Assessing Amazon turker and automated machine forecasts in the Hybrid Forecasting Competition

Andreas Beger and Michael D. Ward

2018-12-03

Abstract

The Hybrid Forecasting Competition (HFC) is an ongoing project to develop hybrid geopolitical forecasting systems that combine human- and machine-generated forecasts. The first project trial period took place in 2018, during the course of which volunteer participants, Amazon Mechanical Turkers, and an automated time series forecasting module we developed provided forecasts for more than 150 questions covering a diverse set of topics. We investigate two questions: (1) how does the performance of turker forecasters compare to volunteers, and (2) what impact did access to the machine forecasts have on forecaster accuracy?

The Hybrid Forecasting Competition (HFC)[1] is an IARPA program that seeks to develop methods for hybrid geopolitical forecasting system that combine human and machine forecasts to answer a broad range of questions about economic, political, health, and other events and trends. The first trial period, or RCT, took place in 2018, during the course of which hundreds of volunteer and Amazon Mechnial Turk forecasters, as well as automated machine models, answered more than 150 questions covering a broad range of issues.

We worked on one of the competition teams, and specifically by contributing a time series forecasting module. Out of the large set of interesting questions one could examine with the results so far, given our specific focus on this project, we will try to examine in this paper two questions: (1) how does the accuracy of turker forecasters compare to volunteer forecasters, and (2) what was the impact of the machine forecasts on human forecaster accuracy?

The Hybrid Forecasting Competition

The goal of the HFC is to find ways to optimally combine human and machine forecasts. For example, machine forecasts can be reliable and scalable, but are constrained by available data, idiosyncratic questions, and cold start problems when a corpus of historical data is not available. Human-generated forecasts on the other hand are more flexible, but also more costly to scale and subject to various cognitive biases.(e.g. Kahneman and Egan 2011).

The competition is organized around Good Judgement-style forecasting tournaments (P. Tetlock 2005, P. E. Tetlock, Mellers, and Scoblic (2017)), where forecasters answer and are scored on a diverse pool of questions. Examples from the first trial period, RCT-A, include:

- What will be the long-term interest rate for South Africa (ZAF) in July 2018?
- How many deaths perpetrated by Boko Haram will the Council on Foreign Relations report for June 2018?
- What will be the daily closing spot price of Brent crude oil (USD per barrel) on 31 May 2018, according to the U.S. EIA?

[1]Funding acknowledgement: This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Each question includes 2 to 5 answer options to which forecasters must assign weights summing to 1. Performance for a single forecast is then based on an ordered multinomial Brier score.

In addition to this basic "human forecaster" tournament, competitors were expected to implement features that would in some fashion or agument these human forecasts with machine-generated tools and forecasts. An explicity requirement for the latter is that they are automated systems, e.g. an ad-hoc hand-tuned and expert-implemented machine model to forecast on a question would not be allowed, rather it would have to be a system that can generate such a model. On the other hand, if a forecaster has the skills and inclination to use data and model as a forecasting tool, that is perfectly allowed.

SAGE approach

One of the distinguishing features of our team's approach was a system that could automatically associate some of the RCT-A questions with a clearly corresponding time series and display these to a forcaster. This is based on a data platform which automatically collects and updates data from a variety of sources, and which can associate questions, based on their title, to an appropriate transformation of a data set if it is in the platform and matches a known, pre-specified pattern or template. If data is found for a question, it can then be shown to users as a simple time series chart accompanying the relevant question.

Additionally, we could use the time series to generate a machine forecast based on a univariate ARIMA model, and which then would either be shown to a user, and/or submitted seperately as a standalone forecast. The module generating these forecasts, "basil-ts", was based on the Auto ARIMA model in Hyndman and Khandakar (2008), which consists of a ARIMA-family model with an automated algorithm for determining a reasonble specific model structure, e.g. whether and how many differencing orders to apply, AR and MA orders, and some other parameters. This is wrapped in additional functionality needed to meet the automation requirements, e.g. recognizing how far a forecast needs to extend, data pre-processing, converting time series to answer option forecasts, updating forecasts with additional information if available, etc.

Research design for RCT-A

Since the scientific goal of HFC is to evaluate various hybrid forecasting techniques, the pripmary aspect of our research design for RCT-A was designed to assess what impact exposure to the time series charts and model forecasts would have on human forecaster accuracy. Incoming forecasters were assigned to one of three experimental conditions. The first group, A, served as control group and only had access to a basic version of the online platform showin the question information and tools to enter weights for the answer options. The second group, B, could also see the time series charts, and the third group, C, could see the chart and machine forecast. All forecasters, regardless of group, were forecasting on the same set of questions.

There were elements in the research design to assess other design choices but they are not relevant to the set of questions we seek to examine here, thus we will not discuss them.

Table 1: Summary of original research design.

Condition	Treatment
A	None; control
B	Chart
C	Chart and machine forecast

Comparison of groups A and B should have shown the effect of seeing a time series chart, and of groups B and C for the effect of seeing a model forecast. In practice, there were issues that complicated the effective design or forecaster groupings and treatments:

- Amazon Mechanical Turk forecasters. Due to lower than expected activity levels, turker forecasters

started to be provided several weeks intot he first trial period. Given activity levels at that time, a decision was made to assign turkers to either condition A or C, but not B. This mean that the group whose treatment was to only see charts (B) consisted only of volunteers, while the other two groups contained mixed populations of volunteers and turkers, thus adding a confounding factor for assessing the impact of charts and machine models.

- Gaps in data and machine coverage. Only about 1/3 of IFPs had time series data available, meaning that within each condition group, most questions did not have a chart nor model. There was also a small number of instances where data was available, but a machine forecast was not. One example were questions related to FluNet influenza case counts, where a change in the data source broke the ability to update chart data, which would have required models to forecast over excessive time horizons. Since the availability of data was related to the type of question, it is not possible to rule out the possibility that questions with data were systematically different in their difficulty from questions without data.
- Changes in the data and machine forecasting platforms over time. Both the data platform and the machine forecasting system suffered from various bugs and related issues, especially during the earlier portions of RCT-A. These problems in some cases resulted in incorrectly aggregated or otherwise inaccurate data, insufficient updating which led to data in the platform falling behind source availability and thus requiring forecasts over longer time periods than neccessary, and bugs in the machine forecaster that led to no or bad forecasts. The quality of the data and machine forecasts displayed to some users thus varied over time and IFPs, in ways that are difficult to reconstruct in retrospective.

In respect to the first two problems, which we can quantify, the effective treatment groups were more complicated and are showin in Table XX.

Table 2: Effective treatments by group after addition of turkers to conditions A and C.

Group	Condition	Forecaster	IFP_Group	Sees chart?	Sees model?
1	A: no chart	Turker	No TS data		
2	A: no chart	Turker	Chart only		
3	A: no chart	Turker	Chart and model		
4	A: no chart	Volunteer	No TS data		
5	A: no chart	Volunteer	Chart only		
6	A: no chart	Volunteer	Chart and model		
7	B: chart only	Volunteer	No TS data		
8	B: chart only	Volunteer	Chart only	X	
9	B: chart only	Volunteer	Chart and model	X	
10	C: chart and model	Turker	No TS data		
11	C: chart and model	Turker	Chart only	X	
12	C: chart and model	Turker	Chart and model	X	X
13	C: chart and model	Volunteer	No TS data		
14	C: chart and model	Volunteer	Chart only	X	
15	C: chart and model	Volunteer	Chart and model	X	X
16	Machine	Machine	Chart and model		

Design, data, and method

DGP

Question comes from T&E to our platform.

Now on two tracks:

1. Human

Table 3: Average Brier score by forecaster group

Forecaster	avg_Brier	n
Machine	0.39	1975
Turker	0.43	39140
Volunteer	0.32	7816

- sees question, and depending on condition the chart and machine forecast
- forecasts, possibly multiple times per day
- Mean daily Brier score

2. Platform

- parses question, finds associated data
- sends question and data to forecaster
- forecaster parses question and data, TS forecast, convert to MN forecast
- mean daily Brier score and/or shown to user

Data

RCT-A lasted from March 7th to September 7th 2018. However, Turkers did not enter until May 2nd, and for an unrelated reason we also discard data after August 2nd. This leaves a total of 48,931 forecasts—7,816 from volunteer forecasters, 39,140 from Turkers, and 1,975 from machine—for 156 IFPs, of which 101 did not have TS data, 6 had TS data but not a machine forecast, and 49 had both TS data and a machine forecast.

Results

Volunteers who had only charts did the best

Table 4: Average Brier score by forecaster and condition groups.

Forecaster	A: no chart	B: chart only	C: chart and model	Machine
Machine	NA	NA	NA	0.386
Turker	0.436	NA	0.432	NA
Volunteer	0.297	0.25	0.372	NA

Table XX summarizes the average Brier scores for all 16 treatment/forecaster/IFP groups in Table XX. As it is quite unwieldy, we include it for reference and will discuss specific insights in more details below.

Table 5: Average Brier scores for all 15 distinct human IFP/treatment groups, as well as the machine forecasts.

Group	Condition	Forecaster	IFP_Group	avg_Brier	sd_Brier	n
1	A: no chart	Turker	No TS data	0.45	0.50	6690
2	A: no chart	Turker	Chart only	0.55	0.56	304
3	A: no chart	Turker	Chart and model	0.39	0.40	2907
4	A: no chart	Volunteer	No TS data	0.31	0.54	510
5	A: no chart	Volunteer	Chart only	0.36	0.46	18
6	A: no chart	Volunteer	Chart and model	0.28	0.34	274
7	B: chart only	Volunteer	No TS data	0.26	0.48	1702
8	B: chart only	Volunteer	Chart only	0.42	0.42	57
9	B: chart only	Volunteer	Chart and model	0.23	0.31	973
10	C: chart and model	Turker	No TS data	0.45	0.50	19502
11	C: chart and model	Turker	Chart only	0.56	0.52	990
12	C: chart and model	Turker	Chart and model	0.38	0.40	8747

Table 6: foo

Model	Statistic	Value
Model 1: Linear model	N	46956
	R^2	0.013
	Adj. R^2	0.013
Model 2: Linear model with IFP random intercepts	N	46956
	Marginal R^2	0.011
	Conditional R^2	0.283

Group	Condition	Forecaster	IFP_Group	avg_Brier	sd_Brier	n
16	Machine	Machine	Chart and model	0.39	0.36	1975

Turkers generally did worse

Turkers overall had worse forecast accuracy than either the machine models or volunteers. To rule a spurious relationship, we can compare the accuracy in directly comparable pairs, e.g. rows 1 and 4 in Table XX consists of the average Brier scores for questions without TS data and in condition A treatment group for turkers and volunteers respectively, and show a difference of 0.14.

Figure shows the results of pairwise comparisons of the average Brier scores for all groups relative to Turker forecasters. Each row in the plot corresponds to a linear model for forecasters from the corresponding condition and IFP group, and the pointranges show coefficient estimates and 95% confidence intervals for Turkers, Machine, and Volunteer forecasters. The Turker values correspond to the average Brier for them in this grouping, while the Machine and Volunteer estimates indicate the relative performance. I.e. for them, values below zero indicate an improvment. Note that the x-scale is reversed, so that better is further to the right.

In most groups volunteers have an advantage sufficient to generate p-values below 0.05, although the effects are minute in comparison to general between-forecaster differences: all models have very small adjusted R^2 values, below 0.01.

Chart and model effects

Teasing out the chart and model effects is a bit harder through pairwise comparison. Instead, we directly coded whether a forecaster saw a chart or model forecast. Forcasters in conditions B and C and for questions that had a chart saw a chart, and forecasters in condition C looking at a question which had a machine forecast will have seen a model forecast. We also leave out the machine forecasts from the data, leaving almost 47,000 forecasts, of which about 25% were made with a chart available, and 21% also had a model forecast available.

We estimated two linear models to predict Brier scores for a forecast. The first had the specifiction:

$$\text{Brier} \sim \alpha + \beta_1 \text{Sees model} + \beta_2 \text{Sees chart} + \beta_3$$

The intercept α corresponds to the average Brier score for a Turker forecasting on a IFP that did not have chartable data, and who neither saw a chart nor machine forecast.

Relatively consistent resulst. First, turkers have worse performance. Second, seeing a chart somewhat helps, seeing a model somewhat hurts.

Figure 1: Pairwise comparisons of turker versus machine and volunteer forecast accuracy. Each row shows estimates from a linear regression model with turkers as reference group, for different groupings of the experimental conditions and IFP data/model availability. The turker estimates (red dots) correspond to the average Brier for turker forecasters, while the other estimates correspond to the relative difference from the turker average for other forecasters. Points further to the right correspond to lower Brier scores, i.e. better peformance.

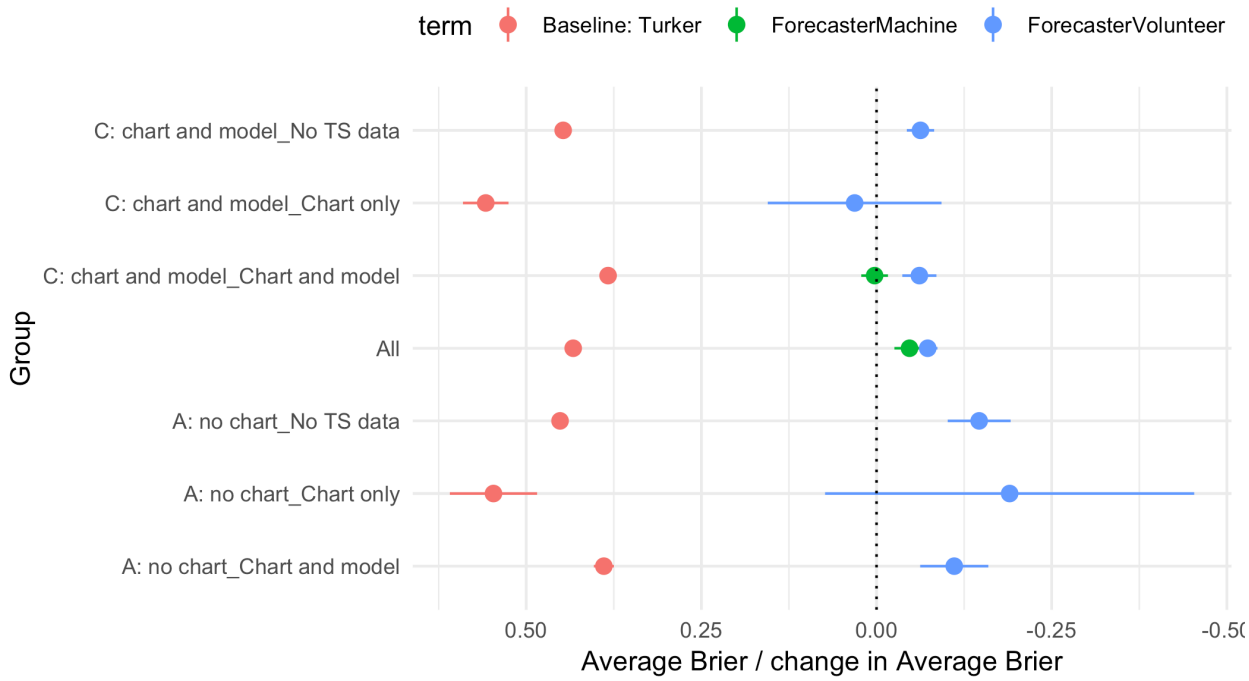
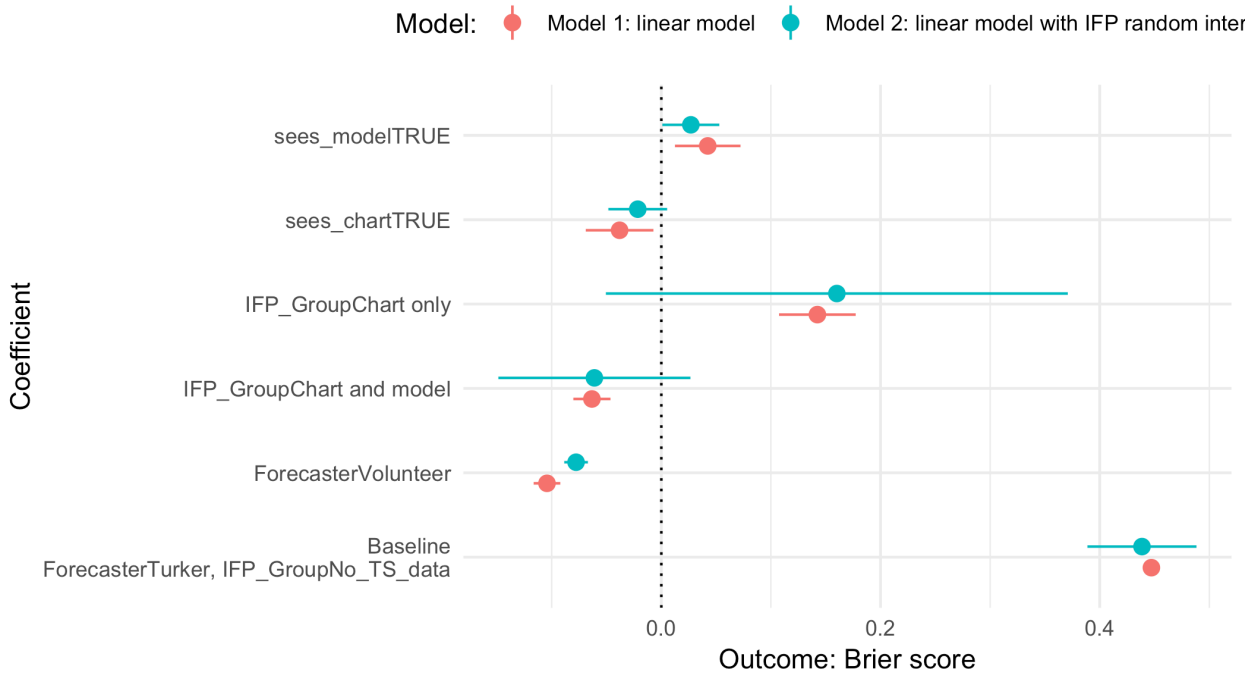


Figure 2: Model estimates



Models generally were disappointing, but did they do well in certain areas?

This is now only looking at IFPs for which a machine forecast was available.

Put the model based version with stars here instead of this:

Data_source	N_IFPs	Machine	Turker	Volunteer
ACLED 2op	7	0.611	0.533	0.396
ACLED 5op	4	0.371	0.554	0.790
earthquakes	2	0.285	0.207	0.199
eia	5	0.184	0.204	0.145
fao-fpi	3	0.386	0.220	0.121
fred	2	0.580	0.488	0.290
gold	3	0.184	0.274	0.171
imf	1	0.222	0.252	0.325
nigeria-security	2	0.355	0.504	0.648
oecd	7	0.430	0.534	0.356
opec	5	0.283	0.222	0.209
other	2	0.416	0.574	0.188
privacyrights	2	0.231	0.321	0.517
stocks	4	0.204	0.204	0.223

Models did relatively well with ACLED 5op, IMF, NST, privacyrights. Poorly on ACLED 2op, FAO, FRED, other?.

Why did condition B volunteers do so well?

Condition B volunteers had the best overall performance out of the 16 groups shown in Table XX. One immediate and plausible story to explain why condition B forecasters did better than both A and C, even after controlling for the impact of Turker forecasters in the other groups, goes like this: we know the machine forecasts were disappointing and less accurate in general. Users who thus saw the chart and machine forecasts were adversely pulled to forecast in line with the machine prediction. On the other than, users who only saw the chart had a good baseline and were able to somehow accurately extrapolate from there.

There are several inconsistencies in the data with this story. First, condition B forecasters did not only do better than other groups on questions that had data and a chart, but also on questions that did not have a chart.

evidence here

Maybe there is a "spillover" effect, where seeing the charts somehow made users just generally better forecasters.

Do we see this in condition C for questions that didn't have a model? No. Condition C Turkers forecasting on a question without a chart did as poorly as condition A Turkers on the same set of questions (rows 1,2,10,11). Condition C volunteers on questions without a machine forecast did worse than both condition B and A volunteers on the same question sets.

evidence here

Another piece of evidence against the spillover effect is that accuracy is actually higher among users who only forecasted a small number of times, which suggests that learning did not take place.

evidence here

Not only were condition B volunteers uncannily accurate on questions even without a chart, they were uncannily accurate on questions where we know the displayed data were wrong.

evidence here

One possibility that we should not discard is that we are digging in noise. All of the effects—for turkers versus other forecasts, for the disappointing performance of the machine forecasts, for conditions or different groupings of IFPs—pale in comparison to differences in accuracy between individual forecasters and between forecasts for different IFPs.

maybe this is a good time to talk about the superactive forecasters and see whether excluding them changes results

Conclusion

Are we digging in noise?

Many of the questions involve monthly time series but are only open for periods spanning a single or few months. Some of the advantage of the human forecasters may thus have been due to operating on a sub-monthly time scale.

We also know that model performance suffered from issues related to automation. Were we able to run the RCT-A questions with the current system and with a data platform that optimally ingests data sources on the bleeding edge of their own updates, our forecasts would have been …

check with updated results from test bed what the performance gain would have been, i think the site visit stats were not accurate anymore

References

Hyndman, Rob J., and Yeasmin Khandakar. 2008. "Automatic Time Series for Forecasting: The Forecast Package for R." Journal of Statistical Software 27 (3). Monash University, Department of Econometrics; Business Statistics.

Kahneman, Daniel, and Patrick Egan. 2011. Thinking, Fast and Slow. Vol. 1. Farrar, Straus; Giroux New York.

Tetlock, Philip. 2005. Expert Political Judgment: How Good Is It? How Can We Know? Princeton, NJ: Princeton University Press.

Tetlock, Philip E, Barbara A Mellers, and J Peter Scoblic. 2017. "Bringing Probability Judgments into Policy Debates via Forecasting Tournaments." Science 355 (6324). American Association for the Advancement of Science: 481–83.