In general, I would not put too much weight on one round. You need a lot more observations to get an idea of your model's quality.

However, if you experience one or more weeks in which your performance is a lot worse than your worst performance on your internal validation, this is an indication that you are overfitting your validation data or that your validation data is not very representative of the live data. In that case, I would reasses my model-building pipeline.

Mind you, both of these issues are hard to avoid. To avoid overfitting your validation data, you can use cross-validation on the training data and only rarely compare models on the validation data. You can also designate validation as well as test data: use validation data for early-stopping and other decisions and use test data only for final model selection. What data is representative of live data? Who knows? One thing you can do is look at historical live performance of other long-term participants. Then split your data into a couple of folds and run a couple of models to get an estimate of performance on the different eras in your data. Pick eras on which your models perform roughly as well as the average participant has in the last few years (corr and sharpe)… or, if you're more optimistic, pick eras that reflect performance on the whole dataset you have available.

These are just a few ideas, decide for yourself what works and makes sense.