O'REILLY®

ON OUR RADAR AI BUSINESS DATA DESIGN ECONOMY OPERATIONS SEE ALL

DATA SHOW PODCAST + FOLLOW THIS TOPIC

Programming collective intelligence for financial trading

The O'Reilly Data Show Podcast: Geoffrey Bradway on building a trading system that synthesizes many different models.

By Ben Lorica. June 15, 2017

Tornado drawing. (source: meredithsteele on Flickr).

Subscribe to the O'Reilly Data Show Podcast to explore the opportunities and techniques driving big data, data science, and AI. Find us on Stitcher, TuneIn, iTunes, SoundCloud, RSS.

In this episode of the Data Show, I spoke with Geoffrey Bradway, VP of engineering at Numerai, a new hedge fund that relies on contributions of external data scientists. The company hosts regular competitions where data scientists submit machine learning models for classification tasks. The most promising submissions are then added to an ensemble of models that the company uses to trade in real-world financial markets.

To minimize model redundancy, Numerai filters out entries that produce signals that are already well-covered by existing models in their ensemble. The company also plans to use (Ethereum) blockchain technology to develop an incentive system to reward models that do well on live data (not ones that overfit and do well on historical data).

Here are some highlights from our conversation:

Coordinating data science and AI in finance

> At Numerai, we believe there are other people in the world who are better data scientists than we are, but we have the financial backgrounds. So, we can take very good financial data and actually encrypt it in such a way that the structure is preserved enough to do machine learning on it, but you can't tell what it is. Because we can do that, we can actually release our data set in data science tournaments, and then users can download our data, which just looks like a giant CSV with a bunch of features and targets. Then they can train their own models, try to predict what will happen in the future, and upload that to our

> website.
>
> We're trying to set it up in such a way where our users don't have to know a lot about finance, and they can just be very, very good data scientists. They can leave much of the financial data munging up to us.
>
> ... A big problem we were running into is that we were sort of paying off users based on how well they did on a backtest. The problem with that is, you can overfit to your backtests. You do a model; it scores you; it tells you how well you do. Then you slightly tweak your model, it scores you better, and you can keep doing that until you get really, really good on a backtest. But that just destroys your ability to generalize into the future.
>
> ... What we wanted to do is create a mechanism that actually makes it irrational for users to want to overfit. That took the form of a cryptocurrency that we call the Numeraire. The idea is, you get this token that has some value, and you can use that to essentially stake your predictions. So, you can say: ‘Hey, I think these predictions are really good.’ If your predictions do turn out to be good, in the sense that they perform better than random on live data, we'll give you your tokens back, and we'll give you some additional payout.

Applications of differential privacy and adaptive data analysis

> This business model has many potential applications, basically, in any setting where the data is very, very sensitive. This could be finance data, this could be health care data, but this could also just be internal corporate data. For example, Amazon wouldn't want to necessarily open source their logistics data, because that's very, very valuable for them. That's something that gives them an edge; so, anything where it would be fantastic to have models from many sources, but sharing the data is hard.
>
> There is a related field in computer science called differential privacy. It is a

> subfield that talks about how you release a data set so that nobody can mine it for sensitive information. And then there's a related field that has to do with adaptive data analysis, where you run a model, you get feedback, you run another model based off of that feedback, and you keep doing that. How can you make those statistically sound?

Related resources:

- Findata Day returns to the Strata Data conference in New York City, September 25-28, 2017.
- What Kaggle has learned from almost a million data scientists: Strata Data conference keynote by Kaggle co-founder, Anthony Goldbloom
- Data preparation in the age of deep learning: featuring Crowdflower co-founder Lukas Biewald
- Building human-assisted AI applications

Share Tweet Share 31 Share 120

Ben Lorica

Ben Lorica is the Chief Data Scientist at O'Reilly Media, Inc. and is the Program Director of both the Strata Data Conference and the O'Reilly Artificial Intelligence Conference. He has applied Business Intelligence, Data Mining, Machine Learning and Statistical Analysis in a variety of settings including Direct Marketing, Consumer and Market Research, Targeted Advertising, Text Mining, and Financial Engineering. His background includes stints with an investment management company, internet startups, and financial services.

more

DATA SHOW PODCAST

Turning PhDs into industrial data scientists

By Ben Lorica

Angie Ma's startup, London-based ASI, runs a carefully structured "finishing school" for science and engineering doctorates.

DATA SHOW PODCAST

Topic models: Past, present, and future

By Ben Lorica

David Blei, co-creator of one of the most popular tools in text mining and machine learning, discusses the origins and applications of topic models.

DATA SHOW PODCAST

Using Agile development techniques for data science projects

By Ben Lorica

The O'Reilly Data Show Podcast: John Akred on building data platforms and enterprise data strategies.

DATA SHOW PODCAST

Building enterprise data applications with open source components

By Ben Lorica

The O'Reilly Data Show podcast: Dean Wampler on bounded and unbounded data processing and analytics.

ABOUT US

Our Company

Teach/Speak/Write

Careers

Customer Service

Contact Us

SITE MAP

Ideas

Learning

Topics

All