Model evaluation is important as we want to know whether stake on a model or whether to further improve it. Sadly, daily scores don't give us much of an idea about the model quality. Diagnostics only diagnose with past data and sometimes even that is extremely computationally intensive.

Here I am proposing a method to use multiple daily scores to estimate future average final scores.

First I downloaded daily resolving scores of many models. For every day from 1 to 19 I calculated standard deviation. Than I gave the algorithm my resolving scores. Knowing the standard deviations and my scores I calculated a million possible resolving scores for each day. Then I averaged them to calculate the most likely future average score. Also averaging the values across days I calculated a standard deviation of million possible average scores. A chose a confidence interval to calculate interval of future scores from the most likely score and a margin of error.

I am not as bright as many of you so it's possible that this approach is very wrong. Also bugs in my code are possible. Thanks jrAI for sharing his smart code for fetching daily resolving scores from past rounds.

You can try the app here:

REMRS