Analysis of Metaculus Accuracy Just Before Resolution.

Scott Lucier

Econometrics 3421

Professor Andres Gallo

7 December 2022

1. Introduction

Imagine you wanted to know the outcome of a question in advance. Perhaps the outcome of an election, or whether a new medical device will get to market in the next five years. You could just guess, you could maybe even get quite good at it, but you’d be alone and without feedback. Metaculus is a forecasting aggregation site designed to do just this. Users of the site input predictions to user created questions, with the goal of producing an aggregation of inputs into a prediction more accurate (on average) than what one person could do themselves, while providing feedback on personal accuracy through a points system. (Metaculus FAQ).

As a forecasting site, the accuracy of said predictions is probably the most important metric on which to judge it. This paper will examine accuracy at a specific time, just before resolution.

1. Review of Literature

There are several preexisting analyses of Metaculus accuracy by Charles Dillon, “How does forecast quantity impact forecast quality on Metaculus?” and “Data on forecasting accuracy across different time horizons and levels of forecaster experience”. The first looks at how the number of predictors on a question impacts brier score, taking predictions only from the first 25% of the question’s lifetime and excluding questions that ran for more than a year. His finding was that accuracy leveled off after about 10 predictors (Dillon, How Does Forecast Quantity Impact Forecast Quality, November 2021). Dillon’s other analysis is about accuracy and overconfidence over different time horizons (how long the question was open). He finds that accuracy as measured by brier score decreases the longer a question runs, though not by that much (0.151 for under 1 year, 0.167 for 2-4, and 0.169 for 5-10.) (Dillon, Data on Forecasting Accuracy, May 2021).

This paper will look further ahead in a question’s lifespan, just before it resolves. While knowing a the outcome of an event just before it happens isn’t as valuable as knowing it far in advance, being just a few hours or days away from a sporting event, or election, or rocket launch isn’t that uncommon of a situation to find oneself in, and this paper will analyze how well Metaculus performs in those circumstances.

In addition, accuracy just before resolution can be taken as a barometer of how well Metaculus is integrating new evidence as time goes on, if there isn’t an improvement over earlier in question lifespans something quite strange is happening, arguably in violation of the basic laws of probability—more evidence is supposed to improve predictions[[1]](#footnote-1). This paper will check whether the Metaculus community is falling prey to any severe bias when integrating new evidence.

1. Data

Data was scrapped from the Metaculus API using a python program supplied by Metaculus user rodeo\_flagellum[[2]](#footnote-2) (Metaculus API). The program downloaded only resolved binary questions, and further sortation reduced the data set down to the following variables: final community brier score, number of votes, comments, and predictions made, planned close time, and open time. I originally planned to use question topic data (i.e. questions on economics, covid, technology, etc.), but unfortunately those aren’t currently available through the API, and applying categories manually for thousands of questions was untenable[[3]](#footnote-3). Using the full time series data was also considered, but the sheer quantity of data was quite literally impossible to run with the software I have access to[[4]](#footnote-4). Instead of using individual user predictions and batches of user predictions in the time series, I used the latest median prediction from the history section of the API (Metaculus API). I also removed questions that resolved ambiguously, 67 out of the 1275 resolved questions. I didn’t have any way to work with ambiguous questions, there’s no brier score for “no resolution.” Thankfully, only about 5.255% of questions resolved ambiguously, and a user with a good eye can often spot them in advance by poorly defined resolution criteria.

Final community brier score is a measure of how accurate the community prediction (a weighted median favoring recent user predictions) was at the moment the question closed. It is represented by:

Where S is the score between 0 and 1, p is the probability forecasted for the event (weighted community median in this case) and f is a binary resolution, 1 for a yes and 0 for a no. A lower score means better accuracy (Metaculus FAQ). This variable wasn’t in the raw data, but is trivially derivable from the final community prediction and the binary resolution, which are included in the API.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. Dev. | Minimum | Maximum |
| Prediction Count | 1208 | .106671 | .1476521 | .0001 | .9025 |
| Table 1: Brier Score Community Summery | | | | | |

The number of votes is the sum of user up-votes and down-votes on a question, used as a rough measure of question quality as seen by the userbase. Comment count is the number of user comments a question received, and can be thought of as a proxy for user engagement. Predictions made is how many individual predictions were made on a question. Tables and graphs for each of those are below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. Dev. | Minimum | Maximum |
| Votes | 1208 | 11.99503 | 19.59499 | -20 | 443 |
| Table 2: Votes Summery | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. Dev. | Minimum | Maximum |
| Comment Count | 1208 | 24.30712 | 69.79471 | 0 | 1857 |
| Table 3: Comment Count Summery | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. Dev. | Minimum | Maximum |
| Prediction Count | 1208 | 257.5637 | 442.4033 | 2 | 6693 |
| Table 4: Prediction Count Summery | | | | | |

Chart, scatter chart

Description automatically generated

Figure 1: Graph of votes vs brier score

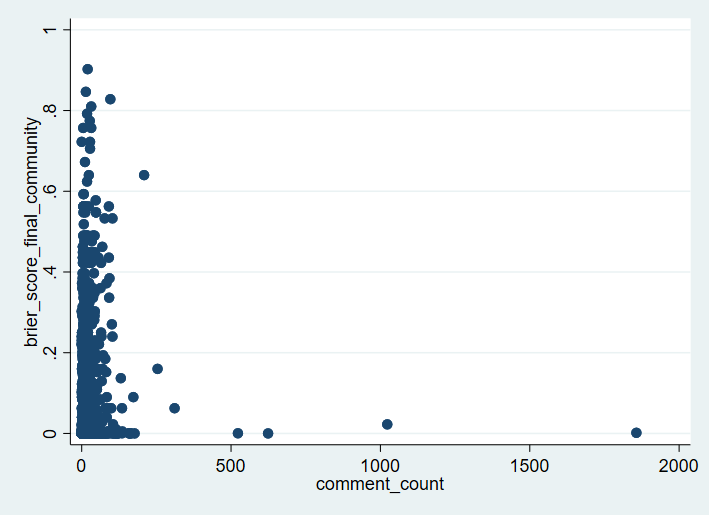


Figure 3: Graph of comment\_count vs brier score

Chart, scatter chart

Description automatically generated

Figure 3: Graph of prediction\_count vs brier score

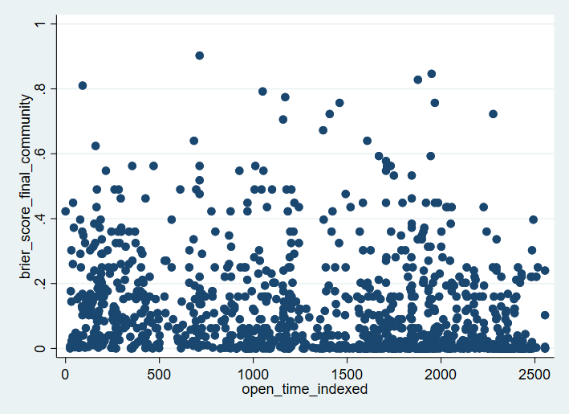


Figure 4: Graph of open\_time\_indexed vs brier score

Open time is the time the question originally opened for predictions. It’s useful for tracking how old Metaculus was at the time the question was opened. In the API it’s in a standard date format, I have indexed it to the date of the first question and converted it to an easier to work with integer format. The earliest question in the data set is at 10/2/2015, and open\_time\_indexed increases by one for each day since then, capping at 2554 (10/6/2022). Table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. Dev. | Minimum | Maximum |
| Open Time Indexed | 1208 | 1396.965 | 736.2859 | 0 | 2554 |
| Table 5: Open Time Indexed Summery | | | | | |

1. Methodology and Results

First, most of the variables were converted to log form, mostly for linearity reasons, but there is good reason to think that additional votes, comments, and predictions have the most marginal impact when their absolute number is low, for the simple reason that the thousandth contributor doesn’t change the prediction as much as the second, or third, and intuitively is less likely to bring as much new evidence to the table. Using the Metaculus prediction[[5]](#footnote-5) instead of the community prediction was considered, but the Metaculus prediction was not instituted until partway through the data’s runtime and is missing on a good portion of the early data.

Regression equation below.

Where Y is the final community brier score, B0 is the intercept, X1 is votes, X2 is comment count, X3 is prediction count, and X4 is open time (indexed to first question).

First, we test for heteroscedasticity using the Breusch—Pagan test with a significance level of .05.

Table

Description automatically generated

Table 6: Breusch —Pagan Test

Since the p-value given by the test is less than .05, we reject the null hypothesis of homoscedasticity. We will use Stata’s robust[[6]](#footnote-6) feature to compensate for the heteroscedasticity of the data.

Next, multicollinearity. Typically, a Variance Inflation Factor of above 3 or 5 is used as the cutoff point for the presence of multicollinearity, we’ll split the difference and use 4.

Table

Description automatically generated

Table 7: VIF of variables in model (robust mode on, but they’re all below 3 either way).

We conclude that multicollinearity is not an issue for this model.

Our final regression is:

Table

Description automatically generated

Table 8: Robust Regression of Brier Score = lnvotes, lncomment\_count, lnprediction\_count, open\_time\_indexed

As seen in the output, all variables except votes are significant, but with low coefficient sizes and large variability in the 95% confidence interval for their coefficients. R2 is also very low. These results indicate missing variables or a highly random event[[7]](#footnote-7). We will confirm this with a RESET test.

Graphical user interface, text

Description automatically generated

Figure 5: RESET Test

The RESET test confirms our suspicions. The model is missing variable(s).

Another version of the model with the logarithm of open time indexed is presented below, but was rejected for having a lower R squared and variable t-values.

Table

Description automatically generated

1. Conclusions

First of all, the mean brier score of .106671 is much lower than those seen in Charles Dillon’s prior works (.151, in the most optimistic case), which likely means that the additional information from the additional run time of the questions is being integrated into the prediction successfully. This makes sense, as it should be easier to predict the outcome of an event closer to its occurrence, as generally there is more information about the event, and is also supported by the results of the RESET test, which shows missing variables. One of the obvious candidates for a missing variable is how effective the Metaculus community was at integrating new information into the prediction for the question as said information became available. Other obvious missing variables include question difficulty[[8]](#footnote-8), question category[[9]](#footnote-9), how well written the question was[[10]](#footnote-10), total score of the individual forecasters who participated in the question (perhaps separated by topic), and score of the user who made the question.

Two of the measured variables correlate with a lower brier score, prediction count and open time. This was expected, more predictors increasing accuracy is in some sense the entire point of the Metaculus website, and as the website has aged it has acquired more predictors and (one would hope) better trained predictors. Votes and comment count, on the other hand, correlate with a higher score, which was not expected at all. Why would having a higher rating for a question correlate with less accuracy? And why the same for comments?

Several explanations come to mind. Votes and comments may be given to “less obvious” questions that are just genuinely harder. More comments may track with higher controversiality, and questions may be controversial for being harder to predict, or for a political nature that causes part of the userbase to predict on ideological grounds. There are some groups of questions dealing with elections or other multi-answer but non-continuous things in the dataset that all have minimal vote and comment counts[[11]](#footnote-11), but often have high accuracy from all the “definitely not going to win” candidates. This alone might be enough to correlate low votes and comments with high accuracy.

Overall, results are inconclusive. Higher prediction count is clearly correlated with higher accuracy, and from the mean brier score alone it’s clear that accuracy is higher just before resolution than at earlier periods, but the problems with vote and comment count put the analysis as a whole on shaky ground.

Works Cited

Dillon, Charles. “Data on Forecasting Accuracy across Different Time Horizons and Levels of Forecaster Experience.” *Forum.effectivealtruism.org*, 27 May 2021, forum.effectivealtruism.org/posts/hqkyaHLQhzuREcXSX/data-on-forecasting-accuracy-across-different-time-horizons. Accessed 2 Nov. 2022.

Dillon, Charles. “How Does Forecast Quantity Impact Forecast Quality on Metaculus?” *Rethink Priorities*, 1 Oct. 2021, rethinkpriorities.org/publications/how-does-forecast-quantity-impact-forecast-quality-on-metaculus. Accessed 2 Nov. 2022.

Metaculus.com. “Metaculus API.” *Metaculus.com*, 2022, www.metaculus.com/api2/. Accessed 10 Oct. 2022.

Metaculus.com. “Metaculus FAQ.” *Metaculus.com*, 2022, www.metaculus.com/help/faq. Accessed 30 Oct. 2022.

‌

1. Obviously humans aren’t ideal reasoners, but we’re usually not *that* far away. [↑](#footnote-ref-1)
2. The 36th highest ranked Metaculus user as of 11/2/2022 [↑](#footnote-ref-2)
3. Perhaps in a future project. [↑](#footnote-ref-3)
4. A ~93 megabyte text file that crashed notes, and far too many entries for Stata. [↑](#footnote-ref-4)
5. A special algorithm that weights higher ranked user’s predictions higher. [↑](#footnote-ref-5)
6. This uses broader assumptions that work for non-normal variables. [↑](#footnote-ref-6)
7. Which, technically, is a matter of missing variables, if not ones we can get. [↑](#footnote-ref-7)
8. How exactly to define this is a hard question. [↑](#footnote-ref-8)
9. Unfortunately not present on the API [↑](#footnote-ref-9)
10. Also hard to define but could matter for drawing in experienced forecasters who don’t want to waste points on a question that might resolve badly. [↑](#footnote-ref-10)
11. I checked and there are comments and votes on the webpage, just not the API. [↑](#footnote-ref-11)