Opinion

Open Access
CrossMark

Exploring the toolkits of predictive analytics practitioners-part I

Volume 4 Issue 1 - 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.ca

Received: August 31, 2016 | **Published:** September 02, 2016

Opinion

Tools, tools, and more tools continue to explode in the analytics marketplace as being key enablers in facilitating the many tasks and functions of the data scientist. These tools are not only used in conducting advanced mathematical routines and visualization reports but also in creating the analytical file. The use and consideration of certain tools will depend on what is being done within the predictive analytics process. For example, are we building an analytical file, conducting advanced mathematical routines or communicating our results to key business stakeholders? In this article, we will examine the first scenario of building the analytical file.

Looking at the data audit process

Within this stage, practitioners on my team often refer to the need to get "intimate" with the data. In previous articles, we discussed the need of data audits and data discoveries as being key enablers in providing this data "intimacy". Creating the record of interest and the derivation of variables at this record level represents the ultimate goal in generating the analytical file. Most practitioners acknowledge that this work, although bereft of any real analytics, encompasses close to 90% of the work. Historically, many tools were developed with the intention of providing analytics (advanced and non-advanced) once the analytical file was created. Although this allowed the practitioner to experiment with different mathematical and visualization tools in a very effective manner, minimal tools were provided which facilitated the process of creating the analytical file. Of course, this has changed as vendors began to appreciate and understand the significance of this phase within the data science process. Many vendors in an attempt to minimize the level of required programming knowledge developed GUI type interfaces to reflect the functions and tasks in building the analytical file. However, it must be emphasized that these tools are still not meant for the general business analyst or domain expert. The practitioner of these tools must have a deep understanding of the data and the processes that are required to create the analytical file. This implies an understanding of being able to create output that is necessary in any data audit process and what the output means. Issues related to cardinality, missing values, distribution of values and outcomes, as well as basic statistical diagnostics need to be addressed in each file that is being considered as an initial source of information. Many of the tools in the marketplace allow practitioners to create the necessary data audit output albeit in an adhoc manner.

Using the data audit learning to create the analytical file

From the data audit output, the learning and insights provide the necessary knowledge in creating the analytical file such as:

a. Which source files to join and how to join them

b. What is the target variable and how to create it?

c. What derived variables to create and how to create them.

These three above perspectives are very labor-intensive for the practitioner and are indeed compounded if programming skills are a requirement. But GUI type tools can now be deployed to achieve the same functionality without the person writing any programming code. Checking syntax and programming logic is now eliminated as part of the process. Yet, the practitioner still needs to check or view output to ensure that the required functionality is still achieving the desired outcome.

Besides the time saving component of not having to write and check code, another time saving device is the ability to better organize the flow of work. Within the programming world, a practitioner's work will be organized with comments pertaining to blocks of code that ultimately create the roadmap for the desired outcome. This does reduce the time and effort both for the existing programmer in reviewing his or her work but also when transferring code to another programmer. Yet, the use of a GUI type interface also provides the practitioner with a better visualization of how the flow of work achieves a desired outcome which is the analytical file. Blocks of code representing certain tasks are depicted as icons which are now connected to each other via arrows thereby providing an overall picture of how the analytical file is created. The ability to reduce programming effort and to more visually organize the flow of work have empowered organizations to utilize more people in their data science efforts. The need is less for hard-core programming skills but rather a deeper understanding and knowledge of data and how to work with it within a variety of different data science projects. But the need still remains for the hard-core data science professional with programming skills.

Why programming skills are still required

With the newer types of data such as machine to machine, web data and social media data becoming more accessible, semi-structured and unstructured data is now the norm in many data science projects. Programming skills in terms of parsing the data are now mission-critical in being able to extract the right data from these type of environments. Yet, even in deriving new variables once the necessary data has been extracted, there are an almost infinite number of scenarios where the data scientist has to work the data in order to create some "specific" information that might be considered in a data science solution. Most of the GUI type tools offer the "programming option" as a default to the more advanced technical user/data scientist as there will always be scenarios where the need to manipulate the data is beyond the scope of the current GUI tool. Organizations that have

Submit Manuscript | http://medcraveonline.com

© 2016 Boire. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

Copyright: ©2016 Boire

the hard-core data science programming capabilities complemented by data scientists who are well-versed within the GUI functionality of the tool are simply better equipped to handle both the volume and complexity of data science projects. In my next article, I will look at how these tools have evolved in order to provide more mathematical and visualization capabilities that yield not only better solutions but solutions which are more easily socialized amongst the organization's key stakeholders.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

Citation: Boire R. Exploring the toolkits of predictive analytics practitioners-part1. *MOJ Proteomics Bioinform.* 2016;4(1):168–169. DOI: 10.15406/mojpb.2016.04.00109