participants to submit reports summarizing their solutions to qualify for a place on the final leaderboard for the organizers to analyze and share learnings.

Analyze Momentum

Momentum means that past winners of forecasting competitions will continue winning in similar environments. For example, the winners of an energy forecasting competition in January in Texas would have momentum if they continued to win in subsequent months in Texas. Generally, momentum implies that a forecasting method has a successful strategy (or representative patterns) in the longer term when applied to similar contexts.

This effect has also been observed between competitions, where successful participants are likely to participate in future competitions with a similar structure, as observed in energy forecasting competitions.

Avoid Bad Forecasts

In clinical drug trials, the FDA uses a sequential test to reject vaccines when a certain number of adverse cases (e.g., deaths) are reached. Every time a new adverse case happens, the pharmaceutical company records whether the adverse case happened in their vaccine group (competitor method) or control group (benchmark method). If a disproportional number of adverse cases occurs in the vaccine group, then the vaccine is rejected. For example, the vaccine is rejected if 7 out of 7 adverse cases came from the vaccine group. However, if only 6 out of 7 came from the vaccine group, the vaccine can continue until a new safety threshold is hit (e.g., 8 out of 9 would reject the vaccine). Once a large number of adverse cases is reached or a low number of total adverse cases are attributable to the vaccine group, the vaccine is accepted (Dragalin and Fedorov, 2006). The theory underlying this process comes from the field of sequential statistics, which ensures that the statistical significance cutoffs sum to a chosen false positive rate (e.g., 5%) across all decision stages.

Similarly, if competition organizers define a bad forecast (adverse case) up front, it may be possible to evaluate forecasting methods based on their ability to avoid bad forecasts (or accuracy at specified time periods). Competition organizers could count the number of times a benchmark forecasting method makes a bad forecast compared to a competitor's forecasting method. For a competitor's method to be accepted as safe for continued use, the benchmark method would need to make significantly more mistakes over time than the competitor's method.

Link Forecasting with Value-Add

The value of forecast improvement is only realized when it leads to improved decision making – resulting in financial or similar gains, or reduced risk. Forecasting

competitions should aim to reward forecast value. This could range from selecting a relevant error metric to simulating a relevant decision-making process. Both M6 and HEFTcom attempted and exposed the correlation, though not perfect correlation, between forecast skill (averaged over the competition period) and performance in decision making. While the incentives created by such a setup need to be carefully considered – see the Limit Extreme Behavior section above – they offer greater opportunities for learning and insights into what may work in practice than forecasting-only competitions.

REFERENCES

Dragalin, V. & Fedorov, V. (2006). Multistage Designs for Vaccine Safety Studies, *Journal of Biopharmaceutical Statistics*, 16(4), 539-553.

Hyndman, R.J. (2020). A Brief History of Forecasting Competitions, *International Journal of Forecasting*, 36(1), 7-14.

Farrokhabadi, M. et al. (2022). Day-ahead Electricity Demand Forecasting Competition: Post-covid Paradigm, IEEE Open Access *Journal of Power and Energy*, 9, 185-191.

Lichtendahl Jr., K.C. & Winkler, R.L. (2007). Probability Elicitation, Scoring Rules, and Competition Among Forecasters, *Management Science*, 53(11), 1745-1755.

Makridakis, S., Fry, C., Petropoulos, F., & Spiliotis, E. (2022). The Future of Forecasting Competitions: Design Attributes and Principles, INFORMS *Journal on Data Science*, 1(1), 96-113.

Schneider, M.J., Rankin, R., Burman, P., & Aue, A. (2024). Benchmarking M6 Competitors: An Analysis of Financial Metrics and Discussion of Incentives. ***arXiv preprint arXiv:2406.19105***

Witkowski, J. et al. (2023). Incentive-compatible Forecasting Competitions, *Management Science*, 69(3), 1354-1374.

Matthew J. Schneider holds a PhD in statistics and is currently an Associate Professor of Business Analytics at the LeBow College of Business at Drexel University. His research focuses on the intersection of data privacy, time series forecasting, and marketing analytics. He has consulted for various financial services, pharmaceutical, energy, and FinTech clients on connecting AI/ML/Stats methodologies to value-add for their business use cases. His first academic paper in the *International Journal of Forecasting* evaluated the results of the M3 competition, and he was the program chair of the International Symposium of Forecasting in Charlottesville in 2023.

mjs624@drexel.edu

Jethro Browell received a PhD in wind-energy systems from the University of Strathclyde (UK) in 2015 and is currently Professor of Statistics and Data Analytics at the University of Glasgow. His research interests span all aspects of data analytics and forecasting with a focus on applications in the energy sector. Jethro has worked extensively with industry in the UK and Europe developing methods for forecasting wind power and electricity demand, among other things. Several utilities are using his forecasts and decision-support tools today. Jethro led the organization of the 2024 Hybrid Renewable Energy Forecasting and Trading Competition.

jethro.browell@glasgow.ac.uk

Rufus Rankin earned a doctorate of business administration (DBA) from Grenoble Ecole de Management in 2013. He has created more than two dozen investment funds and allocated over $2 billion in institutional capital to hedge funds, CTAs, equities, and fixed income. Rufus is currently the Director of Investment Strategy at ProShares. He has published research in multiple venues, including a book on using principal component analysis for portfolio diversification.

rufusrankin@gmail.com

This article originally appeared in *Foresight*, Issue 75 (forecasters.org/foresight) and is made available with permission from *Foresight* and the International Institute of Forecasters.