Superforecasters: A Decade of Stochastic Dominance

Christopher W. Karvetski, PhD
karvetski@goodjudgment.com

1. Introduction

The emergence of "Superforecasters" occurred during the initial multi-year geopolitical forecasting tournaments that were sponsored by the research arm of the US Intelligence Community (IARPA) and ran from 2011 until 2015. These Superforecasters routinely placed in the top 2% of accuracy among their peers and were a winning component of the experimental research program of the Good Judgment Project, one of five teams that competed in the initial tournaments. These elite Superforecasters exceeded the many forecasting accuracy targets set by IARPA, and notably were over 30% more accurate than US intelligence analysts forecasting the same events with access to classified information. While the story of Superforecasters has been chronicled in numerous media and academic publications, it is best told in Tetlock and Gardner's book *Superforecasting: The Art and Science of Prediction*[1].

While the official IARPA geopolitical forecasting tournaments ended in 2015, many from the original cohort of Superforecasters, as well as newly-identified elite forecasters, have continued to make forecasts as professional forecasters within Good Judgment Inc (GJ Inc), the commercial successor to the Good Judgment Project that serves a variety of public and private sector clients. Our goal within this technical whitepaper is to fill the gap between 2015 and 2021 and benchmark the accuracy of Superforecasters versus a large peer group of online forecasters.

2. Analysis

For our comprehensive analysis, we compiled forecasting data over a six-year period on 108 geopolitical forecasting questions that were posted simultaneously on the GJ Inc Superforecaster platform as well as the Good Judgment Open (GJ Open) forecasting platform, an online forecasting platform and community that allows anyone to sign up, make forecasts, and track their accuracy over time and against their peers. The 108 forecasting questions were diverse, spanning elections, economics and policy, COVID-19, and other key geopolitical events. The keywords from the questions are presented in Figure 1.

Figure 1. Phrases that comprised the forecasting questions.

The average time each forecasting question was open was 214 days. In total, as shown in Figure 2, there were 219 (1.5%) Superforecasters versus 14,179 (98.5%) GJ Open forecasters that made at least one forecast on one of the 108 questions. Despite being relatively small in number, the Superforecasters were much more prolific, making 59,756 (35.6%) forecasts on these questions versus 107,866 (64.4%) forecasts for GJ Open forecasters.

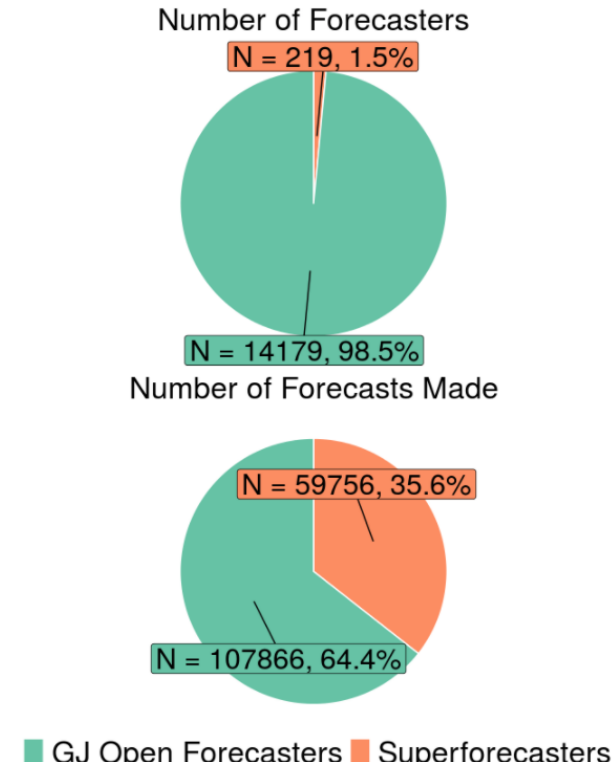

Figure 2. Composition of forecasters and forecasts.

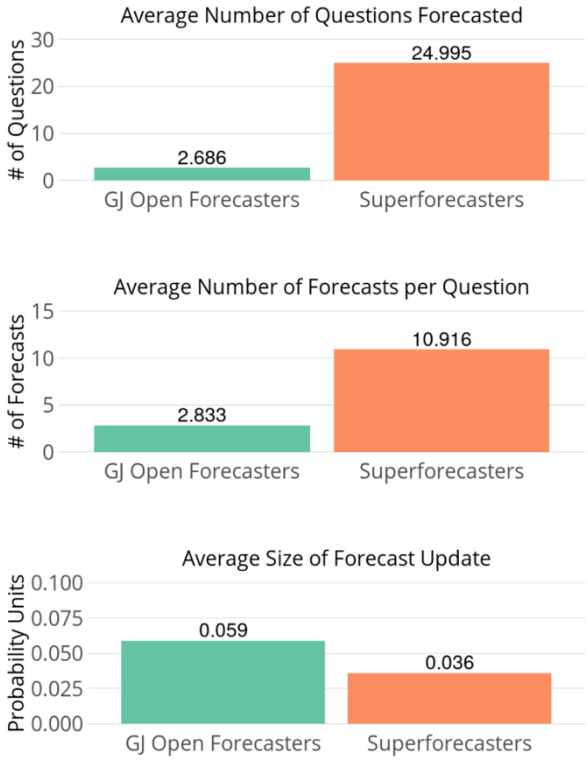

Figure 3. Question level statistics.

As shown in Figure 3, each Superforecaster made, on average, forecasts on 25 of the 108 different questions, versus less than three for each GJ Open forecaster. Consistent with the reporting in Grant's book *Think Again: The Power of Knowing What You Don't Know*[2] and the study of Atanasov et al.[3], Superforecasters were also much more likely to update their beliefs via small, incremental changes to their forecast. Superforecasters made almost 11 forecasts per question, with an average change of 0.036 probability units per update, versus 2.8 forecasts per question for GJ Open forecasters with an average change of 0.059 probability units. To isolate forecasting skill, we aligned and compared forecasts from both groups made on the same question and on the same day. This approach ensures that each forecaster was benchmarked against other forecasters who had the opportunity to access the same information at the same point in time, as opposed to comparing forecasts on the same question but made at differing time points within the question, which would yield a potential informational advantage to the forecaster making later forecasts.

In total, there were 16,358 days where at least one Superforecaster and at least one GJ Open forecaster made a forecast on the same question, which produced a sample of 53,001 forecasts from Superforecasters and 92,155 forecasts from GJ Open forecasters. We scored every forecast using Brier scoring (or a comparable scoring method for ordinal binned questions); and then for each day, we calculated the average Superforecaster score and the average GJ Open forecaster score. Finally, with 16,358 pairs of averaged scores, we took an overall average as our single score for each group, which is shown in Figure 4. We see that, on average, Superforecasters' daily average error scores were

35.9% more accurate than their GJ Open counterparts (0.166 versus 0.259, respectively).

While these figures account for difference in accuracy among the individual forecasters, we also wanted to benchmark the accuracy when the forecasts were aggregated. Using GJ Inc aggregation methods, we calculated the aggregate forecast for each of the two forecasting groups on the 16,358 days, then scored each aggregate forecast in a similar manner, and finally averaged the daily aggregate error scores. These results are shown in Figure 4, where we see that aggregation had a notably larger effect on GJ Open forecasters, yet the Superforecaster aggregate forecasts were, on average, 25.1% more accurate than the aggregate forecasts using GJ Open forecasts (0.146 versus 0.195, respectively).

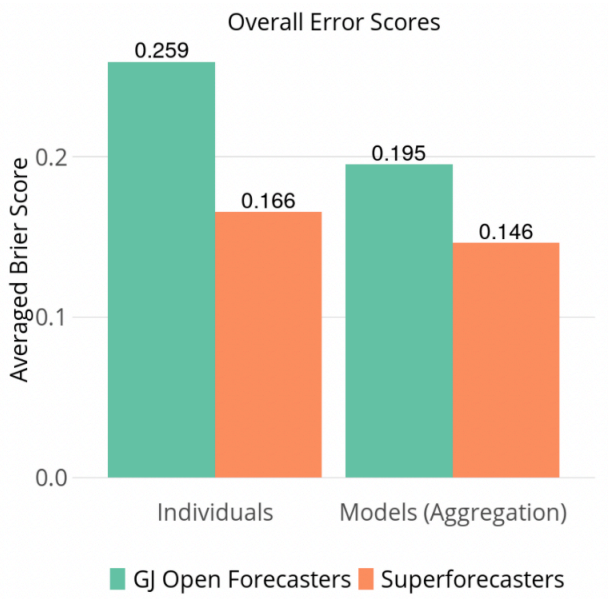

Figure 4. Averaged error scores.

Using the same 16,358 days with at least one forecast from both groups on the same question, Figure 5 shows the average daily error score of Superforecasters' forecasts when charted against the average daily error score of forecasts from GJ Open forecasters that were made on the same day. We see that when forecasts are trivially easy (i.e., error scores near zero), the two error scores are comparable. Yet, as the difficulty of the question increases, the Superforecasters achieve lower Brier error scores, signifying greater accuracy and earlier insight.

To provide reference, we see that when GJ Open forecasters achieve an error score near 0.5, which corresponds to forecasting 50/50 on a binary outcome, Superforecasters have an average score of 0.28, which implies providing a forecast of 0.63/0.37 on the same event and having the first option resolve as true. We see that in cases when GJ Open forecasters are categorically wrong and achieve a score of 2.0 (i.e., putting 100% wrong on the wrong outcome), the average score of Superforecasters is 1.22, which corresponds to a 0.78/0.22 binary forecast with second option resolving as true.

Figure 5. Average daily forecasting error comparison.

In Figure 6 we track the average error score as it relates to the days until the respective questions' outcomes were resolved. We see that across all time points Superforecasters had an average error score that was significantly less than that of their GJ Open forecasting counterparts. The average error score for GJ Open forecasters just before resolution (zero days) was comparable to average error score for Superforecasters on forecasts made 30 days out from resolution. More impressively, the average error score for GJ Open forecasters at 30 days from resolution was larger than any of the average error

scores of Superforecasters on any day up to 300 days prior to resolution.

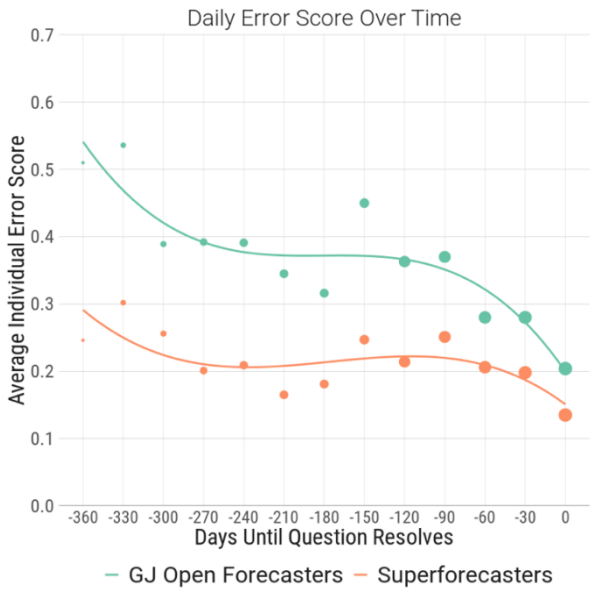

Figure 6. Average daily error scores over time.

To get a better idea of why Superforecasters were more accurate, we calculated forecasting metrics of calibration, discrimination or resolution, and noise for the two sets of forecasts from the individual forecasters. Using the entire set of 167,622 forecasts (those described in Figure 2), we plotted the overall calibration curve for each group of forecasters in Figure 7. For example, seeing a value of 6.5% for the GJ Open curve at the forecasted point near zero implies that while the average of all bin forecasts near zero was actually 0.4%, the percentage of times the corresponding bins resolved as true was 6.5%, and therefore GJ Open forecasters were overconfident in these forecasts.

In general, we see that GJ Open forecasters follow the typical over-confident pattern observed across many other forecasting studies[4], where they forecast between 0% and 20% when there is a greater possibility of the event occurring and similarly over-predict the likelihood of events occurring for forecasts greater than 50%. This pattern is notably absent for the Superforecasters. The miscalibration score in Table 1 translates the calibration curve into a score and reflects the average absolute distance between the forecasted values and the occurrence rates. GJ Open forecasters differ from perfect calibration by an average of 0.068 probability units, whereas Superforecasters differ by only 0.014 probability units, a 79% improvement.

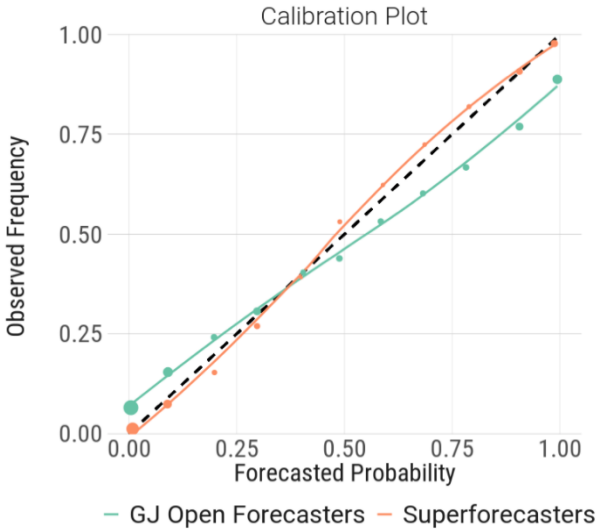

Figure 7. Calibration curves of each forecasting group.

By divvying up the forecasts that were assigned to bins/events that eventually resolved as true, versus those that eventually resolved as false, we calculated two metrics of forecast discrimination. In general, better discrimination implies a clearer dividing line among forecasts for events that resolved true versus forecasts on events that eventually resolved false.

The first metric that measures discrimination is *Area Under the Curve* (AUC), which ranges from 0.5 (completely random guessing) to 1 (perfect discrimination). In Table 1, we see forecasts from GJ Open forecasters had an AUC value of 0.867 versus a value of 0.950 for Superforecaster forecasts. A second metric of discrimination, *d-prime*, corresponds to the notion of effect size between forecasts on bins/events that resolved as true versus those that resolved as false. The metric is oriented such that 0 implies no discrimination and the metric increases as discrimination improves. We see the GJ Open forecasters achieved a d-prime score of 1.764 whereas Superforecasters had a d-prime score of 2.837. The mean forecast value of Superforecasters' forecasts on events that eventually resolved as true

was 0.741, versus a mean forecast value of 0.116 on events that eventually resolved as false. In contrast, the mean forecast value of GJ Open forecasters for events that eventually resolved as true was 0.676, versus a mean forecast value of 0.164 for events that eventually resolved as false.

Table 1. Metrics of comparison.

Metric	GJ Open Forecasters	Super-forecasters
miscalibration	0.068	0.014
AUC	0.867	0.950
d-prime	1.764	2.837
Noise (SD)	0.102	0.046

Finally, in an effort to measure forecasting noise, we selected all days where at least two Superforecasters and two GJ Open forecasters forecasted the same question on the same day (N = 9,586 days). While some deviation is expected between forecasters due to difference of opinion or interpretation of the evidence, excessive noise implies lack of forecasting reliability and can be one of the largest components of forecasting error[5][6]. By looking at the average standard deviation (SD) from each group of forecasts made on the same question and same day, we see that GJ Open forecasters had an average forecast standard deviation of 0.102 probability units, whereas Superforecasters had an average standard deviation of 0.046 probability units, a 55% reduction in noise. This is consistent with the findings of Satopää et al.[7], in analyzing data from the original IARPA tournaments, that Superforecasters were superior to their forecasting peers in tamping down judgmental noise when producing their forecasts.

3. Summary

Our analysis showed that Superforecasters, while a comparatively small group, were significantly more accurate than their GJ Open forecasting peers. The gains in accuracy were relatively consistent across different forecasting timelines and different levels of forecasting difficulty. In further investigating three different components of forecasting error, we found that Superforecasters excelled in comparison to GJ Open forecasters across the board. In particular, Superforecaster forecasts were well-calibrated and without noticeable biases such as over-confidence, which were displayed by the GJ Open forecasters. This implies a forecast from Superforecasters can be taken at its probabilistic face value. Superforecasters had greater resolution in identifying the eventual outcome. Finally, the amount of between-forecaster noise was minimal, implying greater consistency and reliability in translating the variety of different signals—whether from myriad news or social media feeds of varying credibility, or mined from relevant events from history—into a numeric estimate of chance.

References

[1] Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown Publishers.

[2] Grant, A. (2021). *Think Again*: *The Power of Knowing What You Don't Know.* New York, NY: Viking.

[3] Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, 160, 19–35.

[4] Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, 63(11), 3552–3565.

[5] Kahneman, D., Sibony, O., Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. New York, NY: Little, Brown Spark.

[6] Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

[7] Satopää, V. A., Salikhov, M., Tetlock, P., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science*, in press.