Data Science Bowl:

A Proven Strategy to Advance Research Through Machine Learning

Received from:
Josette Oder Moynihan
(703) 869-4403 - ph
(703) 880-0456 – fx
oder_josette@ne.bah.com
Saidi, Nilofer [USA]" <Saidi_Nilofer@bah.com>,
"josette@studiod11.com" <josette@studiod11.com>

Executive Summary
Year over year, the Data Science Bowl (DSB) is proof of the power of data science competitions to advance research initiatives. The global competitions converge crowdsourcing, machine intelligence and social engagement to create an excellent environment for young researchers (and experienced ones) to stretch their skills on tough issues, from ocean health to heart health and lung cancer. Our talk will cover approach, technical details of competition design, and outcomes.

The Data Science Bowl Story
The DSB was first conceived by Booz Allen Hamilton, a leading provider of management and technology consulting services. A recognized pioneer in the data science field, Booz Allen's client and pro bono work proved the impact of analytics and machine intelligence to transform how we think about and solve problems. We decided to take action on a much grander scale: harness the power of the data science community to solve a seemingly impossible societal challenge, something bigger than what could be achieved solo.

The idea of tapping into the growing data science community also appealed to Kaggle, the world's largest online global data science community with over 700,000 members. Together, Booz Allen and Kaggle brought the first DSB to life in 2014; the fourth competition is ahead.

A Learning-Centric Approach Yields Impressive Results
The DSB unites data scientists, organizations, industry experts, and citizens to design revolutionary solutions to global problems. This online competition brings together all backgrounds and skill levels in a simultaneously supportive and competitive environment.

From the onset, we designed these challenges to fuel an unparalleled environment of learning and experimentation. The DSB continues to welcome both new and young researchers and analysts as well as seasoned experts across fields. In this environment, one learns from the other all while applying bold thinking and techniques to push the state of the art. These machine learning challenges are also ideal for inspiring an organization's existing staff and recruiting new talent.
For the DSB, the common denominator year over year is a robust data set provided by a research-based organization. When participants are unleashed, they become part of a dynamic community in action. In addition to the Reddit AMAs we put on each year with subject matter experts explaining the problem, there were more than 1,000 kernels (aka online tutorials developed by Kaggle community members) to share analysis and explore models. The sheer number of tutorials created is 5x the number of kernels in similar Kaggle competitions – proof that many people see the DSB as a way to learn from others and grow and hone their skills.

A Working Strategy to Drive Impact

The DSB runs for three months, is open to all data scientists worldwide, is judged with a strictly objective performance metric, and offers cash prizes to the teams producing the top-performing machine learning algorithms. Within this framework, the challenge problem is different each year.

Here are highlights of DSB focus areas and impact to date:

- The first DSB (completed in spring 2015, with a total cash purse of $175,000 provided by Booz Allen) was focused on ocean health through classification of 100+ varieties of plankton species as seen in a large trove of subsurface ocean images. **The winning algorithm reduced time of data analysis by 50%, increased accuracy by 10%, and sparked a cascade of academic activity in the US and Europe to utilize the solution in plankton imaging systems. Tutorials and sample code from the competition also advanced the state of the art in computer vision and Deep Learning.**

- The second DSB (completed in spring 2016, with a total cash purse of $200,000 provided by Booz Allen) was devoted to improving heart health diagnosis through cardio imaging analytics of MRI heart scans**. The competition delivered the first-ever open source algorithm to accurately measure cardiac function and provide cardiologists with an objective, reference diagnostic model to help eliminate measurement bias and streamline diagnosis. The NIH is sharing the approach in the medical community.**

- The third DSB (completed in spring 2017, with a total cash purse of $1 million provided by the Laura and John Arnold Foundation, was aimed at improving early detection of lung cancer in radiological CT scans. **The scientific community is developing ensemble models of winning solutions and integrating algorithms into clinical prototypes. Preliminary findings show that the top teams' solutions outperformed the current, best-in-class clinical classification models by an average of 10%.**

A common feature of the first three DSB challenge problems has been the type of data being investigated: large images stored in a non-traditional domain-specific data format. Some of the machine learning techniques that have been applied to these imaging analytics challenges include machine vision, deep learning, and convolutional neural networks. Another common characteristic is the necessity for feature engineering – designing, extracting, and testing different features in the images for input into the image classification models. The diversity of techniques is a unique characteristic of online competitions like the DSB.

Many researchers of citizen science projects (in which a nontrivial science challenge problem is posed to the public) have argued that it is precisely the diversity of perspectives, approaches, backgrounds, insights, skillsets, and internal biases that lead to the best (most accurate, most insightful, unbiased) crowdsourced solutions.

Data Science Bowl Impact Statement

The DSB "Data Science for Social Good" competition is the largest and most prestigious of its kind, garnering increasing involvement of tens of thousands of participants and advocates as well as corporate, academic and non-profit partners and sponsors. In 2017, we achieved a 53% increase in the number of participants while the prize purse jumped four-fold and online engagement skyrocketed. In 2017, approximately 10,000 DSB

participants joined together to form over 2,000 teams, who contributed over 18,000 algorithm solutions, which employed an estimated 160,000 hours of volunteer effort. Factor in how tutorials proliferated throughout the competition and it's clear the DSB delivers as a promising learning opportunity.

The DSB taps into a growing, global data science community that's passionate about making an impact on our world. From young researchers to experienced ones, they bring bold thinking and a most diverse range of machine learning tools, techniques, and talents to each challenge.