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

Andreas Beger and Michael D. Ward

2018-11-29

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

Turker

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.

Time series data was not available for all questions

Additionally, chartable data was only available for about a third of the questions. The chart and model treatments thus are really part of a nested three layer set of factors: whether a question had chartable data, then whether a chart was displayed to a user, then whether a machine forecast was also displayed to a user. Whether a question had associated data depended very much on the type of question, ranging from questions like "What will be the gold price on [date]" where there is a clearly relevant time series that is easy to acquire and update, to "How many battle deaths will ACLED report in [country] in [date]?" where data are available but have to be transformed to a relevant series, to, on the other end, questions like "Who will win the World Cup?" where it is exceedingly difficult to acquire relevant time series data. Thus it is also possible that questions with chartable and without chartable data were systemtically different in their difficutly and average acccuracy levels.

The effects of the two randomly assigned treatments, seeing a chart and seeing a model, are thus confounded by two other non-ranom factors: whether a forecaster was a volunteer or turker, and whether a question had chartable data or not. This effectively gives us an experiment design like in Table XXXX.

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

Condition	Is Turker?	Chartable data	Sees chart?	Sees models?
A.1				
A.2	X			
A.3		X		
A.4	X	X		
B.1				
B.2		X	X	
C.1				
C.2	X			
C.3		X	X	X
C.4	X	X	X	X

Varying quality of data and machine forecasts over time

To complicate matters further, the quality of the data acquisition and machine forecasting modules changed over time during the RCT. Once a broader set of questions started flowing into the platform when the RCT started, a range of issues affecting both data quality and machine forecast quality arose. Examples include:

- Incorrect aggregation of event data like ACLED and ICEWS that produced incorrect monthly counts of events or casualties.
- Several questions were about monthly oil production rates for various countries. The resolution data are in PDF tables in monthly OPEC oil market reports. As this is hard to automatically scrape, an alternative source was used, but the values in this source do not always match the OPEC values. Nor, for that matter, do OPEC values always match in successive monthly reports.
- Parsing failures that led the forecaster to produce non-sensical forecasts like negative values for count time series, or complete failures, e.g. when one question had a date like "April (04) 2018" instead of the expected "April 2018".
- A limitation in the way Elastic Search aggregates non-standard fixed length time periods, e.g. periods of 40 days, required that for some questions data aggregation had to occur in the forecasting module,

as opposed to data platform module, unlike planned.

Most of these issues and bugs were fixed during the course of the RCT, and thus the quality of both the data and machine forecasts should have improved over time.

Design, data, and method

DGP

Question comes from T&E to our platform.

Now on two tracks:

1. Human

- 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

The first trial period lasted from March 7th to September 7th 2018. The data used in the analysis below consist of 70,006 forecasts made by 1,374 distinct users on 186 questions during this period. The first turker forecasts were on May 2nd.

Results

Basic findings

Brier over time

Make point here that Brier generaly decreased over time? Maybe questions easier, maybe forecasters learn over time? Just establish that all analysis from here is for May 2nd and on only.

Other

Machine forecasts were not more accurate than human forecasts.

Table 3: MMDB by forecaster, data availability, and condition.

Forecaster	Question	A: no chart	B: chart only	C: chart and model
Volunteer	IFP without data	0.15	0.09	0.12
Volunteer	IFP had data	0.25	0.22	0.28
Turker	IFP without data	0.15	NA	0.14

Forecaster	Question	A: no chart	B: chart only	C: chart and model
Turker	IFP had data	0.28	NA	0.27

The low MMDB for condition B forecasters appears to be primarily driven by good performance on questions that *did not* have charts (keeping in mind that 2/3 of questions would not have had a chart).

Table 4: Mean Brier scores by condition

Condition	Turker	Has data	Saw chart	Saw model	Mean Brier	Questions	Forecasts
A.1	No	No	FALSE	FALSE	0.17	124	1660
A.2	Yes	No	FALSE	FALSE	0.16	110	9567
A.3	No	Yes	FALSE	FALSE	0.34	62	846
A.4	Yes	Yes	FALSE	FALSE	0.26	51	3596
B.1	No	No	FALSE	FALSE	0.11	124	4249
B.2	No	Yes	TRUE	FALSE	0.24	62	2152
C.1	No	No	FALSE	FALSE	0.14	124	6045
C.2	Yes	No	FALSE	FALSE	0.15	110	28604
C.3	No	Yes	TRUE	TRUE	0.29	62	2217
C.4	Yes	Yes	TRUE	TRUE	0.25	51	11070

Turker performance compared to volunteer forecasters

Machine forecast impact

Figure 2 looks only at volunteer forecasters in conditions B and C, showing again mean daily Brier scores over time. The right chart shows that forecasters who saw both a chart and model had slighly worse performance throughout the RCT compared to condition B forecasters who only saw the chart. This matches the hypothesis that seeing the machine forecasts hurt performance. However, if we look at the chart on the left, which shows performance for questions for which no data was available, i.e. where both condition B and C forecasters saw neither a chart nor a model, condition C performance was also worse up until sometime in late July.

Discussion/conclusion

Based on the quick check I did during the site visit, probably something like this:

- the machine forecasts were not accurate enough, although sans data and code issues they may have been
- Turkers don't forecast as well as volunteers
- seeing the machine forecasts had a slight detrimental impact on accuracy, although maybe this changed as the forecasts became more accurate later in the RCT
- seeing the charts improved performance, but I don't think it is clear why. Seeing the chart for instances where the data are hard to acquire (e.g. some data transformation are needed) did not boost accuracy more than in other series.

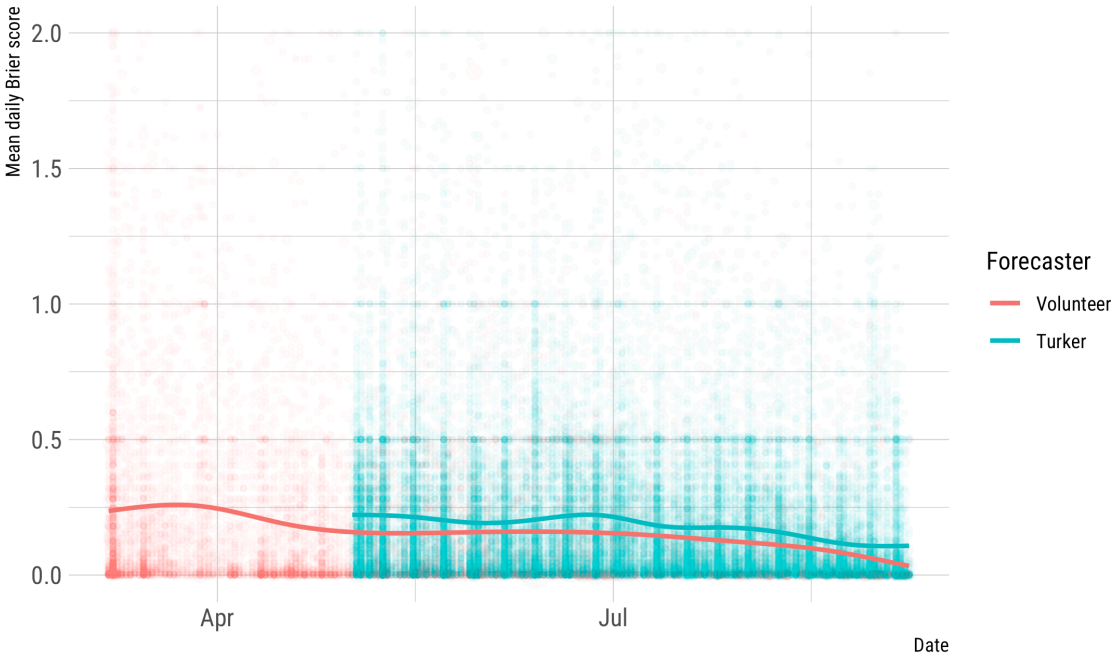

Figure 1: Average mean daily Brier score (MDB) by forecaster type over time. Both volunteers and turkers improved over time.

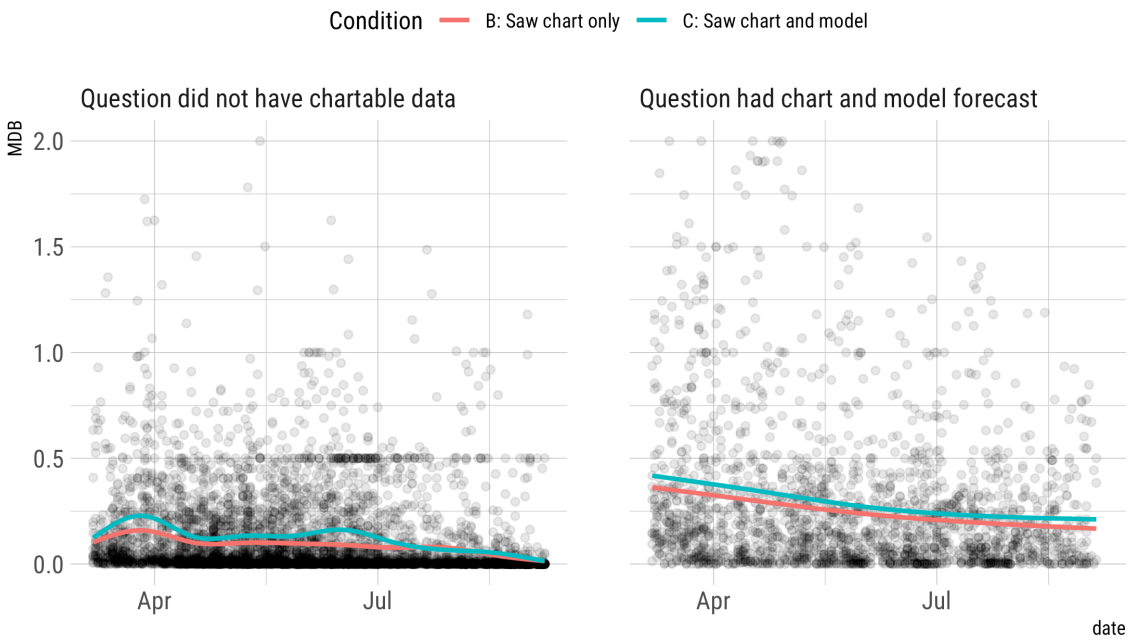

Figure 2: Average mean daily Brier score (MDB) by forecaster type over time. Both volunteers and turkers improved over time.

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.