

No Free Hunch (<http://blog.kaggle.com/>)



🏠 ([HTTP://BLOG.KAGGLE.COM](http://blog.kaggle.com/)) > MARCH MACHINE LEARNING MANIA, 4TH PLACE WINNER'S INTERVIEW: ERIK FORSETH

⬅️ ([HTTP://BLOG.KAGGLE.COM/2017/05/09/DSTL-SATELLITE-IMAGERY-COMPETITION-3RD-PLACE-WINNERS-INTERVIEW-VLADIMIR-SERGEY/](http://blog.kaggle.com/2017/05/09/DSTL-SATELLITE-IMAGERY-COMPETITION-3RD-PLACE-WINNERS-INTERVIEW-VLADIMIR-SERGEY/)) ➡️ ([HTTP://BLOG.KAGGLE.COM/2017/05/01/DATASETS-OF-THE-WEEK-APRIL-2017/](http://blog.kaggle.com/2017/05/01/DATASETS-OF-THE-WEEK-APRIL-2017/))

March Machine Learning Mania, 4th Place Winner's Interview: Erik Forseth

Kaggle Team (<http://blog.kaggle.com/author/kaggleteam/>) | 05.05.2017



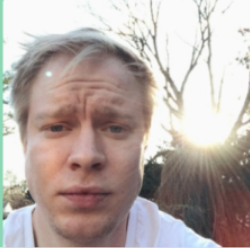
The annual [March Machine Learning Mania competition \(https://www.kaggle.com/c/march-machine-learning-mania-2017\)](https://www.kaggle.com/c/march-machine-learning-mania-2017), which ran on Kaggle from February to April, challenged Kagglers to predict the outcome of the 2017 NCAA men's basketball tournament. Unlike your typical bracket, competitors relied on historical data to call the winners of all possible team match-ups. In this winner's interview, Kaggle [Erik Forseth \(https://www.kaggle.com/errofo\)](https://www.kaggle.com/errofo) explains how he came in fourth place using a combination of logistic regression, neural networks, and a little luck.



The basics

What was your background prior to entering this challenge?

My background is in theoretical physics. For my PhD I worked on understanding the orbital dynamics of gravitational wave sources. While that work involved a healthy balance of computer programming and applied math, there wasn't really any statistical component to it. In my spare time, I got interested in machine learning about two years prior to finishing my degree.




Erik Forseth

United States

Joined 5 months ago · last seen 3 days ago

in



Competitions Novice

Home

Competitions (1)

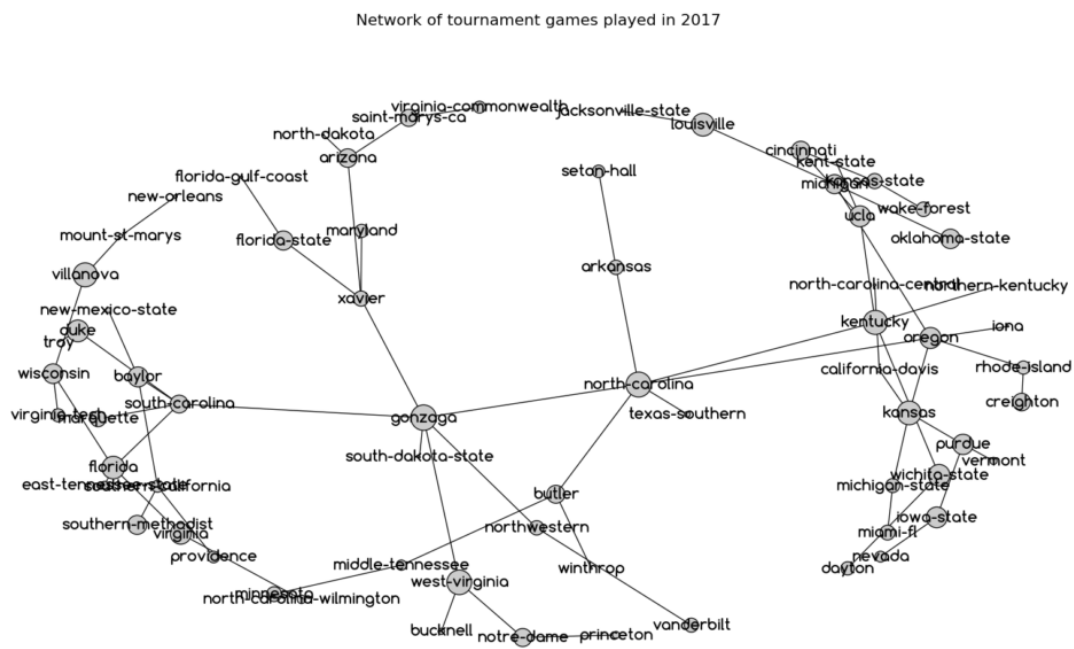
Contact User

Do you have any prior experience or domain knowledge that helped you succeed in this competition?

As a matter of fact, sports prediction – college basketball prediction in particular – has been a hobby of mine for several years now. I have a few models which are indefinite works in progress, and which I run throughout the season to predict the outcomes of games. So on one hand, entering the competition was a no-brainer. That said, March Madness is a bit of a different beast, being a small sample of games played on neutral courts by nervous kids.



Let’s get technical:



My 4th-place entry used the combined predictions of two distinct models, which I’ll describe in turn.

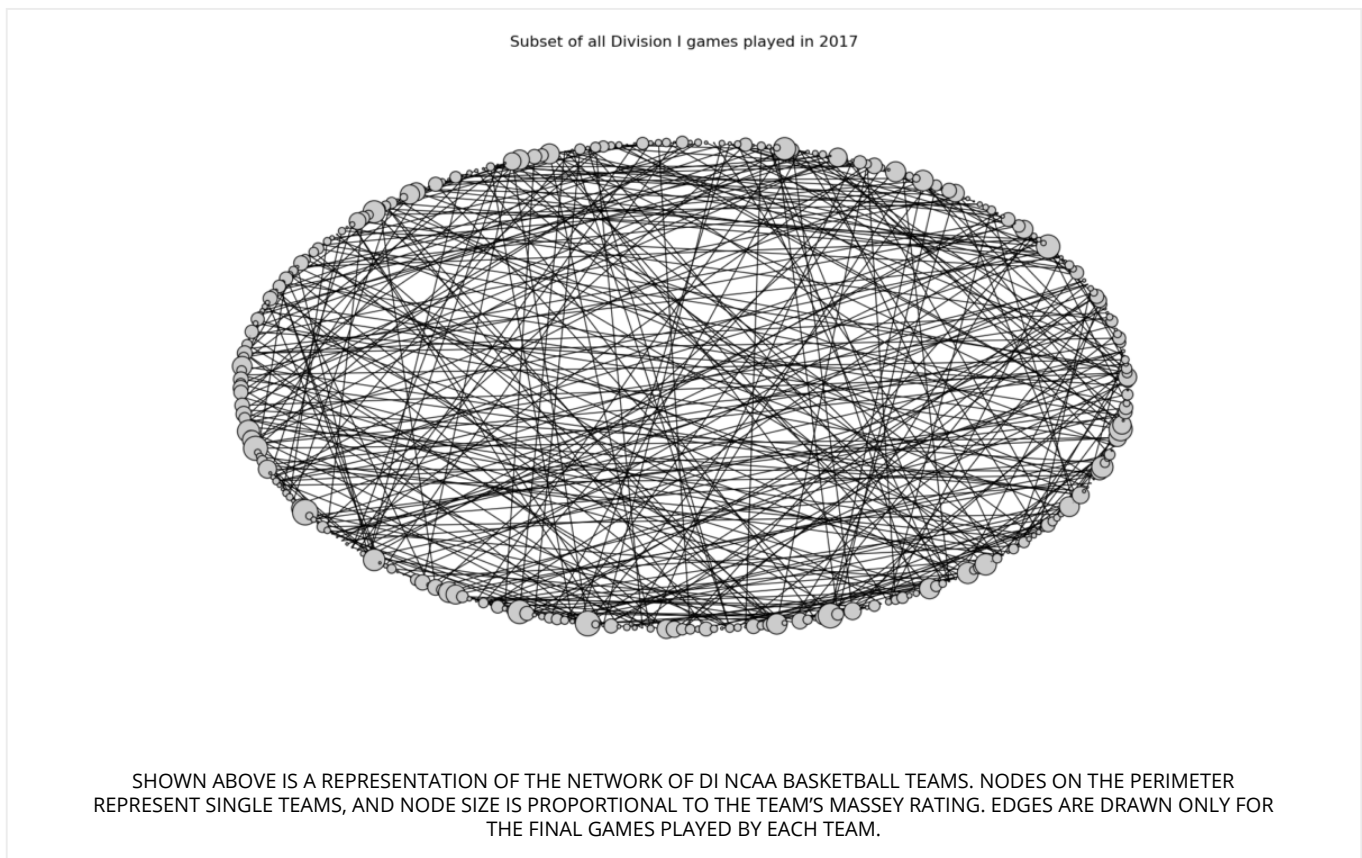
Model 1

Needless to say, there's a rich history and a large body of work on the subject of rating sports teams. Common to most good ratings is some notion of "strength of schedule;" whether implicitly or explicitly, the rating ought to adjust for the quality of opponents that each team has faced.

Consider for example the Massey (<http://www.masseyratings.com/>) approach. Each team is assigned a rating r , where the difference between teams' ratings purports to give the observed point differentials (margins of victory) m in a contest between the two teams. So, one constructs a system of equations of the form:

$$r_i - r_j = m_{ij}$$

for teams i and j and outcomes m_{ij} , and then solves for the ratings via least squares. The Massey ratings are a compact way of encoding some of the structure of the network of the roughly 350 Division I teams. They take into account who has beaten who, and by how much, for all games played by all teams.



And so, my first model was a straightforward logistic regression, taking a set of modified Massey ratings as input. My modified ratings differ from the original versions in that: (1) I augment the system of equations to include games played during the prior season, and (2) I rescale the system in such a way that recent games are given greater weight. These are all relatively straightforward computations done in Python using NumPy.

Model 2

The second model is a neural network trained on raw data instead of derived team strength ratings. I'm building these in Python with Theano. As far as I'm aware, most everyone who does sports prediction uses linear models based on team ratings of various flavors (Massey, ELO, RPI, etc., see above), and there hasn't really been a compelling reason to do anything fancier than that. So, I've been interested in the question of whether or not I can get something as good or better using a different approach entirely. One of the main challenges for me here has been to figure out how to present the model with raw data in such a way that it can build useful features, while at the same time keeping memory under control (I'm confined to training these on my laptop for the time being). This is still very much a work in progress, as I'm continually playing with the input data, the architecture, etc. Nevertheless, prior to the competition I managed to get something which performed as well as my latest-greatest linear models, and so in the end I averaged the predictions from the two.

Finally, my 4th-place entry involved a bit of "gambling." As I pointed out earlier, 63 games is a really small sample, and to make matters worse, you're being scored on binary outcomes. You could get a little more resolution on the entries if Kaggle instead posed a regression problem, where competitors might be asked to predict point differentials and were then scored according to mean-squared-error. Or, you could even have competitors predict point differentials *and* point totals, equivalent to predicting the individual scores of each team in each matchup.

Regardless, the current formulation of the contest is interesting, because it requires a certain amount of strategy that it might not otherwise have if the only goal were to come up with the most accurate classifier on an arbitrarily large number of games. In this case, my strategy was:

- There are only 63 games here....
- I'm rewarded for being correct, but punished for being wrong.
- Nevertheless, I believe in my underlying models, so I'm going to "bet" that their predictions tend to be right.
- Therefore, I will take all of the predictions, and push those above 0.5 toward 1, and those below 0.5 toward 0. (I came up with a hand-wavy rule of thumb for pushing those near the extremes more than I pushed those near the middle. In other words, I wasn't so sure about the predictions near 0.5, so I wanted to more or less leave those alone, whereas I wanted to get the most out of my stronger picks.)

What was the final effect of perturbing my predictions this way? I submitted the unperturbed predictions as my second entry, and it would've placed about 25th on the leaderboard, or still close to top 5%. I think this all goes to show that pure luck and randomness play a big role in this competition, but that there is headway to be made with a good model and a sound strategy.



Words of wisdom:

Do you have any advice for those just getting started in data science?

My advice to anyone with an interest in data science is to give yourself a project you're interested in. Rather than setting out and trying to learn about a specific method, pose an interesting problem to yourself and figure out how to solve it. Or, if you're absolutely intent on learning more about some particular toolbox, at least give yourself some interesting context within which you can apply those tools. To that end, writing yourself a web scraper can vastly increase your ability to get usable data to play around with.



Just for fun:

If you could run a Kaggle competition, what problem would you want to pose to other Kagglers?

I don't have a specific problem I'd pose; it's neat to see the various challenges that crop up. Though, we've seen a lot of image recognition tasks recently. I think it would be interesting to have more time series prediction and sequence classification problems.

[MARCH MANIA \(HTTP://BLOG.KAGGLE.COM/TAG/MARCH-MANIA/\)](http://blog.kaggle.com/tag/march-mania/)

[SPORTS ANALYTICS \(HTTP://BLOG.KAGGLE.COM/TAG/SPORTS-ANALYTICS/\)](http://blog.kaggle.com/tag/sports-analytics/)

 **BUFFERED VPN**

Customer Service
like no other VPN

GET STARTED
30-DAY MONEY-BACK

Voted as
No. 1 on

BestVPN.com



Revitalize Your Body
& Improve Your Health!





0 Comments

No Free Hunch


Sorted

1 Login

Recommend 1

Share





Sort by Best



Start the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS ?



Be the first to comment.

Omaha, NE's NEW RULE

Drivers With No Tickets In 3 Years Are In For A Big Surprise

[Learn More](#)

Sponsored by Comparisons.org





(<https://www.facebook.com/kaggle>)



(<https://twitter.com/kaggle>)

