THE WALL STREET JOURNAL.
WSJ.com

TECHNOLOGY

MARCH 15, 2011, 5:59 P.M. ET

May the Best Algorithm Win… With $3 Million Prize, Physicians Group Raises Stakes on the Data-Crunching Circuit

By JENNIFER VALENTINO-DEVRIES

Amid a larger effort to use medical data to improve health care, one company is trying something new: offering $3 million in prize money for the algorithm that can best predict when people are likely to be sent to the hospital.

The algorithm contest, the largest of its kind so far, is part of a trend toward using such prizes to help find the best answers to complicated data-analysis questions.

Data-mining competitions have been around for a while—most notably the $1 million Netflix Inc. prize awarded in 2009 for a model to better predict what movies people would like. But the $3 million health prize, sponsored by California physicians group Heritage Provider Network Inc., raises the stakes. And the start-up handling the competition, Kaggle Pty. Ltd., is aiming to build a business by conducting even more.

The Heritage competition, which begins April 4 and is set to last about two years, will provide contestants with "anonymized" insurance-claims data so they can develop a model to predict how many days a patient is likely to spend in the hospital over the next year.

The goal is to eventually use this model to "identify people who can benefit from additional services," like visits from a nurse or preventive care, thus preventing hospitalization and saving costs, said Jonathan Gluck, a senior executive at Heritage, which has about 700,000 patients.

"We just wanted to spur innovation and different ways of thinking in health care," he said. Heritage, which works with insurers but doesn't provide insurance, says the data won't be used to charge higher premiums.

Holding a contest, even with a whopping $3 million prize, could end up being a good business decision, Mr. Gluck said. "Let's say you could hire 20 or 30 Ph.D.s for $3 million. Well, for $3 million as a prize, you're going to get a lot more than 20 people participating."

Not only that, but by opening the contest to all comers, Heritage can draw on people it would likely never have hired based on their résumé. "You get a lot more creative thinking," he said.

Kaggle, which designs such competitions and prepares the raw data, is seeing a surge in interest from other companies. Since the Australian firm started up 11 months ago, it has conducted about 15 contests, including an effort by Ford Motor Co. to use vehicle data to determine when a driver was distracted, as well as one to help identify writers of Arabic documents.

"We've discovered it's a powerful way to do predictive analytics," said Anthony Goldbloom, Kaggle's founder and chief executive. In coming months, Kaggle, which is in the process of moving to San Francisco, is set to run a contest for the National Aeronautics and Space Administration and Wikipedia, among others.

Most of the prizes are small—$100 to $10,000—but they have all yielded algorithms that beat the previous benchmark, Mr. Goldbloom said. The number of participants has ranged from dozens to a few hundred. For businesses, Kaggle typically charges about $10,000 for the use of its contest platform, not including consulting fees.

Kaggle's success speaks to the growing need for people who can take massive amounts of data and crunch the numbers to make something meaningful. The percentage of job starters on LinkedIn with titles related to analytics and data science has increased more than 40% over the past year, according to the business social-networking site.

"All these companies are having trouble finding data scientists, but I had no trouble when it came to choosing one," simply by looking at Kaggle's leader board, said Mr. Goldbloom, describing how he found a recent hire.

Preparing for the Heritage prize has required months of work by two Kaggle employees, in part because the real-world data is complicated and messy, Mr. Goldbloom said.

"You might see somebody who was diagnosed with a breast neoplasm—a lump on the breast—and the treatment code was a circumcision. There's clearly something not right there," he said.

In addition to cleaning up the data, Kaggle is working with a specialist to render the information anonymous, so it can't be traced back to any patients. This includes not only removing names and addresses, but taking out information about dates and even treatment codes. The focus on privacy comes after Netflix had to scuttle a second prize when researchers were able to extract personal information from some of the Netflix data.

For health-care data, the potential for problems is even greater. "If no one wins, that's not the end of the world. But if the data is de-anonymized, that is," Mr. Gluck said of the Heritage contest.