Anthony Goldbloom gives you the secret to winning Kaggle competitions

January 13, 2016 Andrew Fogg Big Data

Kaggle has become the premier Data Science competition where the best and the brightest turn out in droves – Kaggle has more than 400,000 users – to try and claim the glory. With so many Data Scientists vying to win each competition (around 100,000 entries/month), prospective entrants can use all the tips they can get.

And who better than Kaggle CEO and Founder, Anthony Goldbloom, to dish out that advice? We caught up with him at Extract SF 2015 in October to pick his brain about how best to approach a Kaggle competition.

The only 2 winning approaches

According to Anthony, in the history of Kaggle competitions, there are only two Machine Learning approaches that win competitions: Handcrafted & Neural Networks.

Well, that should make things simple...

Handcrafted feature engineering

This approach works best if you already have an intuition as to what's in the data. First, a competitor will take the data and plot histograms and such to explore what's in it. Then they'll spend a lot time generating features and testing which ones really do correlate with the target variable.

For example, a chain of used car dealers wanted to predict which cars sold at a second-hand auction would be good buys and which ones would be lemons. After much trial and error, by many different applicants, it turned out that one of the most predictive features was the car color. By grouping standard color cars and unreliable colored cars, they found that unusual colored cars were more likely to be reliable.

The way they found this answer was to test lots and lots and lots of hypotheses. The vast majority of them didn't work out, but the one that did won them the competition.

As long as Kaggle has been around, Anthony says, it has almost always been ensembles of decision trees that have won competitions.

It used to be random forest that was the big winner, but over the last six months a new algorithm called XGboost has cropped up, and it's winning practically every competition in the structured data category.

Neural Networks and Deep Learning

For any dataset that contains images or speech problems, deep learning is the way to go. The people who are winning these competitions (the ones without well-structured data) are spending almost none of their time doing feature engineering. Instead, they spend their time constructing neural networks.

For example, one Kaggle competition asked participants to take images of the eye and diagnose which ones had diabetic retinopathy (one of the leading causes of blindness). Incredibly, the algorithm that won had the same agreement rate with an ophthalmologist (85%) as one ophthalmologist has with another.

Which method should you use?

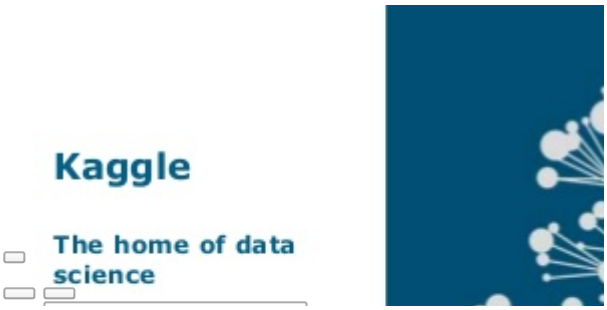
So, faced with a Kaggle competition, how should you spend your time? Should you do a lot of testing on which features affect the outcome? Or should you spend all your time building and training neural networks.

For most competitions it's pretty obvious. If you have lots of structured data, the handcrafted approach is your best bet, and it you have unusual or unstructured data your efforts are best spent on neural networks.

But what about datasets that fall somewhere in the middle?

One such competition that internal Kaggle employees weren't sure of initially asked Kaggle users to take EEG readings and determine whether someone was grasping or lifting.

The answer? Neural networks!



[contentblock id=6 img=gcb.png]

How Kaggle competitions work

Companies come to Kaggle with a load of data and a question. For example, GE might come to them with a load of data about heat and vibration and ask their users to help predict when an airplane is going to fail.

As part of the problem, the company would provide a set of training data where the outcome you are trying to predict is known to both them and the Kaggle competitor. They also provide a test dataset where the outcome competitors are trying to predict is known only to the company. It's how companies know how accurate your machine learning model is.

As competitors upload their algorithms, Kaggle shows them in real time how they are doing in relation to the other competitors. A competitor can upload up to 5 entries in a day and typically competitions last for around 2 months.