Numerai's Super Massive Data Release

Unlock a new dimension of performance and research

Anson Chu

Follow

Numerai

--

1

Listen

Share

Super Massive Data Release

The Numerai dataset contains decades of historical data on the global stock market. Machine learning models trained on the dataset learn to predict stock returns and earn cryptocurrency (NMR) based on performance in the Numerai Tournament.

The performance of the models on Numerai are driven by two things — the quality and quantity of information embedded in the dataset, and the skill and creativity of the model creators used to turn that information into predictions.

Today, we are releasing a new version of the Numerai dataset that massively increases the amount of embedded information with 3x features

and 5x training data,

and unlocks a whole new dimension of research possibilities with 20x new targets

.

We expect the new dataset to give all models a huge boost in performance and profitability, and we can't wait to kick off a new season of research with all of you in the community!

New Example Scripts

Whether you are a new or existing user, the easiest way to get started with the new dataset is with our example scripts repo on Github.

GitHub - numerai/example-scripts: The official example scripts for the Numerai Data Science…

The official example scripts for the Numerai Data Science Tournament - GitHub - numerai/example-scripts: The official…

github.com

In this repo you will find examples of how to download the data, train the new official example models, compute validation metrics, generate predictions, and upload submissions back to Numerai.

Also included is a new analysis and tips notebook that will guide you through the dataset, and explain advanced concepts such as feature exposure and neutralization, era-wise time-series cross-validation, and ensembling.

API and Website Changes

Improved Data API

You can download each file in the new dataset individually in both CSV or Parquet file format. Available now in our GraphQL API and Python client.

Improved Diagnostics Tool

You can test the performance of your models on the new validation data with our improved Diagnostics tool — available 24/7, runs in <60s, and an upgraded UI with full access to all historical runs.

New Season of Research

To kick things off, the masterminds behind the new dataset, Michael Phillips (MikeP) and Michael Oliver (MDO) share their research process building the new example model and notebook, and show you their analysis on how the new data can improve model performance and profitability.

Super Massive Data Release: Deep Dive

Highlights We have just released the biggest upgrade to Numerai's dataset ever. The new dataset has 4x the number of…

forum.numer.ai

Both of the Michaels will also be doing a live AMA over Zoom on Thursday September 9th at 10:30AM PDT / 1:30PM EST

. Submit your questions and comments here.

Join us at our community chat and stay tuned for the details!