Hello Data Scientists, it's month-end and that means we have some announcements! We have new features ready for you, the deprecation of an old feature, and some big news.

New Feature Metadata

As of round 287, we are making public the long awaited Feature Metadata file. This is downloadable each round through the dataset API and the dataset zip as "features.json". Inside the file are two sections: "feature_stats" and "feature_sets".

The 3 feature sets in the file are lists of feature names:

- Legacy

: 304 of the original 310 features that were carried over to the new dataset. You can use this set to achieve nearly the same model as the legacy data.

- Small

: 38 features that were found to be the most important based on Shapley values calculated by Michael Oliver in this forum post. You can achieve relatively high performance using these features without many compute resources. Check out our example model to see how to use this!

- Medium

: 420 features that can capture a large amount of information that exists in our dataset. This is not a strict superset of the "small" dataset and does not use Shapley values to pick these.

The feature stats are a map of feature names to statistics about that feature:

["feature_dichasial_hammier_spawner": { "legacy_uniqueness": 0.1778140232616494, "spearman_corr_w_target_nomi_20_mean": -0.0006870366768374209, "spearman_corr_w_target_nomi_20_sharpe": -0.06827874939205324, "spearman_corr_w_target_nomi_20_reversals": 7.490937957833377e-05, "spearman_corr_w_target_nomi_20_autocorr": -0.021193429374978014, "spearman_corr_w_target_nomi_20_arl": 3.3248407643312103 }, ...]

Here's an explanation of each statistic:

- Legacy Uniqueness

: how unique this feature is compared to the legacy dataset. This is 1 - max(corr(feature, legacy_features)), so a legacy_uniquness of 0 means that feature is perfectly correlated with at least one feature in the legacy dataset.

- Spearman Correlation w/ Target Nomi (20-day)

: these are aggregated statistics over the per-era correlation between this feature and target_nomi_20. Mean and Sharpe (mean divided by standard deviation) are given as well as the following statistics:

- Reversals

: how many times the feature changes between correlated and uncorrelated with the target. A low value implies consistency of correlation.

- Autocorr

: Similar to autocorr from diagnostics, this is the correlation of lagged per-era corr values with themselves. If this is high, the feature is more likely to be consistently correlated (or uncorrelated) with the target.

- Average Run Length (ARL)

: The average number of contiguous eras that this feature stays highly correlated with the feature.

- Reversals

: how many times the feature changes between correlated and uncorrelated with the target. A low value implies consistency of correlation.

- Autocorr

: Similar to autocorr from diagnostics, this is the correlation of lagged per-era corr values with themselves. If this is high, the feature is more likely to be consistently correlated (or uncorrelated) with the target.

- Average Run Length (ARL)

: The average number of contiguous eras that this feature stays highly correlated with the feature.

Numerai-CLI 0.3.2

The newest version of the numerai-cli includes small updates to the way we build Dockerfiles allowing you to be more modular with your code and share local packages between nodes. It also provides support for Cron-based Nodes when running “numerai node test”.

Before this update, the CLI only gave Docker access to whatever folder you configured for a given node. Now, it gives Docker access to whatever parent directory you run “numerai node deploy” in. This is especially useful when copying code from a sibling directory; for an example of how to accomplish this, check the wiki.

Deprecating Signals Submission Diagnostics

Until last quarter, the only way to get diagnostics for your signal was to submit it and wait over 10 minutes to see the results. Now that we have a new Diagnostics Tool (which runs in 2 minutes or less), we no longer need to run the expensive and redundant diagnostics when you submit.

Starting October 25, 2021, we will no longer be providing diagnostics when you submit a signal to Numerai, instead you can use the always-available diagnostics tool.

Coming This Quarter

For Q4 2021, we want to continue unification of Signals and Tournament by starting to score Tournament submissions against a 20D2L target (20 days long w/ a 2-day lag). This target is already included in the new data, we are just switching our scoring to use it. This change will likely take place sometime in November.

We are also preparing more data. You all experienced the amazingSupermassive Data Drop last Quarter, but we aren’t done yet. Numerai Data V3.2 is coming December 25, 2021; prepare yourselves and your computers.