Opinion

Exploring the toolkits of predictive analytics practitioners-part2

Keywords: big data, biobank, cloud computing, cancer, electronic medical records, genomics, proteogenomics

Volume 4 Issue 2 - 2016

Richard Boire
Environics Analytics, Canada

Correspondence: Richard Boire, Environics Analytics, 33 Bloor Street East Suite 400, Toronto, ON M4W 3H1, Canada,
Email Richard.boire@environicsanalytics.com

Received: September 20, 2016 | **Published:** October 19, 2016

Opinion

Continuing on our discussion from last month on toolkits for practitioners, you will note that I purposely do not make reference to specific brand names and companies. By googling data science software, the user can easily obtain a long list of brand names and companies as part of their toolkit search. Instead, it is my intention to discuss functionality and usage when exploring toolkits. In exploring this functionality and usage, the user can then use these insights when conducting their search.

In the last article, we began the discussion of exploring the toolkits of predictive analytics practitioners. One common approach used by some of the leaders in software analytics products is to look at these toolkits in terms of three overall functional areas:

i. Data Management/Integration

ii. Advanced Analytics

iii. Reporting/Visualization

In last month's article, we focused on the first functional area which is data management/integration. For many of the more traditional software players as well as the newer players, this is an area of new development. The tools have evolved to empower more people to conduct these data management activities themselves. Instead of requiring someone to have a computer programming background and the ability to write code, the tools themselves display specific functionalities which are captured within icons that can be dragged and dropped onto a screen in order to perform a specific function. These functionalities represent the many activities that are required to create the end deliverable which is the analytical file. This increased level of empowerment then frees up the more technical person with computer programming skills to conduct the more complex data management activities which the existing software is unable to provide. Under both scenarios in integrating data with and without computer programming skills, detailed data science knowledge is still a requirement if the end deliverable is an analytical file.

Advanced analytics

This article will now focus on the other remaining functional areas which relate to advanced analytics and reporting/visualization. Let's focus first on advanced analytics. For the practitioner, this means that toolkits must have access to multiple mathematical techniques. Some of the more traditional software firms in data science and predictive analytics have provided this functionality for years. In providing this functionality, these firms would vet their mathematical techniques amongst a variety of highly-trained mathematicians. Once the vetting process was complete, the techniques would become a new mathematical procedure which would then be commercially available within its suite of products. As data science evolved, the software firms evolved their products particularly their mathematical techniques into more user-friendly type routines. Instead of having to write code to invoke the appropriate mathematical routine, users could now simply drag and drop the appropriate icon containing the desired mathematical technique.Ensemble type modelling is one of the more recent mathematical developments where a variety of models can be combined in an attempt to capitalize on the benefits of a variety of techniques with the goal being a more optimal solution. The use and development of ensemble modelling can be more easily conducted within this environment as the user just simply connects the various models into one overall node which is entitled ensemble modelling.

For each one of these techniques, users would have access to the statistical diagnostic reports which would assess the mathematical performance of the routine from a statistical perspective. Alongside these reports, users would also have access to reports that would the depict model's performance from a business standpoint. Decile reports/gains charts as well as AUC curves demonstrate the "lift" in performance in using a model as opposed to the status-quo scenario. At the same time, these lift curves can be translated to hard dollar numbers in the form of incremental ROI.

Open-source products

Yet, the growth in predictive analytics and data science has resulted in the use of open-source type products. Before the advent of these open-source products, one had to have the financial means to purchase one of the more commercially available packages discussed in the previous paragraph. This resulted in either large organizations that had the investment capacity to purchase these packages alongside the hiring and retention of staff that had an appetite for programming and code. Like most things in life, necessity is the motherhood of invention. The growing demand for predictive analytics/data science practitioners has resulted in the increase of open-source products that are readily available and virtually free to anyone with a computer and appetite for programming. The field of predictive analytics and data science is now exploding as many people now have the means to try out these tools with the only limitation being the investment in time to learn the coding rules and structure of the particular package. We are still in the early stages of open-source products as an option in building data science solutions. But one needs to be very clear in that there is no real formal vetting process for many of the procedures or libraries that contain these mathematical techniques. In fact, with an open-source product, I could write code to create a mathematical technique and then share it with all the other users of this product. Having said that, some of the older open-source products are improving their methods of vetting out mathematical routines that are created by their user community. Furthermore, the universities and colleges can also provide a more formal network for vetting these techniques.

Reporting/visualization

The last area of reporting/visualization has resulted in a large number of organizations providing this specific type of functionality. In fact, many of the players that provide functionality in both data integration/management and advanced analytics also provide functionality in the third area of reporting and visualization. Yet, a number of players only provide functionality in the third area which may appear to be a limitation. However, the fact that their focus is only on this specific area of reporting and visualization complemented with their attractive cost options does present viable options for organizations in their search for tools. The real competitive advantage in using these tools is how intuitive is it for the user. By this, I mean "Can the user input an analytical file and quickly create visual charts and graphs without too much of a learning curve. Another key feature is flexibility which is essentially the options to drill down and create more detailed visualizations that might deliver some insight. For example, you might create annual trend sales for the years 2011-2016. You find that sales have experienced a noticeable decline in 2016. You then do this by product and notice that a particular product has experienced heavy decline in the last year. You then explore this further by looking at geography and then discover that the declining sales were more prominent in one city. If the data is organized in a very granular manner, you might actually look at product sales of all the stores within that city to see if you can identify a trend within a store or group of stores. The key, though, in designing effective reporting/visualization software is navigation. How can the user easily navigate through the system in order to arrive at the right answer?

The tools continue to improve which allows organizations to complete more projects and to provide solutions that are more easily communicated to the business stakeholders. But again, these are just tools and will never be as important or as significant as the "human" practitioner who is the data science architect behind the development of the solution. As a fellow colleague of mine once said "A fool with a tool is still a fool".

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

Citation: Boire R. Exploring the toolkits of predictive analytics practitioners-part2. *MOJ Proteomics Bioinform.* 2016;4(2):201‒202.
DOI: 10.15406/mojpb.2016.04.00116