Open Source for Predictive Analytics and Machine Learning: Understanding the Pros and Cons

By Fern Halper, TDWI VP Research

WHAT IS OPEN SOURCE FOR PA/ML?

Background

The open source model is a collaborative development one where code is freely available and the copyright holder has the rights to study, change, or distribute the code. Open source tools for analytics have been available for decades, but there has been a recent surge in use as younger analysts and data scientists fashion their own identity and more organizations make the move to analyze big data.

In fact, open source has become quite popular because it is a low-cost source community for innovation which appeals to many data scientists and analytics application developers—especially those who like to code. These toolkits are often used to build predictive analytics/machine learning (PA/ML) models.

Open source comes in a few flavors:

- A free version of each tool is available for download (e.g., the Apache or GNU tools). This version provides no support except for community forums. (Some open source analytics packages come with support licensing.)
- Commercial open source analytics options offer more bells and whistles such as GUIs, data preparation, and visualization capabilities. These are marketed as open source products, although they have licensing fees.
- Many commercial analytics vendors that have traditionally provided proprietary software offer open source options, often allowing users to connect seamlessly with open source analytics tools. Typically, if the vendor has a drag-and-drop visual interface, it will often let a user connect to a model developed in an open source package and drag that into the workflow.

THE CURRENT STATE OF OPEN SOURCE

There are numerous open source analytics tools available for download. TDWI research indicates that many organizations are active supporters of utilizing open source tools for PA/ML. Three of the most popular open source analytics packages in use today include R, Python, and Spark (see Figure 1).

- **R** is a language and environment for statistical analysis and is part of the GNU free software/open source project. R has been in use for several decades and is widely used as a statistical environment by universities as well as organizations. It includes data handling and storage facilities, a large set of tools for data analysis, tools for graphical analysis, and its programming environment. Algorithms number in the thousands.
- **PYTHON** is an interpreted, interactive, object-oriented scripting language now available through the Python Foundation. Like R, it was developed in the 1990s to be an easy-to-read language. Like R, it also has a library for analytics. For example, Scikit-learn includes machine learning and data mining tasks, including clustering, regression, and classification. Theano includes neural network algorithms.
- **SPARK** is an open source big data processing framework that is part of the Apache project.

The framework provides processing capabilities for multiple kinds of big data (text, graph, and streaming). Spark also offers analytics libraries, including a machine learning library (MLlib).

- **OTHER TOOLS** are available for PA/ML. For instance, Scala, an open source programming language, works with ML libraries including SMILE (Statistical Machine Intelligence and Learning Engine) and Deeplearning.scala. TensorFlow, which originally came out of the Google Brain team, has gained popularity for use in deep learning. In our surveys, about 20% of respondents are using it.

Recently, commercial analytics vendors have begun to support R and other open source tools in their analytics products (e.g., in a data science workbench or as part of an analytics platform.) The data scientist can build the model in the library of their choice and insert it into a commercial product workflow. Alternately, some vendor products provide the ability to program the commercial product from inside the open source environment (for instance, in a Jupyter notebook). Others provide a GUI that enables business analysts or data scientists to build a predictive model using an open source language without having to write a script to develop the model.

INCREASED ADOPTION ON THE HORIZON

For those looking to perform PA/ML, R and Python are still at the top of the list of open source technologies for those who plan to build models, in line with the percentages of those already building PA/ML models (see Figure 2). Typically, we see that not all respondents stick to their plans, but interest in the technologies is definitely high.

Additionally, adoption within an organization may increase as more developers look to build intelligent applications; this will increase penetration of open source tools into the organization. In fact, the majority of organizations TDWI surveys believe that business analysts as well as data scientists are and will be using open source to build PA/ML models.

COMMERCIAL VERSUS OPEN SOURCE: IMPORTANT POINTS TO CONSIDER

As organizations move to build and deploy predictive analytics models, they need to determine which kinds of tools to use—commercial versus open source (or both). Some important advantages and disadvantages of open source for analytics are listed on the following pages.

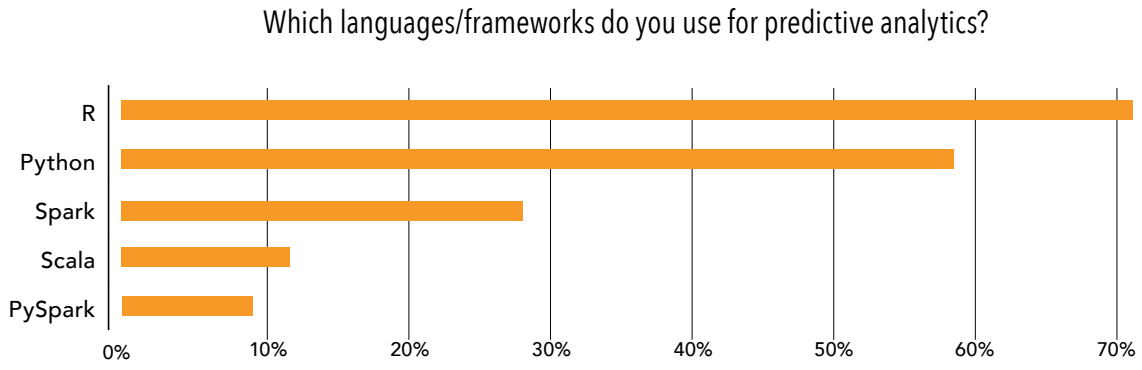

Figure 1. *Open source languages/frameworks for use in predictive analytics. Source: 2018 TDWI Survey, n=244 respondents who are actively building predictive models.*

Advantages of Open Source

- **IT IS FREE.** Open source is a low-cost option if you do not have the budget to pay for a commercial product. There are numerous algorithms for analytics available including statistical, machine learning, NLP, and many other kinds of techniques. This can be a cost-effective way to get started on projects.
- **IT IS A COMMUNITY OF INNOVATION.** There are huge numbers of people contributing to open source analytics projects and that can mean many good and creative ideas that spur its growth. Some open source projects (such as R) have an active community. Vendors such as Google and Amazon are providing new algorithms for use in advanced analytics. Companies that use these communities also often contribute to it because they want to give back.
- **MANY DATA SCIENTISTS ARE TRAINED IN OPEN SOURCE.** Open source packages such as R and Python are more frequently being taught in higher-education settings. Universities are producing data scientists (and others) who know how to use the tools to build PA/ML models. Finding this talent may be easier than finding people trained exclusively on commercial products, although most organizations provide a set of tools so data scientists can use the tools they want to use.
- **OPEN SOURCE CAN HELP ATTRACT TALENT.** In line with the previous benefit, many organizations see the use of open source as a way to attract talent—it allows them to draw data scientists and others who want to be at the cutting edge of innovation in machine learning. It also enables them to interact with others outside of the organization, which some see as a learning opportunity.
- **IT CAN IMPROVE AGILITY.** Open source code can be shared quickly so users get the latest updates quickly and do not need to wait for commercial vendors to release new versions of analytics software. This can improve agility. Additionally, many open source proponents like how open source reduces issues associated with vendor lock-in with proprietary software.
- **OPEN SOURCE ACCELERATES DEVELOPMENT.** Open source can help to speed development. For example, it can be used to experiment to see if a particular modeling approach will work to solve a specific problem. Additionally, with the number of open source tools available for PA/ML, it can be a simple matter of searching open source development platform "marketplaces" (such as Github) to find the apps you need to embed in your own application (e.g., anomaly detection or image recognition).

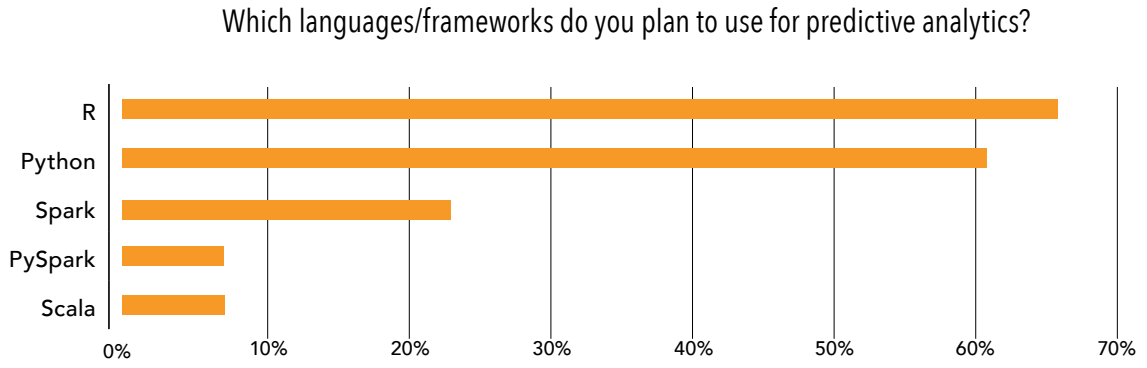

Figure 2. *Planned usage of open source technologies. Source 2018 TDWI Survey, n=180 respondents actively investigating predictive analytics technologies.*

Disadvantages of Open Source

Although open source has many advantages, it has disadvantages as well.

- **IT IS NOT EASY TO USE.** Open source can be hard to learn. In fact, many users say the learning curve is steep. Some open source packages are coding environments; others use scripting languages. Although many vendor products provide intelligence and handholding baked into the software, such is not the case with open source. Users will need to learn how to use open source products to build models. Additionally, algorithms may not all be in one place. Users might have to draw from several packages when analyzing multistructured data such as text or images. Finally, traditional vendor/customer relationships are built on customer satisfaction and customer feedback is incorporated to improve the product and maintain relationships. In the open source world, if you want something done, you may need to do it yourself if the community can't help.
- **MODELS ARE DIFFICULT TO MANAGE.** As organizations build many PA/ML models, they will need to manage, govern, and monitor these models; this requires version control and metadata management to ensure that the right model is put into production. It will include tools for monitoring models once put in production for signs of degradation. The tools for managing models are not yet available in the open source world, yet they are critical to success in the long run. Additionally, open source requires your coders to be good citizens and maintain current versions of languages and fully document their code. Data scientist turnover rates are high and version issues and cryptic code mean models are often rebuilt from scratch with new data scientists or developers, which can be a big productivity hit.
- **DEPLOYMENT CAN BE HARD.** In addition to dealing with model management, deployment can be difficult with open source. It is one thing to build a model using free open source tooling. It is another to deploy it into production. Free open source tools are typically not good at dealing with data handling and the other unglamorous parts of advanced analytics, such as deployment. Deploying models into production requires coding skills (e.g., building model wrappers in a REST API) and this again plays to the talent issues. Likewise, deploying models at scale may be difficult, depending on the tools used. The debate rages about the ability to deploy R at scale across a distributed environment. Vendors typically have refactored their algorithms to work at scale. This may or may not be the case with some open source packages.
- **OTHER DISADVANTAGES.** Other disadvantages of open source include problems if some projects are not cloud ready, forcing enterprises to deal with special considerations for cloud (such as elasticity and failover) as well as code that can run in parallel. Additionally, although some specific tooling is available in open source marketplaces for specific problems, typically these are not domain specific. This is different than a vertical solution a vendor might build to help speed the path to value.

FINAL THOUGHTS

Many commercial vendors are adapting to the idea that commercial products and open source can and should play together in the same space. Vendors should be providing integration points to a variety of data sources and deployment options to support hybrid, best-of-breed analytics environments. Organizations will look to strike a balance between open source and commercial products. It will be important to choose a vendor that can provide flexibility along a continuum of open source and proprietary functionality.

RECOMMENDATIONS FOR USING OPEN SOURCE FOR PA/ML

Make sure you have the right skills

- For PA/ML, skills include understanding how to engineer features, what algorithms make sense for what use cases, and how to interpret and explain model output
- Ideally, you need the same skills when using commercial products, but open source typically does not provide the bells and whistles of a commercial product
- Be prepared—especially if there's a chance you will be coding your models rather than using a GUI or some other interface

Think about deployment

- Plan for packaging the models for use in another system or application
- Consider how you will track and monitor models in production
- Check to see if your organization will support using open source in conjunction with a commercial tool
- Ask whether your commercial vendor allows you to integrate open source models into commercial products or has the ability to bring commercial tooling into open source environments (for monitoring, etc.)

Use open source and commercial products together

- TDWI research indicates those organizations that utilize open source typically also use a commercial package; ask about this so you can have the tools you want and that best serve your needs
- See if your organization will run models built in open source and those built using commercial tooling side by side so you can compare them
- Find out if your models built in open source can be translated to run on vendor platforms; this can help to support parallelizing or enabling run in stream, which provides an advantage for open source deployment
- Ask your commercial vendor to provide model governance so you can control open source model languages and versions as well as to unify your analytics

ABOUT OUR SPONSOR

sas.com/open

SAS is a leader in analytics. Through innovative software and services, SAS empowers and inspires customers around the world to transform data into intelligence. SAS software is engineered to extract maximum value from analytics while supporting the entire analytics life cycle—from data to discovery to deployment.

By combining the power of the SAS Platform with open source technologies, you can unify disparate toolsets and analytics assets into a streamlined, collaborative environment that fosters productivity, business agility, and tangible results.

To learn more, visit sas.com/open.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT THE AUTHOR

FERN HALPER, PH.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).