San Francisco startup makes data science a sport

April 15 2012, MARCUS WOHLSEN , Associated Press

In this Thursday, April 12, 2012 photo, Kaggle's president and chief scientist Jeremy Howard poses near a water tower in Mountain View, Calif. The founders of San Francisco startup Kaggle believe the problems data scientists solve are so important that they should be paid like professional athletes. By turning data science into a crowd-sourced contest, they hope they have created a way to make that happen. Kaggle hopes top contenders will participate in a sport tailor-made for the 21st century: Competitive number-crunching. (AP Photo/Paul Sakuma)

(AP) -- Strange secrets hide in numbers. For instance, an orange used car is least likely to be a lemon. This particular unexpected finding came to light courtesy of a data jockey who goes by the Internet alias

SirGuessalot, who in fact wasn't guessing at all. Instead, he and his partner, PlanetThanet, relied on the hard math skills that make them top contenders in a sport tailor-made for the 21st century: competitive number-crunching.

The used car defect prediction contest is one of dozens hosted by San Francisco online startup Kaggle, whose creators believe they can tap the global geek population's instinct for one-upmanship to mine better answers faster from the world's ever-rising mountain of data.

"Competitions bring together a wide variety of people into a wide variety of problems," said Jeremy Howard, who became Kaggle's president and chief scientist after winning multiple competitions himself. "You get people looking at stuff they'd never look at otherwise."

While the used car contest was fun, Kaggle has its eye on weightier scientific problems. In one contest, an English major who trained himself in data science built a model for predicting the progress of HIV infections in individual patients. In another, a scientist who studies glaciers for a living won a NASA-backed Kaggle competition to measure the shapes of galaxies by mapping the universe's dark matter.

The data problems that need solving are so important that those who find the solutions should be paid like professional athletes, said Kaggle founder Anthony Goldbloom. By turning data-mining into a crowdsourced contest, he hopes he's created a way to make that happen. Already one of Kaggle's contests offers a multimillion dollar prize.

"We want to see the best data scientists earning more than Tiger Woods," said Goldbloom, who started the company in his native Australia and recently came to San Francisco's South of Market startup haven.

The job market for mathematicians and statisticians has become hot as the sheer volume of data generated by ever faster, cheaper computing resources explodes.

Data storage has become so inexpensive that a 2011 McKinsey and Co. report estimated that a disk drive capable of storing all the world's music would cost about $600. Walmart stores 10 times more data on customer transactions and other parts of its operation than is contained in the entire Library of Congress, according to the same report.

Analyzing the so-called "big data" deluge has become a key task for businesses in an effort to divine everything from which ads online customers will click to how much inventory they need to maintain. Political candidates analyze data to predict voting patterns. Dating websites try to predict ideal mates.

Kaggle competitions focus on creating and testing formulas that can be used to make predictions based on the contents of giant datasets.

The more accurate the formula, the better the chances it will accurately provide answers to complex questions, such as the orange used car being the least likely to break down.

Goldbloom argues that no matter how many data scientists companies hire, relying on in-house data talent means companies can't know if they're getting the best solution.

In a Kaggle contest, competitors find out as soon as they submit their solutions how they stack up against fellow contestants. They can keep trying for the duration of the typically three-month contests, which are highlighted on the company web site.

As the first entries come in, the accuracy of competing models improves by leaps, Goldbloom said. As the contests progress, the improvement curve flattens out. Goldbloom and Howard believe that shows the competitive approach pushes data scientists toward the best solutions within human reach.

"Crowdsourcing allows you to squeeze data dry," Goldbloom said.

Not all competitions are open to all comers, however. About 33,000 contestants have taken part in Kaggle's public competitions, where prize money tends to top out at around $10,000. Winners can get invited to participate in elite private contests, which may include access to sensitive private data sets.

Kaggle's business model depends on deep-pocketed contest sponsors like banks seeking to outdo each other with more lucrative prize purses to attract the best competitors, who themselves in theory could then make their livings off Kaggle competitions alone.

The biggest prize by far open to the public is $3 million offered by the California-based Heritage Provider Network medical group to the data scientist best able to use hospital admission records to predict the profiles of people most likely to end up in the hospital. The next-biggest purse is $100,000 in prizes put up by the Hewlett Foundation for algorithms that can automatically grade student essays.

In its grandest vision of itself, the 11-person company backed by PayPal co-founder Max Levchin will have tens of thousands of competitions running simultaneously. Guilds of data gurus will band together to unleash software that enters competitions automatically. Kaggle becomes not just a way to push humans to perform at their best but to make machines themselves smarter as code-based contestants battle and "learn" from their mistakes.

In this way, Howard said, data competitions become steps along the development of artificial intelligence systems such as self-driving cars.

As for why orange used cars are most likely to be in good shape, the numbers did not hold the answer. One notion was that such a flashy color would only attract car fanatics who would be more likely to take care of their vehicles. That didn't pan out, however, since the least well-kept used cars turned out to be purple.

Citation: San Francisco startup makes data science a sport (2012, April 15) retrieved 25 April 2024 from https://phys.org/news/2012-04-san-francisco-startup-science-sport.html