'Specialist knowledge is useless and unhelpful'

By Peter Aldhous

Interview with Jeremy Howard, *New Scientist* #2893, 7 December 2012

Kaggle.com has turned data prediction into sport. People competing to solve problems are outclassing the specialists, says its president Jeremy Howard

Kaggle has been described as "an online marketplace for brains". Tell me about it.
It's a website that hosts competitions for data prediction. We've run a whole bunch of amazing competitions. One asked competitors to develop algorithms to mark students' essays. One that finished recently challenged competitors to develop a gesture-learning system for the Microsoft Kinect. The idea was to show the controller a gesture just once, and the algorithm would recognise it in future. Another competition predicted the biological properties of small molecules being screened as potential drugs.

How exactly do these competitions work?
They rely on techniques like data mining and machine learning to predict future trends from current data. Companies, governments and researchers present data sets and problems, and offer prize money for the best solutions. Anyone can enter: we have nearly 64,000 registered users. We've discovered that creative data scientists can solve problems in every field better than experts in those fields can.

These competitions deal with very specialised subjects. Do experts enter?
Oh yes. Every time a new competition comes out, the experts say: "We've built a whole industry around this. We know the answers." And after a couple of weeks, they get blown out of the water.

So who does well in the competitions?
People who can just see what the data is actually telling them, without being distracted by industry assumptions or specialist knowledge. Jason Tigg, who runs a pretty big hedge fund in London, has done well again and again. So has Xavier Conort, who runs a predictive analytics consultancy in Singapore.

You were once on the leader board yourself. How did you get involved?
It was a long and strange path. I majored in philosophy in Australia, worked in management consultancy for eight years, and then in 1999 I founded two start-ups - one an email company, the other helping insurers optimise risks and profits. By 2010, I had sold them both. I started learning Chinese, and building amplifiers and speakers because I hadn't made anything with my hands. I travelled. But it wasn't intellectually challenging enough. Then, at a meeting of statistics users in Melbourne, somebody told me about Kaggle. I thought: "That looks intimidating and really interesting."

How did your first competition go?
Setting my expectations low, my goal was to not come last. But I actually won it. It was on forecasting tourist arrivals and departures at different destinations. By the time I went to the next statistics meeting I had won two out of the three competitions I entered. Anthony Goldbloom, the founder of Kaggle, was there. He said: "You're not Jeremy Howard, are you? We've never had anybody win two out of three competitions before."

How did you become Kaggle's chief scientist?
I offered to become an angel investor. But I just couldn't keep my hands off the business. I told Anthony that the site was running slowly and rewrote all the code from scratch. Then Anthony and I spent three months in America last year, trying to raise money. That was where things got really serious, because we raised $11 million. I had to move to San Francisco and commit to doing this full-time.

Do you still compete?
I am allowed to compete, but I can't win prizes. In practice, I've been too busy.

What explains Kaggle's success in solving problems in predictive analytics?
The competitive aspect is important. The more people who take part in these competitions, the better they get at predictive modelling. There is no other place in the world I'm aware of, outside professional sport, where you get such raw, harsh, unfettered feedback about how well you're doing. It's clear what's working and what's not. It's a kind of evolutionary process, accelerating the survival of the fittest, and we're watching it happen right in front of us. More and more, our top competitors are also teaming up with each other.

Which statistical methods work best?
One that crops up again and again is called the random forest. This takes multiple small random samples of the data and makes a "decision tree" for each one, which branches according to the questions asked about the data. Each tree, by itself, has little predictive power. But take an "average" of all of them, and you end up with a powerful model. It's a totally black-box, brainless approach. You don't have to think - it just works.

What separates the winners from the also-rans?
The difference between the good participants and the bad is the information they feed to the algorithms. You have to decide what to abstract from the data. Winners of Kaggle competitions tend to be curious and creative people. They come up with a dozen totally new ways to think about the problem. The nice thing about algorithms like the random forest is that you can chuck as many crazy ideas at them as you like, and the algorithms figure out which ones work.

That sounds very different from the traditional approach to building predictive models. How have experts reacted?
The messages are uncomfortable for a lot of people. It's controversial because we're telling them: "Your decades of specialist knowledge are not only useless, they're actually unhelpful; your sophisticated techniques are worse than generic methods." It's difficult for people who are used to that old type of science. They spend so much time discussing whether an idea makes sense. They check the visualisations and noodle over it. That is all actively unhelpful.

Is there any role for expert knowledge?
Some kinds of experts are required early on, for when you're trying to work out what problem you're trying to solve. The expertise you need is strategy expertise in answering these questions.

Can you see any downsides to the data-driven, black-box approach that dominates on Kaggle?
Some people take the view that you don't end up with a richer understanding of the problem. But that's just not true: the algorithms tell you what's important and what's not. You might ask why those things are important, but I think that's less interesting. You end up with a predictive model that works. There's not too much to argue about there.

Profile
When Jeremy Howard graduated in philosophy from the University of Melbourne, Australia, he was already working as a management consultant for McKinsey & Company. Later he founded email company FastMail and the Optimal Decisions Group, which helps insurance companies set premiums. He is now president and chief scientist of Kaggle, San Francisco.