White Paper

Operationalizing Analytics Solutions and Methods

Andrew Fast, Ph.D.

Chief Scientist

John Elder

Founder

February 2015

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING

Table of Contents

1.0 Operationalizing Analytics

It is a challenge to make an analytic model work in a production environment, as it requires teamwork from IT, Data Management, Analytics, and Business units. The goals of the process are threefold:

1. Build models with repeatable, reliable results that do not depend on any single person or working environment to operate.
2. Make model results available to end-users in a timely and useable manner.
3. Monitor model performance on an ongoing basis to ensure quality and alert analysts to any degradation over time.

There exist different strategies for achieving these goals. In this report, we briefly touch on two required components:

1. Automating the model scoring process through model management and monitoring
2. Providing the results to end users in an interactive fashion.

The ultimate solution depends on the organization’s goals and available resources.

1.1 Model Management and Monitoring

Analysts need to create repeatable model runs in a timely manner. This is the most important requirement when operationalizing a model as strategies for visualization and deployment depend on a strong foundation of model management underneath. By monitoring models, managers know when to retrain a model being used in production.

Model Management involves:

- Versioning – maintaining an approved model across changes over time.
- Scheduling- running the model on an ongoing basis to obtain timely results.

Model Monitoring involves observing and reporting:

- Model Run times – Is the model getting bogged down on large data?
- Model performance – Is it consistent with the scores we observed during training? Changes in performance could indicate a shift in the underlying data population.

1.2 Visualization and Delivery

The goal of visualization and delivery is to take the automated results coming out of a model management procedure and make them available to end-users who may or may not be a part of the analytics team. This step gets model results out of the analytics team into the hands of business users.

2.0 Approaches for Model Management and Monitoring

2.1 Baseline: Managing Models Manually

At minimum, an analyst runs a script manually each time results need to be updated. Though simple, this approach suffers when the demand for the model increases placing undue responsibility on one person. This baseline approach can work however, if the time between model runs is large (quarterly or annually) and the model code (including scripts, streams, etc.) is readily available for multiple people to use and run as needed.

2.2 Using Commercial Tools

Each of the Gartner Leaders in the Magic Quadrant for Analytics (SAS, IBM, KNIME, and Rapid Miner) have a package for deploying and operationalizing models. SAS Model Manager is the most well-developed of the group, and has capabilities for model management (versioning, scheduling, etc.) as well as model monitoring (champion vs. challenger, etc.) The other tools are strong on the model management side, but have fewer capabilities for model monitoring. If an organization has significant investment in a commercial analytics platform, using that tool to push models into operation often provides the best value for the effort.

2.3 Open-Source Options

In recent years, open-source analytics platforms have been increasing in popularity. This demand is driven by their greatly lower prices and somewhat greater dynamic functionality. Many of these packages integrate with an open standard for models called *Predictive Markup Modeling Language*, or *PMML.* Zementis Adapa, for example, is a low-cost software solution for operationalizing models stored in the PMML format. One major caveat is that many commercial vendors do not fully support the PMML standard, and those that do often inject proprietary features in non-standard fields.

2.4 Custom Software Solutions

The most common solution (after manual model management) for operationalizing models is to employ custom software. By definition, these solutions can take on many different forms depending on the situation. We highlight three different strategies we have seen in multiple organizations.

2.5 Using Commercial Tools and Windows Scheduler

This solution could be dubbed “Model Management Lite” as it replaces the commercial model management tools with a collection of system scripts and tools. One common solution is to use Windows Scheduler to run a Windows batch process on a regular basis. This approach has many of the strengths of a commercial system due its reliance on a commercial platform but lacks the integration and ease of use that a commercial tool provides. However, it is low in cost and can be implemented within almost every IT environment.

2.6 Database Stored Procedures

The most efficient way of automatically scoring data is to do in-database scoring as the data is updated. This can be achieved in a database through the use of a stored procedure or materialized view. Every database system, whether a traditional relational database (RDBMS) or a “Big Data” NoSQL solution, has API hooks for integrating custom scoring into a database. This makes sense to use for organizations with high-volume data and a mature IT capability.

2.7 Custom Software Application

Custom scoring software can integrate model operationalization into a middleware software component (or other automated scoring mechanism) as a standalone tool or as part of a larger system. The specifics of this approach will depend on the goals of the system and the available IT environment.

3.0 Approaches for Visualization and Deployment

The many options for visualization and deployment range from the traditional reporting in Excel to fully custom interactive web applications. Here, we briefly overview each of the major strategies.

3.1 Standard Reports

The most basic form of deployment is sending model results to Excel, in CSV or other tabular format. These files are easy to pass around (unless size is an issue) and can be opened by everyone using standard software. These reports are usually automatically generated as part of the model run and sent to, or downloaded by, the intended recipient. The drawback to these methods is that raw numbers in an Excel spreadsheet are not instantly consumable by management.

3.2 Integration with Business Intelligence Software

Business Intelligence tools such as Tableau, MicroStrategies, and COGNOS provide a more interactive method for displaying model results. With these tools one can build an interactive view of the system, but they are sometimes limited in the types of visualizations allowed.

3.3 Custom Visualization Tools

The sky is the limit here in terms of functionality (and perhaps, cost). Anything from a web-based visualization to custom tablet and smartphone apps could be used to provide model results to end-users. All of these tools require a solid foundation for model management and results storage in order to supply these apps with data. Often custom tools can combine results from multiple systems to provide the best user experience. The extra effort to get the functionality can be well worth it, as the models that are actually used are the ones that will be profitable for the company.

4.0 Summary

Successful operationalization efforts combine the automation of the modeling process with a delivery method that can be integrated into the existing business workflow. There are many ways these can be combined successfully. The one that will produce actionable results for your organization is the one to use.

5.0 About the Authors

Andrew Fast, Ph.D. is the Chief Scientist at Elder Research and leads the research and development of new tools and algorithms for data and text mining. Dr. Fast graduated Magna Cum Laude from Bethel University and earned Master's and Ph.D. degrees in Computer Science from the University of Massachusetts Amherst. There, his research focused on causal data mining and mining complex relational data such as social networks. Dr. Fast has published on an array of applications including detecting securities fraud using the social network among brokers, and understanding the structure of criminal and violent groups. Other publications cover modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (work featured on ESPN.com). With colleague Dr. John Elder and others, Andrew has written a book on Practical Text Mining that was awarded the PROSE Award for Computer Science in 2012.

Dr. John Elder, Founder and CEO of Elder Research, leads the most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he’s an adjunct professor. He’s authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served 5 years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose “book of the year” awards.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING