A New Data Science Competition Where Being Different Pays

Michael Phillips

Follow

Numerai

--

Listen

Share

Not every data scientist wants to play “Find the best version of XGBoost”.

Data science competitions have become stale. As the field of data science was emerging, landmark competitions like the Netflix Prize and early Kaggle competitions encouraged new algorithms and creativity. But now there are a handful of algorithms which are known to perform best at certain types of problems. Today, data science competitions typically boil down to: “throw 1000 different XGBoost models at the problem, cross-validate, and see which combination of hyper-parameters + preprocessing steps performs the best”. The process of building a good model on any dataset has become monotonous and automatic. (In fact, Google, Microsoft and others have automated it with cloud services aptly called ‘AutoML’ because these tools require no insight or creativity.

)

The Numerai data science tournament is different. Numerai gives out their grid searched XGBoost model for free and now poses a new challenge to its data science community:Can you build a model that’s different from what everyone else has submitted?

The problem is no longer about finding the best parameters for XGBoost, but about building an original model that hasn’t yet been discovered.

A New Type of Data Science Competition

In the Numerai Tournament, participants will be paid not only for performance — they will also be paid for originality and uniqueness.

The Numerai tournament, if you aren’t familiar with it, is a data science competition where participants are given a dataset that appears to be a simple regression problem. In reality, the data is obfuscated stock market data, and the participants are predicting future price movements. Those predictions are then combined by the Numerai Meta Model, and that meta model is used to control the capital of the Numerai hedge fund.

I now work for Numerai full-time, after being a participant in the tournament for a while before that. I had always avoided other data science competitions because they felt tedious — like my purpose in the competition was to just put the problem into my hyper-parameter tuner, and then throw as much compute as possible at the problem. Numerai drew me in, though. The idea of many different data scientists submitting unique models to help control the capital in a real hedge fund is incredible. Then, when you start playing with the data, you see that it feels nearly impossible to beat the model that they give out for free, and you are forced to think deeper about how other users are managing to climb the leaderboard. Even as unique as it is already, we are now making the Numerai tournament much, much more interesting.

Previously, users had been paid only according to how well their predictions match up with what really happens in the market. If a user’s predictions perform better than random chance, they are rewarded. If they perform worse, they are penalized. The result of this is many users submitting very similar models of the same structure, because that structure is known to perform consistently well.

As it turns out, Numerai doesn’t necessarily benefit from getting 1000 submissions that all predict roughly the same thing… only the first 1 or 2 of those submissions are really useful, and the other 998 might be redundant information. Numerai’s true power emerges from having many unique models that all have different strengths. Then those unique models become individual building blocks, and we can combine them in a way that creates an incredibly powerful and unique portfolio.

We know that the Numerai dataset is rich, and to get all of the information out of it, we need users to try new things. Users are already using modeling approaches on our data that we have no idea how to recreate. We want to encourage everyone to continue to develop models like these. These types of users are extremely valuable to the meta model, but are not being proportionally rewarded yet. That’s all about to change.

That’s why we’ve introduced Meta Model Contribution. Meta Model Contribution estimates how valuable each model is to

the meta model that runs the hedge fund, so that those users can be paid based on their real value added.

The result is an incentive structure that aligns directly with the hedge fund. By reorienting the very objective of the tournament, we are turning all of the data scientists into hyper-efficient data-miners for the hedge fund.

The New Data Science Process

Data scientists will be familiar with the idea of having an optimization function. This is simply the calculation of a metric that measures the performance of your model, so that you can compare your various attempts against one another in an objective, quantifiable, and potentially automatic way. In typical data science competitions, performance is one-dimensional. For example: “Maximize the percent of rows classified correctly”, or “Minimize the average squared distance between each prediction and its respective target”. In any case, the data scientist will try hundreds or thousands of different types of combinations of models, parameters, and pre- or post-processing steps, and see which one gives the best result.

The new process for the Numerai tournament will require consideration of an entirely new dimension — instead of only considering performance, the data scientist will need to consider their predictions’ independence with respect to other user predictions. An intuitive way to think about how to quantify a good model in this new two-dimensional competition might be performance * (1-correlation_with_all_other_models).

Based on the way Numerai rewards participants, a unique model with a score of 0.01 might be rewarded more than a standard model with a score of 0.03.

So a data scientist might first want to generate a handful of what she thinks are the most common types of models. Then she can write an optimization function that penalizes prediction-similarity to these models, and then use that new metric to iterate on her own pipeline to generate the predictions that she will submit. With this approach, she can construct a model that has a great chance of being extremely unique while also performing well on the dataset, and maximizing her payout.

This is a rough first-cut at tackling this new tournament format. We expect thatour community of data scientists will be able to push this to the limits, far beyond any ideas we currently have.

Meta Model Contribution aims to reward the users who are best able to find these unique and valuable approaches.The way we can quantify this is by first building a meta model from all users’ submissions. Then we can take each submission and residualize (or subtract out) the meta model predictions from the submission. Whatever is left over after being residualized to the meta model is what we score versus the true stock market results. This encourages users to find new information in the data that few others were able to find.

Rewards

We’ve paid out $1,100,000 worth of cryptocurrency to users in the last 3 months alone. We want future payouts to be allocated to the data scientists who help the hedge fund the most.

If you’re a data scientist or machine learning whiz, head tonumer.ai to get started modeling, controlling the hedge fund’s capital, and earning your share of the tournament payouts.