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

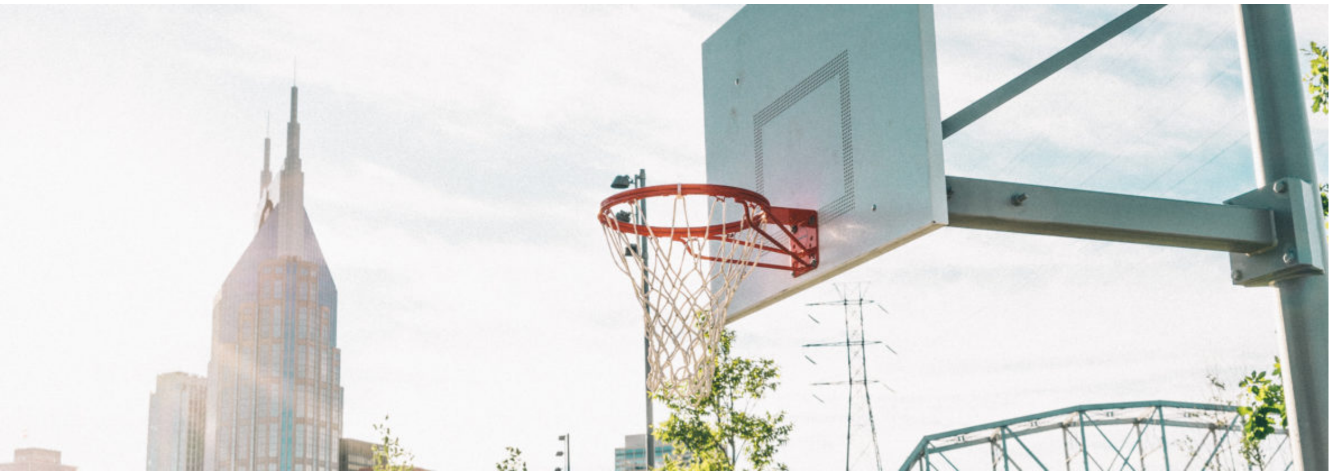
(HTTP://BLOG.KAGGLE.COM) > MARCH MACHINE LEARNING MANIA, 5TH PLACE WINNER'S INTERVIEW: DAVID SCOTT

(HTTP://BLOG.KAGGLE.COM/2017/05/25/TWO-SIGMA-FINANCIAL-MODELING-CHALLENGE-WINNERS-INTERVIEW-2ND-PLACE-NIMA-SHAHBAZI-CHAHHOU-MOHAMED/) (HTTP://BLOG.KAGGLE.COM/2017/05/19/MARCH-MACHINE-LEARNING-MANIA-1ST-PLACE-WINNERS-INTERVIEW-ANDREW-LANDGRAF/)

2 (http://blog.kaggle.com/2017/05/23/march-machine-learning-mania-5th-place-winners-interview-david-scott/#

March Machine Learning Mania, 5th Place Winner's Interview: David Scott

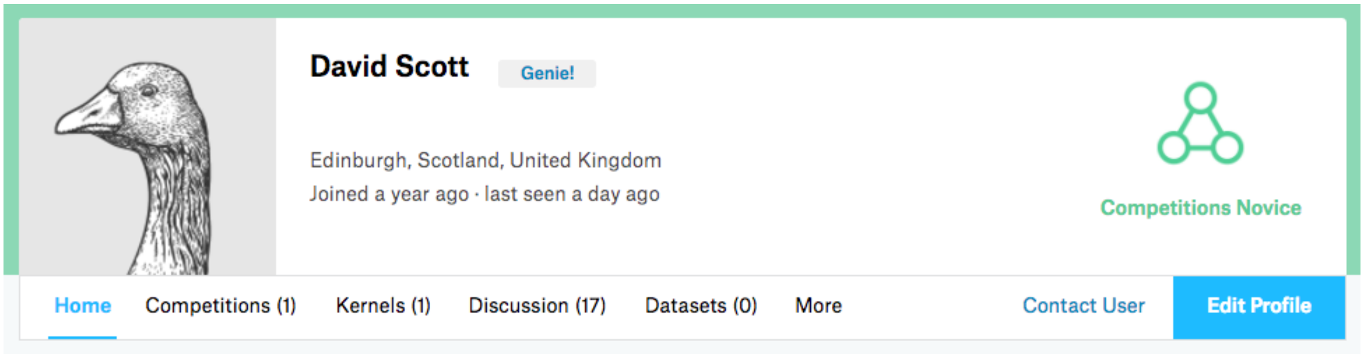
Kaggle Team (http://blog.kaggle.com/author/kaggleteam/) | 05.23.2017



Kaggle's annual March Machine Learning Mania competition (https://www.kaggle.com/c/march-machine-learning-mania-2017) drew 442 teams to predict the outcomes of the 2017 NCAA Men's Basketball tournament. In this winner's interview, Kaggler David Scott describes how he came in 5th place by stepping back from solution mode and taking the time to plan out his approach to the the project methodically.

The basics:

What was your background prior to entering this challenge?



David Scott Genie!

Edinburgh, Scotland, United Kingdom
Joined a year ago · last seen a day ago

Competitions Novice

I have been working in credit risk model development in the banking industry for approximately 10 years. It isn't a massive stretch from my original degree in Actuarial Mathematics and Statistics.

I have been lucky to receive exposure to big data and data science through previous roles but decided I wanted to teach myself R and to improve my machine learning knowledge. The best opportunity seemed to be to utilise Kaggle datasets where I could to help with this.

What made you decide to enter this competition?

I had started using some of the titanic data to learn some R but I am a massive sports nut. So when the opportunity came up to get data to predict March Madness I couldn't resist. This was my first entry in a Kaggle competition outside of the training exercises.

Let's get technical:

How did you approach the problem?

At first I dived into the data. As I played and wasn't achieving the success I was hoping for, I had a realisation. How would I approach this problem at work? I had rushed into solution mode without planning out the project and hadn't considered what I had expected to see. At that point I realised I had to consider 3 important things.

1. Finding out what the experts reviewed to predict March Madness (Experts).
2. Be careful structuring my data to make sure I didn't over fit (Data Construct).
3. Figure out what model development technique I would use for my final model (Model Development).

Experts - Did any past research or previous competitions inform your approach?

No. It would have been a good idea but instead I started listing to podcasts on college basketball. This gave me an understanding of what the commentators use when they are evaluating a good team and looked to make sure this information was included in my final model.

Data Construct - What pre-processing and feature engineering did you do?

I spent most of my time creating a linear model predicting the best teams based on their regular season results using the points' difference as the target variable. This gave me a rank order of teams that was my main predictor in my model. Outside of that my time was spent matching data from other sites to get things like Strength Of Schedule, etc.

I also made sure to split the data into enough segments that the model would not overfit. This included splitting the development data into a build and validation sample and leaving the test data provided for the last 4 years. Each change in the development was evaluated for consistency with the others.

What supervised learning methods did you use?

I kept it simple with a logistic regression. This is something I am very familiar with and something I thought it would work well for this problem.

Words of wisdom:

What have you taken away from this competition?

The main take away from this competition was that data and how you use it was more important than the modelling technique. I stuck with a basic logistic regression technique for the model development and it appeared to work well.

Looking back, what would you do differently now?

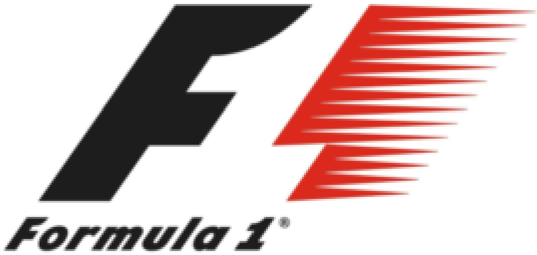
I would have factored in that games at the later stages of the tournament would be close. I ran out of time to consider that if teams met in the final 4 regardless of their rating before the tournament it is likely to be close. This meant that 1 upset towards the end of the tournament could have derailed my finishing position.

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

I don't really have advice for getting started in data science but I would suggest having a go at the problem datasets in Kaggle. In the competitions I would suggest taking time to think about the problem and plan in advance before diving in. If you find out what the experts consider to be useful, chances are you won't be far off.

Just for fun:

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



I have more of a problem I would like to pose to other Kagglers. I am a massive F1 fan and I have always been interested in understanding what the car contributes and what the driver contributes. This way I could figure out the best driver of all time. I have a source of data that I would be happy post on Kaggle for some assistance if people are interested.

What is your dream job?

I am very fortunate that I really enjoy understanding what can be used to predict different events and learning new techniques as I go. My career to date has allowed me lots of opportunities to do this.

2 Comments No Free Hunch 1 Login

♡ Recommend 1 Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Jacob Cupul • 9 months ago

Awesome, I like your practical approach. Please post the F1 dataset! You have at least one person who would be interested in exploring it...

1 ^ | ˅ • Reply • Share ›

Apoorv Jagtap • 9 months ago

Nice, please post the dataset, I am sure the community here would be more than willing to help improve it in any way possible.

^ | ˅ • Reply • Share ›