I have been taking part in the tournament for a month or so now, and was hoping some more experienced users might help me to interpret the model diagnostic values more clearly. I have read all the information I can find, and the defined meaning of each metric is clear, but I haven't found much in terms of applying the values themselves to the decision-making process, or of how well the diagnostic scores on validation have historically translated to live results.

For example, after a few lockdown evenings spent tweaking my code, and training only on training data, Validation Corr on my main (staked) model shows in diagnostics as 0.0224 (Fair, 62.46%) with a Validation Sharpe of 1.2213 (Excellent, 97.37%) and a Feature Exposure of 0.0327 (Excellent, 100.00%). Live corr has outperformed this, but over far too short a period to base any conclusions upon.

By removing the call to normalize_and_neutralize in a second (unstaked) model, Corr improves to 0.0244 (Good, 76.10% - this turns it green, I like green) at the expense of Sharpe 0.7627 (Bad, 32.77%) and feature exposure 0.2914 (Bad, 32.54%).

In this case, I instinctively prefer the norm & neut model despite the lower validation correlation, due mainly to the wider standard deviation and an unpleasant-looking max drawdown, but I don't see anything quantifiable to determine the optimal points between the two, or which of the metrics are truly lagging behind the rest and by how much.

Not everything can be quantified in the art of data science of course, but while a feature exposure of "Excellent 100%" certainly appears ideal at first glance, I'm sure there are good reasons for 0% of other users keeping exposure this low, and searching for optimal diagnostic scores yields scant results. Perhaps I am still selecting too many features, or my neutralization proportion is too high…

Also, I wonder how high a score it is possible to get on validation correlation, without training on validation of course! I'm sure I can still improve it further from here, and have scored higher using trial-and-error methods with no guarantee that the logic behind them transfers to live… but at a certain point time spent trying to improve it further would be time better spent elsewhere, and it's difficult to identify where that point is. The top leaderboard scores of >0.05 are live scores over an extended period, which at least hints at a higher validation corr being possible, but how high can it realistically get without overfitting to validation? Should I move on to something else once validation corr reaches 0.025, for example, or can it be pushed to 0.03, 0.04…?