Standards in Predictive Analytics

The role of R, Hadoop and PMML in the mainstreaming of predictive analytics.

Standards play a central role in creating an ecosystem that supports current and future needs for broad, real-time use of predictive analytics in an era of Big Data.

Just a few years ago it was common to develop a predictive analytic model using a single proprietary tool against a sample of structured data. This would then be applied in batch, storing scores for future use in a database or data warehouse. Recently this model has been disrupted.

There is a move to real-time scoring, calculating the value of predictive analytic models when they are needed rather than looking for them in a database. At the same time the variety of model execution platforms has expanded with in-database execution, columnar and in-memory databases as well as MapReduce-based execution becoming increasingly common.

Modeling too has changed: the open source analytic modeling language R has become extremely popular, with up to 70% of analytic professionals using it at least occasionally. The range of data types being used in models has expanded along with the approaches used for storage. Modelers increasingly want to analyze all their data, not just a sample, to build a model.

This increasingly complex and multi-vendor environment has increased the value of standards, both published standards and open source standards.

In this paper we will explore the growing role of standards for predictive analytics in expanding the analytic ecosystem, handling Big Data and supporting the move to real-time scoring.
Predictive Analytics Today

Predictive Analytics is becoming increasingly mainstream with most companies, most organizations, making it a part of their overall strategy. Whether it is focused on improving customer engagement, managing risk, reducing fraud or optimizing the supply chain, predictive analytics is turning organizations’ data into useful, actionable insight. The results of one recent survey (Taylor, 2013) are shown in Figure 1. Two thirds of respondents have seen a real, positive impact from predictive analytics while fully 43% of respondents report a significant or transformative impact.

**Figure 1: Overall impact from Predictive Analytics**

These are significant numbers and a comparison to the same survey from 2011 shows that the number of organizations seeing a positive impact has increased significantly.

As the impact of predictive analytics has grown in organizations the focus has shifted. In the past the primary use case might have been to occasionally build a predictive model to influence a decision. Today there is a focus on operationalizing analytics. Organizations are building more models and applying these models in their day to day operations. In fact the same study shows that this increase in impact is highly correlated with this shift to a more operational focus as those reporting a transformative impact were much more likely to also report that they tightly integrated predictive analytics into operations. Those reporting predictive analytics as a primary driver for decision-making also outperformed those regularly or occasionally using predictive analytics. As Figure 2 shows, the more tightly respondents integrate predictive analytics into operations the more likely they are to report transformative impact from those predictive analytics.

It should be noted that this need not imply automation of the decision that uses the predictive analytics. While embedding predictive analytic models into automated decisions is a very effective operational deployment technique it is not the only one. Well defined manual decision-making approaches that integrate predictive analytics
can be supported in operational systems. Many organizations find this more palatable and evolve gradually towards increased automation.

**Figure 2: Different integration approach and their impact**

<table>
<thead>
<tr>
<th>Integration Type</th>
<th>Transformative Impact</th>
<th>Significant Impact</th>
<th>Some Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tightly integrated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary driver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regularly used</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occasionally used</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The increased impact of predictive analytics, and increased awareness of this impact, leads to increased demand for both technology and resources. The range of available tools has increased dramatically as has the number of training courses and academic programs for producing “data scientists”. As in other industries this explosion of interest has put a premium on standard approaches that will allow the ecosystem to expand to meet demand. In particular, the open source analytic language R has come to be an essential part of the predictive analytic landscape.

Meanwhile another trend has increasingly merged with the growth of predictive analytics—Big Data. Driven by increased digitization and the Internet, the amount of data available as well as the range of data types and the speed at which that data arrives have all increased. Commonly described as “the 3 Vs” based on an initial Gartner Group paper, Big Data is about handling the increased Volume, Variety and Velocity of data. While Big Data is associated with many trends, it is increasingly clear that the real value of Big Data comes from being able to apply it to business decision-making using data mining and predictive analytics. Simply trying to adapt reporting and dashboard infrastructure to show more data is of limited utility. Organizations are finding that to really exploit Big Data they must turn it into analytic insight.

One of the primary consequences of this is that the data available for predictive analytics is no longer all structured data and it is no longer all stored in traditional databases and data warehouses. Open source technology has evolved to meet this demand by providing a collection of highly scalable approaches to storing and managing data under the Hadoop label. Consisting primarily of the Hadoop Distributed File System and the MapReduce programming framework, Hadoop has rapidly become the platform of choice for storing the large amount of unstructured or semi-structured data that is available to organizations. The role of this open source stack in the predictive analytics market is evolving rapidly as the need to bring this data into the predictive analytics mainstream has grown.
When those building predictive analytics are asked about using the more unusual data types covered by Big Data, they don’t report much usage—most predictive analytic models are still built using structured data. However, when those who have seen a transformative impact from predictive analytics are compared to those who have yet to see any impact, or when those who already have cloud-based solutions deployed are compared to those who don’t, there is a significant increase in the importance of Big Data types as shown in Figure 3.

**Figure 3: Impact of experience on Big Data Usage**

The clear implication of this is that as more organizations get more experience with predictive analytics, the rate at which these Big Data types will be used to build predictive analytic models is likely to grow rapidly.

At the same time there has been an explosion of new technologies for data storage including columnar and in-memory databases as well as Massively Parallel Processing or MPP appliances. Combined with the growth of Hadoop this has put a premium on approaches that allow predictive analytics to be built and executed in a wide variety of proprietary and open source platforms. This, combined with the shift to real-time scoring and away from batch-oriented scoring using predictive analytics has led to a surge in interest in PMML—the Predictive Model Markup Language. This XML format allows a predictive analytic model to be developed in one tool or language against data stored in a particular format and easily migrated to a new environment for execution.

This paper then will examine three core themes:

- The role of R in broadening the predictive analytic ecosystem
- The role of Hadoop in handling Big Data for predictive analytics
- The role of PMML in moving to real-time predictive analytics
Broadening The Analytic Ecosystem With R

Many factors are putting pressure on the predictive analytic ecosystem to expand and grow:

- The use of predictive analytics is rising in companies of every size and in every industry.
- The problems are shifting from those where high cost/high ROI approaches are acceptable, such as risk and fraud, to customer opportunity-facing problems where a different model is required.
- As smaller companies tackle predictive analytics this pressure to lower the cost of solutions is increasing.
- Organizations using predictive analytics are using more models now than in the past, with hundreds or thousands of models becoming increasingly common.

The net result of these pressures is increased demand for analytic professionals, data scientists and data miners, as well as demand for cheaper software alternatives. R is being driven by, and driving, this broadening ecosystem.

Introducing R

R is fundamentally an interpreted language for statistical computing and for the graphical display of results associated with these statistics. Highly extensible, it is available as free and open source software. The core environment provides standard programming capabilities as well as specialized capabilities for data ingestion, data handling, mathematical analysis and visualization. The core contains support for linear and generalized linear models, nonlinear regression, time series, clustering, smoothing and more.

The language has been in development and use since 1997 with the 1.0 release coming in 2000. The core is now at release 3.0. New capabilities can be added by creating packages typically written in the R language itself. Over 5,000 such packages have been added through the open source community.

The Opportunity For R

The biggest opportunity for R is the number of people using it:

- It is widely used in academic programs, creating a pool of new resources every year.
- Because it is free and open source it is widely used in not for profit and government projects, further increasing the pool of potential users.
- As more business and technical professionals see data mining and analytics in their future, it is freely downloadable making it appealing as a tool to learn with.

All this creates a large community of users both new and increasingly experienced.
Furthermore the rate of increase in the adoption of R as an analytic model development language has really taken off. It has risen steadily in usage in the Rexer Analytic Survey every year since the survey first started asking about it. In 2013 70% of respondents now report using it while 24% say it is their primary tool.

Because R is open and designed to be extensible the number of algorithms available for it is huge. Over 5,300 packages have been developed that extend R in some way and it is hard to imagine an algorithm that is not available for R. While proprietary platforms may have algorithms with unique features or better performance characteristics, the likelihood is that R will have something that can solve the same problem. R also emphasizes cross-vendor compatibility through broad support for the PMML standard (see Moving to Real-Time with PMML below).

Related to this huge body of existing material is the fact that R has a large and growing research community. Those developing new algorithms, new ways to process data, in universities and research institutes are far more likely to do so in R than in any other language. This creates a pipeline of new ideas, new approaches, that feeds continually into the R community.

The opportunity for R is clear: Using R gives an organization access to a pipeline of research, a huge body of proven algorithms and an active community while making it possible to recruit at every level of experience from new college graduates to experienced data miners. R has broadened the analytic ecosystem and continues to do so.

**Challenges With R**

It can seem sometimes that the future of analytics lies entirely with R, that there is no opportunity for any non-R based approaches going forward. From the perspective of commercial predictive analytics, however, R has a number of challenges.

First R is an open source project. While this makes it freely accessible it also means that projects can get started, even commit a company to its use, without having the support they need to succeed. Like most successful open source projects this challenge has been mitigated by the rise of companies that provide commercial support and training services around the product.

Parallelism, scalability and performance are also an issue, particularly of the base algorithms. This is mitigated by some of the additional packages developed. It remains a challenge given that one of the benefits of R is the constant supply of new packages which may or may not be scalable. Commercial vendors are mitigating this by providing their own implementations that take R functions and execute them in-database or in-Hadoop as well as on massively parallel infrastructures or appliances.

Tooling is also an issue with the basic R environment being script-based. While many data miners like to be able to write code, script management and reuse are a
problem for R scripts as they are with any programming language. Commercial implementations are integrating R into integrated development environments (IDEs) so that less technical users can develop scripts and to improve management of reuse etc. Improved editors, often integrated with source code control systems and testing environments, are likewise improving development practices.

It can be challenging to deploy R on a production system. The base core package has a large tree of dependencies that must be installed with it. The dependent packages include compiler-related code as well as packages that are not generally required for any other purpose on back-end servers. R deployments can also involve compiled third-party packages that have other dependencies. This can be difficult to manage when the analytic team does not have full control over the analytic deployment and must request all this from IT. The use of PMML to manage deployment, as discussed in Moving to Real-Time with PMML below, can mitigate this for both batch and real-time deployments.

Finally, batch scoring of models is an issue as much of R’s strength coming from its use in batch updates of databases with scores. The increasing use of real-time scoring puts pressure on models developed in R as a result. This is addressed partly by growing support for PMML in the R community (see Moving to Real-Time with PMML) and by vendor support for real-time scoring services that can be created from R models.

**Recommendations For R**

Commercial—not just academic—use of R is increasingly practical and well proven. Organizations should make R part of their predictive analytics adoption and roll out strategy. This might be to focus all analytic development on R or take what is an increasingly common approach to mix and match R with proprietary environments—especially where performance, scalability and/or support are of prime concern. Even for organizations already committed to a commercial platform it makes sense to take advantage of R at some level and organizations should explore their platforms support for integrating R.

Whether R is the core or their strategy or just a part, organization’s should plan on working with a commercial vendor that has a solid plan for R in terms of providing scalable implementations of the algorithms they care about. Organizations should also look for either a better development environment than the default or integration with graphical modeling tools, especially if they want to include more non-data scientists in the process. Make sure also to investigate deployment options though this can be mitigated by adopting technology that supports PMML.
Managing Big Data with Hadoop

The era of Big Data has arrived. As anyone who reads the IT or business press knows, there is more data now than ever before. Dealing with that data, getting value from it, is top of mind in commercial, government and not for profit organizations of every size. Big Data is best described using the “3Vs” first identified by Gartner Group:

- **Volume**—there is simply more data available today than in the past as more of our lives are digitized and as organizations’ digital history reaches back further.
- **Variety**—data is no longer just structured data suitable for storing in a relational database. Unstructured and semi-structured data such as emails, sensor logs and text as well as audio and video are becoming increasingly common.
- **Velocity**—this data is becoming available, arriving at organizations, and changing more rapidly than ever. Last week’s data, even yesterday’s data, is increasingly being complemented by today’s data or even live streaming data.

These three challenges are putting pressure on traditional data infrastructures. Simply trying to expand databases or data warehouses (and their associated processes) to cope is difficult and expensive. Hadoop has emerged as a key technology in dealing with Big Data.

Introducing Hadoop

Hadoop consists of two core elements—the Hadoop Distributed file System or HDFS and the MapReduce programming framework. An open source project, Hadoop development started in 2004 inspired by earlier work at Google, and became an official top-level Apache project in 2008.

HDFS is a highly fault tolerant distributed file system that runs on low-cost commodity hardware, allowing very large amounts of data to be stored cheaply. The data uses a file metaphor so that the structure of the data needs to be specified only at read rather than at write. It maintains multiple copies of the data and the software handles management tasks such as keeping enough copies on different nodes or responding to node failure. Name nodes and data nodes run on dedicated machines and communicate over TCP/IP.

MapReduce is a programming framework that breaks large data processing problems into pieces so they can be executed in parallel on lots of machines close to the data that they need to process. First the input data is broken up into pieces. Then the programming logic runs as multiple parallel processes on the nodes that store their share of the data. This Map step creates key value pairs that are then, shuffled to group pairs with same key so they can be processed (Reduced) to give an answer by aggregating all the pairs with same keys in some way. Solving a given problem may
require multiple rounds of MapReduce coding, especially when the function being performed is complex.

**The Opportunity For Hadoop**

Hadoop provides a distributed, robust, fault tolerant data storage and manipulation environment that is well suited to the challenges of Big Data:

- The use of commodity hardware allows it to scale at low cost and so handle the *Volume* involved.
- Because a file metaphor is used that allows the schema of data to be applied only when it is being read (rather than being pre-determined when it is written to storage) Hadoop is flexible enough to handle the wide *Variety* of data involved.
- The storage and processing are streaming-centric in that they assume data will be continuously added and this enables the environment to handle the *Velocity* involved.

Some newer organizations, web 2.0 companies for instance, are using a pure Hadoop strategy and putting all their data in Hadoop. More common is a mixed database/data warehouse/Hadoop approach. Several distinct use cases are emerging for using Hadoop in such a mixed environment:

- Hadoop is used as a landing zone for data where it can be pre-processed (cleaned, aggregated) before being moved to a data warehouse. This can include analytic processing using algorithms (that run in MapReduce). This approach allows new data sources to be rapidly added to an existing environment.
- Hadoop is used as an active archive where data used less often is moved out of the data warehouse and onto Hadoop. This allows older data that might have been archived inaccessibly in the past to be available for analysis, to build predictive models or forecasts where several years of data might be useful for example.
- Hadoop is used somewhat standalone for rapid data exploration. It is easy to bring data in from lots of sources (including from databases or data warehouses) and various tools exist to explore the combined data to see what might be learned. What is learned might be applied to data in Hadoop going forward or used to determine what new data might be usefully integrated into a more traditional environment.

The explosive growth of the Hadoop community in recent years means that this open source project is supported by several commercial organizations that provide training, supported distributions and consulting. In addition database and data warehouse vendors have increasingly integrated their storage with Hadoop while appliance vendors are delivering Hadoop-based appliances that have integrated security and management capabilities.
Challenges With Hadoop

Hadoop has many features that make it appealing in an era of Big Data but it has its challenges too. Like R, Hadoop is an open source project, so projects can get started and even become central to a company’s data strategy without having the support they really need. This has been mitigated for Hadoop, as it has for R and other successful open source projects, by companies providing commercial support and training services around Hadoop.

In addition, Hadoop has some technical challenges:

- Hadoop is a programmer-centric environment and there is no support for SQL in the base environment making it difficult to use existing analytic and development tools against data stored in Hadoop. Additional projects such as Hive and more recently Cloudera Impala have added support for SQL while open source frameworks allow more traditional development approaches to be used to build data processing applications by abstracting the MapReduce programming framework and allowing developers to write SQL or Java.

- Hadoop structures are better at batch processing than they are at interactive systems. A Hadoop/MapReduce approach will handle great scale in batch processing but is less effective at processing a single record. Most use cases for Hadoop focus on its ability to process large numbers of records in parallel rather than on, for instance, being able to retrieve the information about a single customer to drive a next best action script for an interaction with that customer.

- From a decision management perspective Hadoop also lacks any specific data mining or predictive analytics support. Combined with the lack of support for SQL this means most advanced analytic tools won’t work on data stored in Hadoop. This has been largely addressed, however, by the rapid increase in the number of analytic companies that offer direct access to data in HDFS. In addition several companies offer data mining and analytic algorithms that can execute as MapReduce jobs and there is growing support for scoring HDFS data by applying analytic models defined in PMML.

- Scoring against Hadoop means writing MapReduce jobs that pre-process, score and post-process the data. Writing a model as MapReduce code takes significant time and effort, delaying implementation and increasing cost. Better development approaches and improved vendor support mitigate this. Various vendors also allow analytic models, in PMML or other formats, to be executed in a Hadoop environment for training and for scoring, though this is also batch-centric.
Recommendations For Hadoop

The biggest mistake organizations make with Hadoop is by beginning by focusing on the technology. Hadoop has a lot of potential for companies adopting predictive analytics but it must be applied in context. Instead of beginning with a Hadoop installation and loading data into it, begin with a business problem—a decision that must be made. By determining the analytics that will be required to make that decision more accurately or more precisely organizations can see what kind of data will be required and on what terms. This will lead naturally to identifying needed data that exists already in traditional data infrastructure as well as potentially useful data that does not. This creates a use case for Hadoop as it has identified a business problem that requires data not already available in the existing infrastructure.

For organizations that are familiar with and engaged in open source Hadoop should be no different from other technologies they have adopted. Where organizations lack this familiarity they should consider one of the commercial organizations that support Hadoop, provide a distribution that includes the full range of capabilities and have plenty of experience. This might be a Hadoop pure play or an existing vendor offering Hadoop services.

Finally once Hadoop becomes part of the data infrastructure for an organization it is important that it is supported by the rest of their decision management infrastructure. The analytic tools used must be able to access data in HDFS while any in-database approach should be extensible to in-Hadoop also.

It is possible to create Hadoop-sized problems by focusing on batch scoring. If an organization has many customers to score and thinks in terms of batch scoring—that every customer must be scored every day—then this can sound like a Hadoop problem. However the scoring could be done on-demand in real-time for the much smaller number of customers who interacted with the company on a typical day. This real-time scoring problem might be much smaller in scale and so not demand a Hadoop infrastructure.
Moving to Real-Time with PMML

In the past most analytic models were built and applied with a batch mindset. Once a model had been designed and tested it was applied to all the customers, all the claims, all the products in the database. This scoring step might be repeated every day or every week so that a reasonably up to date score was always available in the database.

In recent years this approach has been challenged. Increasingly the focus for companies is on scoring a transaction or a customer right when a decision is required—if an offer is to be made then the customer churn score, for instance, is calculated as part of determining the offer to be made. The rise of Decision Management as an approach for handling real-time decisions has only increased this focus.

This move to real-time scoring of predictive analytics has put pressure on many traditional analytic environments. When scoring was done in batch it was generally done using the same technology as was used to build the model. Once the final script was ready it could be used to process all the records and they could then be loaded into the marketing system or stored back into the database.

With real-time scoring this becomes impractical. Not only are these analytic environments batch-oriented they are often only loosely attached to the production environment. It has become essential to be able to move models from their development environment to a more real-time, interactive scoring environment. Many platforms have added such a capability. Some organizations want to use a different production environment. They need a way to move models, built in a variety of analytic tools, into their production environments, business rules management systems etc. PMML has emerged as the primary way to do this.

It should be noted that there is nothing inherently real-time about PMML. PMML can and is used in batch scoring scenarios too. The focus of the standard is on interoperability and on replacing custom code when deploying models. This interoperability is a key benefit of PMML, as noted below, independent of an organization’s need to move to real-time scoring.

Introducing PMML

PMML is an XML standard for the interchange of predictive analytic models developed by the Data Mining Group. The basic structure is an XML format document that contains:

- A Header
- A data dictionary containing both continuous and categorical variables
Data transformations
• One or more models each consisting of a mining schemas based on the type of model, a target and other outputs such as measures of accuracy.

PMML started in 1998 with 0.7, moving to a 1.0 release in 1999. Since then the standard has seen multiple releases with 4.1 being the most recent (in 2011). The 4.x releases marked a major milestone with support for pre- and post-processing, time series, explanations and ensembles.

Support for PMML is widespread and growing with an active community. Thousands of analytic professionals participate in the Data Mining Group’s LinkedIn Group and many analytic vendors are either members or provide support. Members of the PMML consortium include IBM, MicroStrategy, SAS, Actian, Experian, Zementis, Equifax, FICO, KNIME, Open Data Group, RapidMiner, Togaware Pty Ltd, Angoss, KXEN, Microsoft, Oracle, Portrait Software, Prudsys, Salford Systems, SAP, StatSoft, and Tibco. In addition organizations including BigML, Predixion, Revolution Analytics, Teradata as well as open source projects such as Weka and R also provide support for the standard.

The Opportunity For PMML
PMML offers an open, standards-based approach to operationalizing predictive analytics. This is a critical need for organizations looking to maximize the value of predictive analytics: unless predictive analytic models can be effectively operationalized, injected into operational systems, then the danger is that they will sit on the shelf and add no value. Similarly if it takes too long to operationalize them—if it takes weeks or months—then the value of the model will degrade even as the cost to implement the model rises. As the results of a model are increasingly needed in real-time, this is a critical need for organizations.

Support for PMML is increasingly broad-based:
• Scoring servers and execution engines running on a wide range of production environments can bring in PMML models and execute them to provide a real-time score.
• Business Rules Management Systems can import PMML and either execute them as a model or translate them into business rules that can be executed as part of a rules-based decision.
• Databases and data warehouses can execute PMML models in-database allowing scores to be applied without moving the data from the operational datastore in which it resides.
• PMML models can be executed on Hadoop nodes e.g. through Hive using a SQL-like approach or by being imported into commercial execution engines that provide in-Hadoop execution.
In addition it is possible to import PMML models into many analytic development environments, allowing new models or ensembles to be based on models built in other tools. Some model management environments also support PMML allowing their model reporting and monitoring capabilities to be applied to a PMML model also.

This wide range of deployment options for PMML models also means that organizations can relax their concerns about multiple development environments. If models can be generated as PMML and that PMML can be executed widely then it is possible to create an environment in which models are developed with any analytic tool and run anywhere. As deployment and execution become more focused on operational environments, and less tied to the model development environment, this is becoming an important capability for analytic organizations.

**Challenges For PMML**

The primary challenge for PMML, as it is for any standard, is to get the vendor community to regard support for it as more than just a “check the box” capability. Once a standard becomes established, as PMML has, organizations will ask vendors about their support for it. However it is easy for this support to be focused on checking the box on the RFP and nothing more. The result is minimalist implementations that do not provide the depth of support required for a real project. This has been a challenge for PMML in the past but the market’s adoption has recently reached a tipping point where organizations are relying on PMML in critical production environments. Large organizations where heterogeneous environments are the norm realize the benefits of an open, vendor-neutral standard. These customers are demanding PMML support from their suppliers which, in turn, is putting a great deal of pressure on vendors to provide solid support.

Standards such as PMML also struggle to get vendors to stay current and support the latest release. For PMML this is particularly an issue for the support in PMML 4.x of pre- and post-processing. According to the Data Mining Group’s list of supporting vendors only about half are supporting PMML 4.x for instance. When an organization has different parts of its ecosystem supporting different versions of a standard this can undermine the value proposition for that standard. For PMML this is somewhat mitigated by the availability of tools for converting between versions. In addition many organizations remain uncertain how best to handle pre- and post-processing. Many do not actually want their analytic platform to handle this automatically for fear that the way the analytic team designed this will not match the way the production environment can best support it.

Finally not everything that an analytic team might wish to do is supported in PMML. Most mainstream model types are covered by PMML but new, research-type algorithms may not be. Projects may therefore have to choose to use an alternative algorithm or lose the ability to generate a PMML model for their project. Some
vendors with specific additional functionality in their models try and extend PMML to avoid this problem for their customers but this of course also means that their products must be used at both ends as no other vendor will understand their extensions.

All of these challenges are typical for a standard and as PMML continues to build support in the user community, and as more organizations commit to it, these problems will typically be mitigated. PMML is in the fortunate position of being the only standard for predictive models that is widely accepted and supported across commercial and open source tools. With no real competition and broad support the long-term benefits of adopting PMML seem likely to greatly outweigh minor challenges in adopting the standard.

**Recommendations For PMML**

All organizations approaching predictive analytics should include PMML in their list of requirements for products. Selecting analytic tools that do a good job of generating and consuming PMML and identifying operational platforms that can consume and execute PMML just makes sense. While organizations committed to a single vendor stack may be able to avoid this requirement, even there the ability to bring models developed by a consortium or third party into that environment may well prove critical while partners may need to execute models but not share the same vendor stack.
Future Developments

R, PMML and Hadoop are all well established standards that can and should be part of a predictive analytics strategy. There are also some future developments that are worth considering—the emergence of the Decision Model and Notation standard, growing acceptance of Hadoop 2 and planned updates to PMML.

DMN

The Object Management Group recently accepted a new standard, the Decision Model and Notation standard. DMN as it is known is now in beta and is expected to be finalized in 2014. DMN provides a common modeling notation, understandable by both business and technical users that allows decision-making approaches to be precisely defined. These decision models can include the specification of detailed decision logic and can model the role of predictive analytics at a requirements level and at an implementation level through the inclusion of PMML models as functions. DMN will allow the modeling of automated and manual decisions in a way that shows how business rules and predictive analytics are being combined. It can also be used to describe the requirements for predictive analytic models.

Hadoop 2.x

Technically this is Apache Hadoop 2.2.0 and it was released in October of 2013. It’s considered a future development in this paper because most Hadoop users are not using it yet. Hadoop 2.x is all about really all about YARN—a resource management system that manages load across Hadoop nodes that allows other approaches besides MapReduce to be used. This allows Hadoop resources to be shared across jobs using different approaches and, in particular, will allow more real-time processing. For instance new engines that support SQL, steam processing and graph processing can and are being developed and integrated into the Hadoop stack using this approach. Other changes include better support for federation.

PMML

PMML Release 4.2 is expected to be released in the first half of 2014. As with 4.1, release 4.2 is expected to improve support for post-processing, model types and model elements. 4.2 is particularly focused on improving support for predictive scorecards (especially those with complex partial scores), adding regular expressions as built in functions, and continuing to expand support for different types such as continuous input fields in Naïve Bayes Models. In addition DMG continues to consider ways to make PMML more adaptable and relevant to its growing community of users.
Conclusions

These standards—R, Hadoop and PMML—are proven tools for broadening the use of predictive analytics, supporting Big Data and operationalizing analytics. All three should be part of an organization’s predictive analytic strategy.

The use of R to develop at least some of an organization’s predictive analytics offers clear benefits in terms of staff acquisition and retention. It also plugs an organization into an innovation pipeline. Working with a commercial vendor that supports R provides the scalable implementations, development and deployment tools, and support that will be required.

Big Data is going to be increasingly important in predictive analytics as a complement to traditional data types. The use of Hadoop to store some of this data makes sense. Making sure this data storage is integrated with existing data infrastructure and working with commercial vendors that can provide support is essential as is ensuring that this data can be accessed for both developing and deploying predictive analytic models.

Finally all organizations approaching predictive analytics should include PMML in their strategy. Analytic tools and operational infrastructure that support PMML should be prioritized. While organizations should not always be constrained to models supported by PMML it is generally both a good idea to do so and not much of a limitation.

The future of these standards, and of new standards like the Decision Model and Notation, is bright. Increasing vendor support and growing awareness means that organizations will have excellent choices that support these standards allowing them to benefit from a broader pool of resources, work better with service providers and more easily integrate the products they want to use.

*The following content is provided by our sponsors, Revolution Analytics and Zementis.*
Revolution Analytics

Champions of an Emerging Programming Standard

Revolution Analytics Delivers R for Big Data Analytics in the Enterprise

The open source, statistical programming language R is in use by over 2 million people worldwide. It is supported by a vibrant and talented community with over 5,000 published packages. R continues to grow in use and popularity across academia and multiple industries.

Its use in corporate environments started as an isolated group which would develop and test hypotheses before having to re-code their work for deployment on a legacy platform for production. However, with adoption of R by some of the largest and most recognizable names in the technology industry, including Google and Facebook, R has emerged as a de facto standard for statistical programming and production deployment in diverse data environments.

Revolution Analytics has helped organizations pave the way to the adoption of R as a programming standard through its initial offerings of support and consulting. Today, with its Revolution R Enterprise 7 (RRE 7) software platform, the company delivers a fully tested and supported version of R that scales beyond its open source origins to deliver fully parallelized deployment of statistical algorithms.

As companies augment and update their data architectures with more and new technologies including enterprise data warehouse appliances, Hadoop, cloud storage and real-time streaming; the potential for latency, data loss and compromised security increases. RRE 7 allows companies to use their data without having to move it or being restricted to using only a small sample of it.

Write Once, Deploy Anywhere

The RRE 7 stack is purpose-built to deliver high performance, Big Data ready analytics to any data platform so you can:

- Eliminate out-of-memory limitations
- Analyze data without moving it
- Build models using large data sets
- Build models using ensemble techniques
- Deploy models without recoding

RRE 7 empowers data scientists to build predictive models to perform on the platform where the data is stored. This ensures that investments in data architecture are protected.

RRE 7: The Big Data, Big Analytics Platform

As Revolution Analytics continues to deliver better performance for a growing number of the most used and effective models, RRE 7 also delivers these powerful packages to a broader, less technical audience. Integration with BI and visualization platforms, including market leaders Tableau and Alteryx, puts the power and flexibility of R directly into the hands of business managers and analysts.

For more information on R and Revolution R Enterprise please visit: http://www.revolutionanalytics.com.
Write Once

Deploy Anywhere

With Revolution R Enterprise 7, create predictive models in the industry-standard R language, and deploy to servers, Hadoop or Teradata Database without recoding. Enjoy the latest innovations of open source R with high performance, big data scalability, and support from Revolution Analytics.

PMML, the Predictive Model Markup Language, allows for predictive models to be easily moved into production and operationally deployed on site, in the cloud, in-database or Hadoop. Zementis offers a range of products that enable the deployment of predictive solutions and data mining models built in IBM SPSS, SAS, StatSoft STATISTICA, KNIME, KXEN, R, etc. Our products include the ADAPA Scoring Engine and the Universal PMML Plug-in (UPPI).

**SOLUTIONS FOR REAL-TIME SCORING AND BIG DATA**

ADAPA, the Babylonian god of wisdom, is the first PMML-based, real-time predictive scoring engine available on the market, and the first scoring engine accessible on the Amazon Cloud as a service. ADAPA on the Cloud combines the benefits of Software as a Service (SaaS) with the scalability of cloud computing. ADAPA is also available as a traditional software license for deployment on site.

As even the god of wisdom knows, not all analytic tasks are born the same. If one is confronted with massive volumes of data that need to be scored on a regular basis, in-database scoring sounds like the logical thing to do. In all likelihood, the data in these cases is already stored in a database and, with in-database scoring, there is no data movement. Data and models reside together; hence, scores and predictions flow at an accelerated pace. The Universal PMML Plug-in (UPPI) is the Zementis solution for Hadoop and in-database scoring. UPPI is available for IBM PureData for Analytics powered by Netezza, SAP Sybase IQ, Pivotal/Greenplum, Teradata and Teradata Aster. It is also available for Hive/Hadoop and Datameer.
BROAD SUPPORT FOR PREDICTIVE ANALYTICS AND PMML

ADAPA and UPPI consume model files that conform to the PMML standard, version 2.0 through 4.1. If your model development environment exports an older version of PMML, our products will automatically convert your file into a 4.1 compliant format. Supporting an extensive collection of statistical and data mining algorithms, ADAPA and UPPI deliver on the promise of a universal deployment capability which is vendor-neutral and platform independent.

“By working with Zementis, a PMML innovator, we are able to offer a vendor-agnostic solution for efficiently moving enterprise-level predictive analytics into the FICO Analytic Cloud. Customers, application developers and FICO partners will be able to extract value and insight from their predictive models and data right away, using open standards. This will result in quicker time to innovation and value on their analytic applications.”

Stuart Wells, Chief Technology Officer FICO

“By partnering with Zementis, we are able to offer high performance, enterprise-level predictive analytics scoring for the major analytics tools that support PMML. With Zementis and PMML, the de-facto standard for representing data mining models, we are eliminating the need for customers to recode predictive analytic models in order to deploy them within our database. In turn, this enables an analyst to reduce the time to insight required in most businesses today,”

Chris Twogood, VP Product & Services Marketing Teradata
About the Research

Research was conducted by James Taylor of Decision Management Solutions in Q4 2013. The opinions expressed are those of the author and not necessarily of the sponsors of the research or of the groups responsible for these standards. We would like to acknowledge our research sponsors without whom this research would not have been possible—Revolution Analytics, Zementis and The Data Mining Group.

References


The Data Mining Group, http://dmg.org


Contact Us

If you have any questions about Decision Management Solutions or would like to discuss engaging us we would love to hear from you. Email works best but feel free to use any of the methods below.

Email : info@decisionmanagementsolutions.com
Phone : +1 650 400-3029
Web : www.decisionmanagementsolutions.com