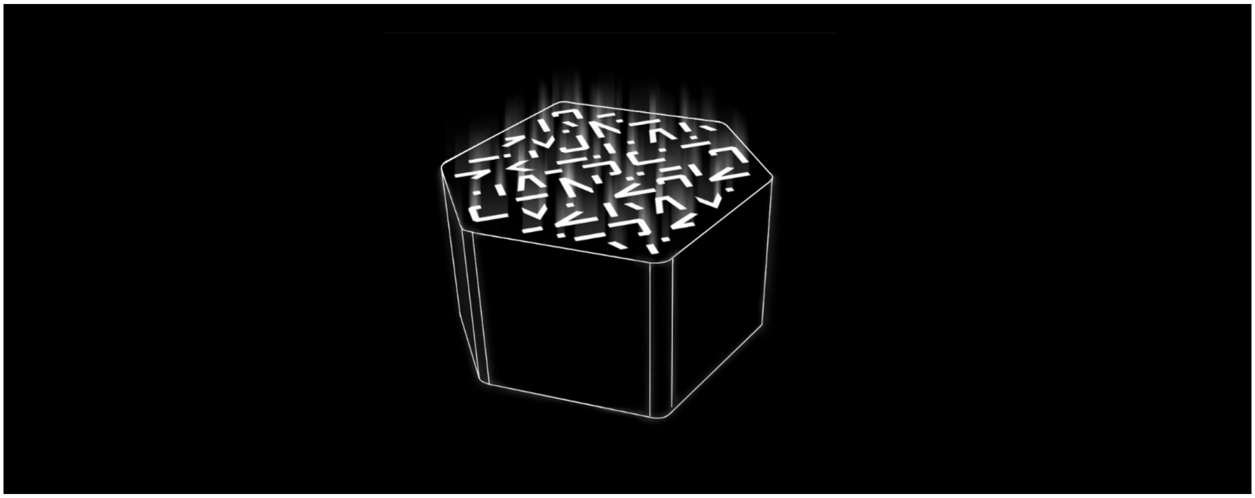
Keno Leon Follow
FRONT END / WEB DEVELOPER — DESIGNER: www.k3no.com
Aug 17 · 10 min read



Numerai walkthrough: Quantitative Analysis & Machine learning for fun and profit.

If you believe that action and practice are better than theory and have little to no experience in the field of machine learning and quantitative analysis (but a healthy interest) , this might be the post for you, for the rest it is a look into a probable future where one more thing could become decentralized and open sourced: ***Wall Street***.

There are 2 main ideas at play here, so I'll try giving you a birds eye view of both :

1. **Numerai:** Within the world of finance, hedge funds are a group of investors that plunk money into a basket of financial instruments and investments with hopes of having a return, a $3 Trillion Dollar competitive industry. Within this industry **Numerai** is one of the first experiments in open sourcing at some level the management and decision of buying and selling investments.
2. **Quantitavie Analysis & Machine Learning:** Quantitative analysis in finance employs statistical & matematical models to predict and evaluate the price and value of financial instruments, in a nutshell

various numbers representing some feature of the instrument go in, and a prediction or score comes out, said prediction or valuation then informs the decision making on the funds capital (buying,selling, holding etc) Machine learning which is everywhere these days is but the latest tool in making sense of the numbers; to a researcher like me or someone more interested in different A.I. problems like pattern recognition this type of analysis is just a new, fun and probably profitable application.

Numerai bridges both worlds and adds a crowdfunded twist by providing the raw data and asking you and me to provide a better estimate, these predictions are then used to control capital and the top performing ones get paid real money for their contribution, additionally there are incentives in the form of staking your own funds (in the form of numerais own cryptocurrency) which in essence is a reward and punishment system for keeping your predictions novel and useful. This is a complex but interesting part which I will sidestep for now to focus on getting started.

> *For the money, for the glory, and for the fun , but mostly for the money.*

In my experience it will take a novice data scientist, coder, or hobbyist a few weeks to get off the ground, at the end of this very short period you will have practical knowledge on applying various statistical and A.I. concepts to a real problem on a platform that you can then expand to your own interests, the competitive aspect and financial incentives can help you stay focused and motivated while learning although I should mention that making a dollar might take months.

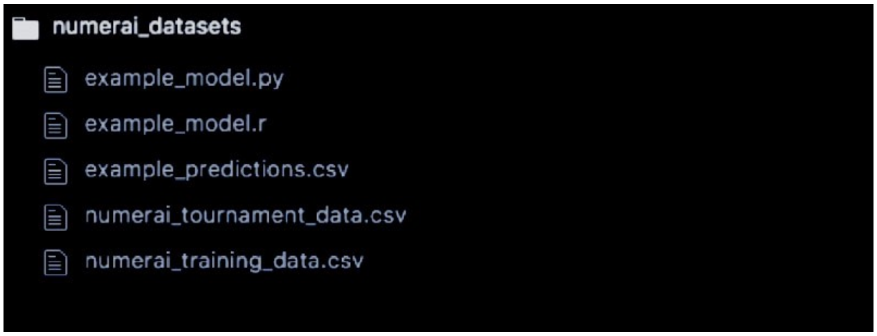
If you are not planning on participating but came for the whole open source wall street thing you can skip the following 2 sections.

Hands on:

This is but one path (the one I took) yours might be different depending on your starting point.

- Sign up to Numerai, read the documentation, get thoroughly intimidated by the documentation, the forums, the slack channel and the sinking feeling that even though you thought you were a smart cookie, this will require learning quite a few new things, take a deep breadth.

- Download the data set, it consists of multiple files which you will use, here's what the latest one looks like:

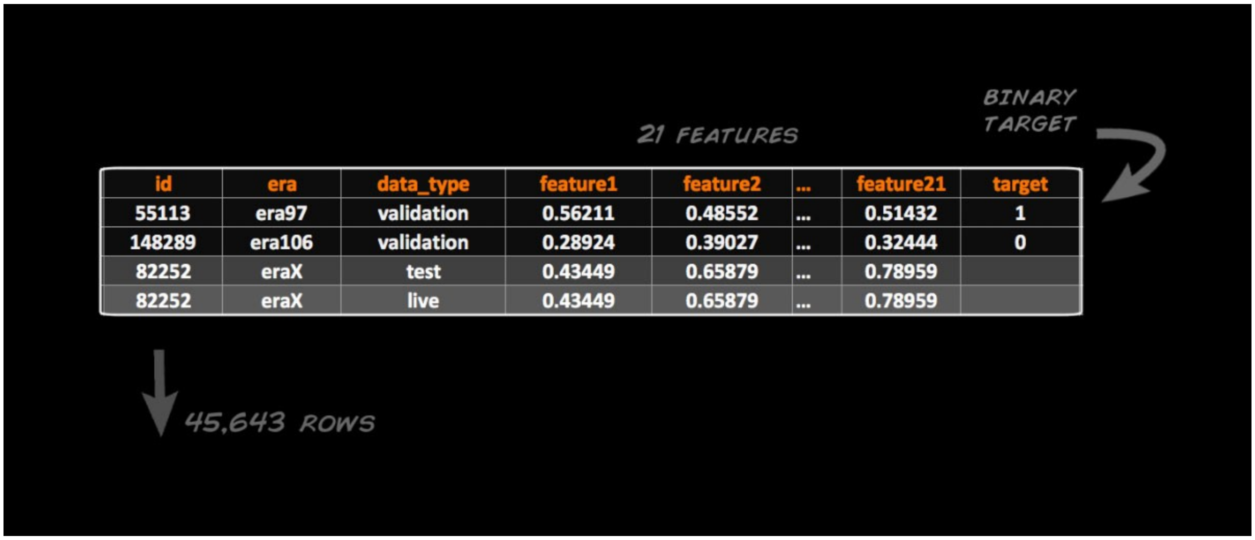


```
numerai_datasets
    example_model.py
    example_model.r
    example_predictions.csv
    numerai_tournament_data.csv
    numerai_training_data.csv
```

- These files might make perfect sense if you've done some data science, if not I'll try giving you a quick overview/introduction:

example_model.py & example_model.r: Numerai is kind enough to provide 2 fully functioning and working examples, these are perfectly good springboards and working examples of a base implementation in R and Python, 2 very popular languages for big data analysis and machine learning, I went with Python since I had some experience with it, although I foolishly tried making a neural net on excel at first, my computer is still recovering.

numerai_tournament_data.csv: There are a few things to this file, so let's start by looking at what's inside:



BINARY TARGET

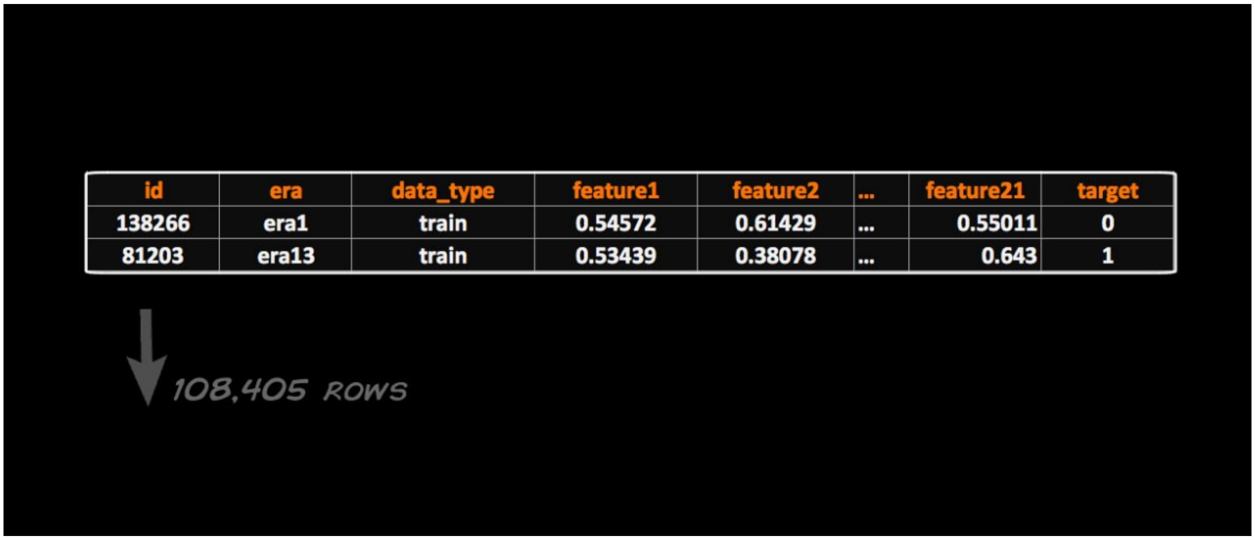
21 FEATURES

id	era	data_type	feature1	feature2	...	feature21	target
55113	era97	validation	0.56211	0.48552	...	0.51432	1
148289	era106	validation	0.28924	0.39027	...	0.32444	0
82252	eraX	test	0.43449	0.65879	...	0.78959	
82252	eraX	live	0.43449	0.65879	...	0.78959	

45,643 ROWS

Importantly there are various types of elements or categories (id's) divided into validation,live and test types, and each one has features, and some date information in the form of eras; the target column is what your model will predict in the form of a binary number (1 or 0). Numerai obfuscates the data, so even though we are assuming these are financial instruments (could be volume, open,close price, etc) there is nothing explicitly apparent, feature 11 on id 12546 could represent stock price at close date last week (era2) or something else, a dataset representing cancer rates or pixels in a dog recognition problem would look the same !

numerai_training_data.csv:

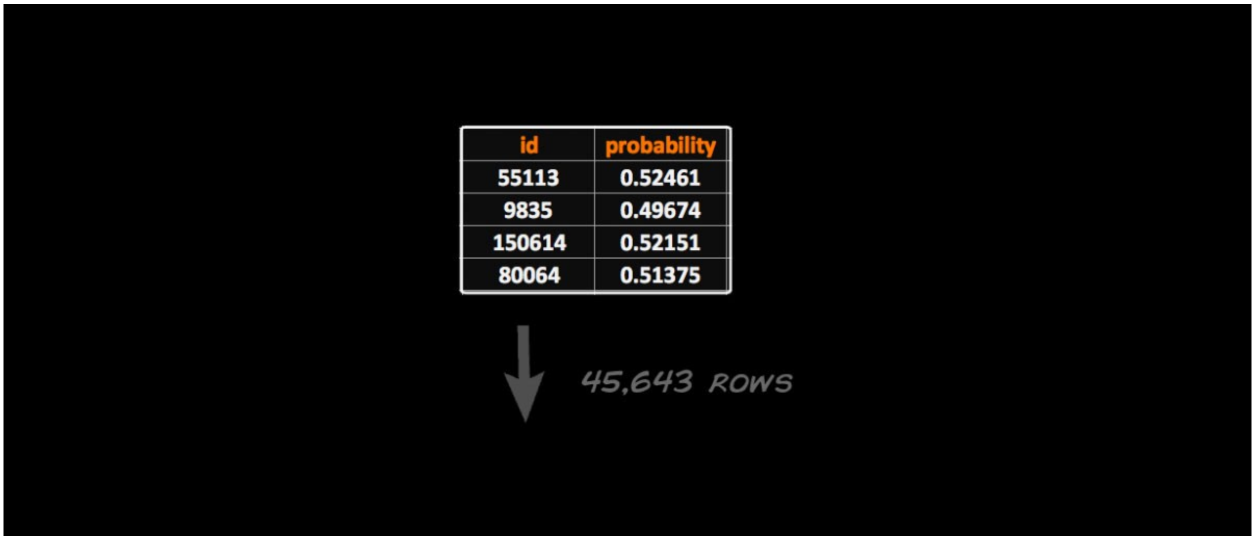


id	era	data_type	feature1	feature2	...	feature21	target
138266	era1	train	0.54572	0.61429	...	0.55011	0
81203	era13	train	0.53439	0.38078	...	0.643	1

108,405 ROWS

An additional dataset, the training set is also provided, if you are familiar with cross validation it will be obvious what this dataset represents and how it is used. In a nutshell, you are given a problems definition and it's solution, (*the training set*), you fit a model to it and then generalize the solution to new data (*the blank test and live rows on the tournament dataset*) and make your predictions. It is the equivalent of learning to ride a bicycle with training wheels (here you are the model) and then trying your skill (predicting) by riding on a bicycle without them, if you fall you have a bad model.

example_predictions.csv *:* This is an example of what you will be submitting, a column of id's (the same id's from numerai_tournament_data.csv) and a probabilistic prediction that the id will be of class 1, so while the target of your model is a binary number, it needs to output probabilities on said target.

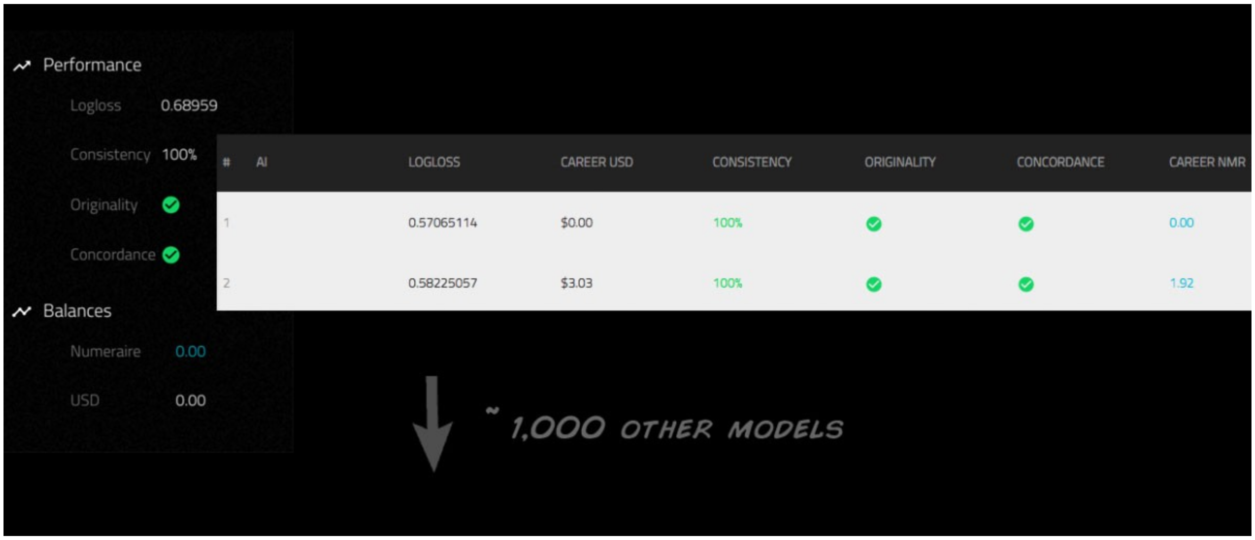


id	probability
55113	0.52461
9835	0.49674
150614	0.52151
80064	0.51375

- Once you have these files, you now need to set up your environment for making, predicting, keeping track of and submitting predictions. In my case it involved installing python plus setting up my editor (I use Atom with script for running python) and creating a few spreadsheets to test,tweak and keep track of predictions, this is time consuming and will probably take a few weeks if not months to streamline.
- **Models :** As mentioned, the example models provided by numerai are an excellent way to get started making predictions and

working with models, the code does everything for you (read files, divide data, define a model, train,predict, write to file) provided you have the right libraries installed, pandas and numpy for dealing with the data and sklearn for the statistical models for instance.

- **The Tournament :** Numerai has a weekly tournament and after each one the clock resets and a new dataset is provided, you can have 3 models submitting predictions and are limited to uploading 25 per day (it's usually plenty once you learn how to calculate logloss and other metrics on your own). They also provide an API if you choose to automate the download/upload.
- Once you have successfully submitted a model, you will be in the leaderboard and your predictions will be scored across various dimensions :



Performance

Logloss	0.68959
Consistency	100%
Originality	✓
Concordance	✓

Balances

Numeraire	0.00
USD	0.00

#	AI	LOGLOSS	CAREER USD	CONSISTENCY	ORIGINALITY	CONCORDANCE	CAREER NMR
1		0.57065114	$0.00	100%	✓	✓	0.00
2		0.58225057	$3.03	100%	✓	✓	1.92

At this point the algorithms for originality,concordance and consistency are not open source, but they are described as follows (*you also need to meet a specific treshold to qualify for the tournament and control capital*)

Consistency: Logloss across era, explanation forthcoming. (*75% an up*)

Originality : Somehow controversial since it might favor early submitters; a check to see if your predictions/model were already submitted. (a pass/ fail flag)

Concordance : A check to see if one model was applied to all data_types. (a pass/fail flag)

Once your model is prequalified by these metrics, you are then scored by your logloss against live data in order to actually win money, a process that takes 4 weeks, the scoreboard is then not a reflection of the actual place your model will place in the tournament. Numerai is not telling the specifics, but they hint at a meta model based on the top predictions, those closests in logloss to their own model and their live data, the time period also plays into favoring models and predictions over a longer time span, since logloss is the main ranking metric, it's important to know about it.

Logloss : (or logarithmic loss) is a graded statistical measure of confidence over your predictions, in real terms something like 0.50 would represent a random guess and 0.00 a perfectly confident prediction, it is graded because the more confident you are about a certain outcome (I will not fall of the bycicle) the more it penalizes you if in fact you fall of the bicycle. In the case Numerai it could be understood as I am certain to this degree about my predictions.

And that's it, once you have your setup ready you are welcome to try any of the multiple statistical models that are open source, and any technique and combination of parameters for analyzing this data, a good place to start is the sklearn library and all their models. I should also mention that ranking within the 100 or so models that are paid is difficult and requires a lot of experimenting, applying advanced techniques and in general getting to know the dataset, so using out of the box models will probably not get you very far.

Folks interested on the whole open source Wall Street thing continue reading here:

Beyond a great way to get into applied machine learning and quantitative analysis, Numerai might become something bigger and arguably better for the world of finance, it is early days and it is on the experimental side like many startups, but the upside potential is there. The hedge fund and banking industry have traditionally been elite,closed and secretive places with a bad reputation, in the same way that Hollywood and the TV networks were once the only source for movies and shows, Youtube has provided a new avenue for both consumers and content creators.

A distributed and open source wall street would allow diverse talent and ideas to contribute, a self taught analyst in a developing country that does not like wearing a suit and the heavy handed bro atmosphere of wall street could make a living and contribute to a hedge fund. A researcher in academia might not want to abandon his full time job or research, but might be able to contribute part time.

The consumers of the hedge fund industry as a whole would also benefit if the experiment proves successful, this bet hinges on the wisdom of the collective quant crowd vs the existing industry and the performance of these new types of funds. I am personally somehow optimistic, but I am also aware that consistent positive performance in the hedge fund industry (and other investments) usually favors an index approach vs stock picking and active management, so in the end a quant approach, even an open sourced one might not be all that profitable.

Finally, I believe the hedge fund industry will probably have nothing to fear,but will coexist peacefully and even contribute if the experiment resolves favorably, established banks might even one day foster this type of funds, competition wise, I also think there is room for everyone since there is plenty of untapped investment capital facilitated by the new ecosystems (cryptocurrencies,globalization, the web) which numerai an other similar ventures might be better positioned to capture; if there is a point of concern for the hedge fund industry, is that of cost, the current 2 and 20 fee structure (2 percent on assets and 20% on profits) is pricey for 2 reasons, the cost of having an infrastructure (brokers, analysts, researchers, back office, etc) and the walled garden of finance that limits participation, it is in this space that a desentralized open source hedge fund can lower costs and compete, so as mentioned we could very well see a new more competitive type of hedge fund appear and the existing ones would need to adapt.

Parting Words

I left a few things out for lack of space and because I wanted to focus this short piece as a both an alternative way of onboarding, which I struggled with and to discuss albeit briefly what this type of fund could herald. You can read about staking via this whitepaper, and keep up with numerai at their blog.

I should also mention kaggle and other writings on numerai, machine learning and quantitative analysis here on medium as a way to start and further your learning (as well as mine).

Small Update:

(Oct 1st 2017)The dataset has been changed, there is now roughly twice the features and 10–20 K more rows, this means that if you are on basic hardware (I am on a i3 4GB RAM HD windows machine) you will most likely experience memory errors on some models, the solution is to subsample. Also, I got my first NMR by ranking within the first 100 !

Medium sized Update:

(Dec 25 2017) Numerai has now released the source for scoring predictions and streamlined the API for automatic submissions along with some other minor changes and a vision statement & road map, read more here : Numerai's Master Plan. Most importantly, payouts have now changed in nature (more NMR, less usd). I personally have been doing great ranking in the top 50 – 60 more or less consistently.

Thanks,

Keno

About the Author :

Born Eugenio Noyola Leon (Keno) I am a Designer,Web Developer/programmer, Artist and Inventor, currently living in Mexico City, you can find me at www.k3no.com