Tournament Dataset V4

On April 5, there will be a new dataset available. It contains 141 new features, as well as targets available for all eras - including what was previously labeled “test”.

What’s New?

- There are 141 new features included, for a total of 1191 features, plus targets available for all eras.
- Test eras are now considered part of validation. This means there is a train.parquet, validation.parquet, and live.parquet file available.
- Each live era will graduate to validation data, and targets are populated in the graduated eras as soon as possible.
- The existing features have been slightly improved, and thus renamed. There will be a map between v3 and v4 features so that you can use your previous feature research on new data without starting over, if you wish.
- The optional targets have also been slightly improved and so are named with a new pattern, eg. “target_jerome_v4_20”.
- Submissions are now only required to have live predictions. Any other predictions submitted will be safely ignored. We will accept submissions from any version, as long as you have predictions for all of the live indices.
- As always, neither the legacy (v2) or the current versions (v3) are being changed in any way, so any automated model you have will continue to work as always.
- The example script repo will be updated with v4 versions of each example model. These models will be functionally identical, but will be updated to read from the v4 dataset.
- There is a new data page at numer.ai/data where you can read about the available files, download the data from each version, and find snippets for how to download the data directly in your code.

New Features

Below is a plot of 3 models’ cumulative performance vs Target Nomi, from the time the training set ends, until present.

The green line is the legacy v2 data with 310 features. The orange line is v3 data which has 1050 features. The blue line is the v4 data which has 1191 features.

[

1434×1210 80.8 KB

](https://forum.numer.ai/uploads/default/original/2X/8/8db7cffbb3f5627b962042a0ee8b96d62ab6b8ce.jpeg)

On an FNCv3 basis, you’ll notice that extra features help a bit more. But not on the very most recent data! All of the models are doing poorly in FNC terms for the most recent period. My hope is that you can use this new data to build models that do better in these sorts of times. There could be a lot of TC for the taking if you can solve that.

[

1428×1214 81.5 KB

](https://forum.numer.ai/uploads/default/original/2X/b/b4c2cf1b3bf190bc5b1361d68d142a9adc5519cc.jpeg)

Continuous Retraining

I’ve added two more lines to the exact same plots above: The red line is the v4 data, but it’s retrained every 52 eras with all of the data available up to that time. The purple line is retrained every 52 eras, but only on the most recent 520 eras, rather than all of history.

You’ll see that retraining annually can provide a nice boost to performance. On a pure corr basis, retraining on only the most recent data is perhaps the best.

[

1436×1208 100 KB

](https://forum.numer.ai/uploads/default/original/2X/c/c7a2a0b82248054278187653a7fb2130cece3a36.jpeg)

However on an FNC basis, you’ll see that retraining only on more recent data hurts a great deal compared to training on all of history.

[

1430×1214 102 KB

](https://forum.numer.ai/uploads/default/original/2X/c/cf989002a9048ccead35c26366145ace886338e4.jpeg)

We hope that users will be able to find more clever ways to use recent data to get an edge, without losing generalization like we seem in the example above.

Legacy Data is still useful

Here I want to show some interesting correlations between each of these 5 models.

Looking at the correlation between their predictions, you can see that v2 only has a correlation of .729 with v4.

[

1600×316 50.4 KB

](https://forum.numer.ai/uploads/default/original/2X/b/b7b0d997699d96d2485e156e16758c1c4f434529.png)

Perhaps more interesting is looking at the correlation between the models' performances. First in corr terms, and second in FNC terms.

[

1600×315 51.3 KB

](https://forum.numer.ai/uploads/default/original/2X/d/dcf27d0f5c961f2e0161b4ed13e1d3c8251ea76a.png)

[

1600×277 47.2 KB

](https://forum.numer.ai/uploads/default/original/2X/e/e5647c352797b61cac7f491a17a0b14a571ba467.png)

What this indicates is that a combination of these versions could be nice.

To make this line of research easier, we've added a feature set to the features.json file which contains only the features from the v2 dataset. This will allow people who are especially fond of the legacy data to start with those, but begin using test data and gradually building out the features they want to use.

eg.

napi.download_dataset("features.json", "features.json")

with open("features.json", "r") as f: feature_metadata = json.load(f) features = feature_metadata["feature_sets"] ["v2_equivalent_features"] read_columns = features + ["era", "data_type", "target"] train_data = pd.read_parquet('train_data.parquet', columns=read_columns)