Embedding Predictive Analytics

Version 8.1, March,2017

By James Taylor

CONTENTS

Overview

Architecture

Capabilities

- **Data Management**
- **Data Preparation**
- **Data Visualization and Analysis**
- **Predictive Modeling**
- **Model Validation**
- **Deployment and Scoring**
- **Model Monitoring**
- **Model Tuning**
- **Repository**
- **In-database Analytics**

Next Steps

Embedding Predictive Analytics one of the five key capabilities needed when building Decision Management Systems. Each can be adopted incrementally, and can scale based on resources and business drivers.

Data mining and predictive analytic models are often grouped with business intelligence (BI), reporting, and visualization under the general term "analytics." Data mining and predictive analytics differ from BI capabilities in key ways:

- Predictive analytics extract meaning about the likely future rather than summarizing or understanding the past. They use historical data to make predictions about what is likely in the future.
- Predictive analytics are probabilistic rather than definitive. Generally, they say how likely something is, by making a prediction with a certain degree of confidence, or rank order a set of possible outcomes from most to least likely.
- Rather than relying on the visual processing power of humans to see patterns in data, they rely on mathematical algorithms to explicitly extract these patterns from the data in a format that can be deployed in information systems.

This last point is key for predictive analytic workbench products being used to develop Decision Management Systems. Presenting the results of a predictive analytic project as mathematics or even as visualizations and reports is not sufficient. It must be possible to use the product to both produce an effective predictive analytic model and embed such a model into an operational system to be useful for building Decision Management Systems.

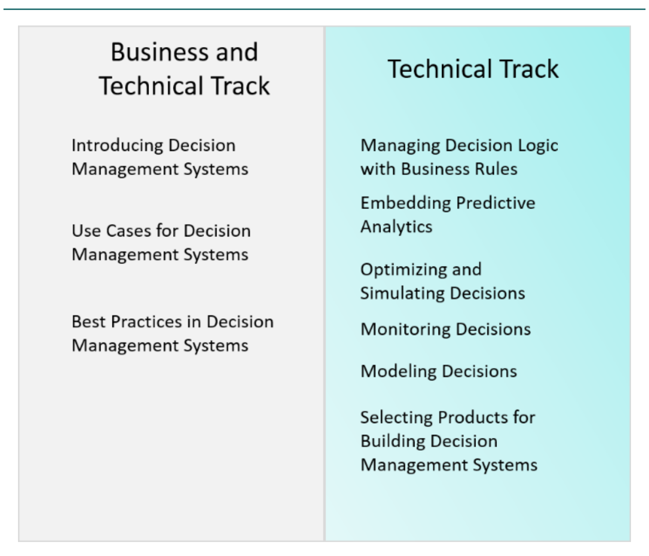
© 2017 Decision Management Solutions

Navigating the Report

The *Decision Management Systems Platform Technologies Report* is a set of documents describing the best practices and technologies for building Decision Management Systems.

1. *Introducing Decision Management Systems*
2. *Use Cases for Decision Management Systems*
3. *Best Practices in Decision Management Systems.*
4. Five Key Capabilities
    4.1. *Managing Decision Logic with Business Rules*
    4.2. *Embedding Predictive Analytics*
    4.3. *Optimizing and Simulating Decisions*
    4.4. *Monitoring Decisions*
    4.5. *Modeling Decisions*
5. *Selecting Products for Building Decision Management Systems*

All readers should begin with *Introducing Decision Management Systems* as it gives an overview of the category, technologies and rationale.



Business and technical readers can continue with *Use Cases for Decision Management Systems* and *Best Practices in Decision Management Systems*.

Technical readers are recommended to read the five Key Capabilities documents (*Managing Decision Logic with Business Rules, Embedding Predictive Analytics, Optimizing and Simulating Decisions, Monitoring Decisions and Modeling Decisions*) to better understand the component technologies of Decision Management Systems. *Selecting Products for Building Decision Management Systems* will be useful as part of assessing technology needs.

More information on the report, its scope, reproduction and more is in the final section **About The Decision Management Systems Platform Technologies Report.**

© 2017 Decision Management Solutions

Overview

Embedding predictive analytics requires a software component for the creation, validation, management, deployment, and ongoing re-building of predictive analytic models. Such a Predictive Analytics Workbench allows a data miner, data scientist, analytics professional, or business analyst to explore historical data and use various mathematical techniques to identify and model potentially useful patterns in that data.

For the purposes of this report, we are not concerned with the use of data mining or predictive analytic workbenches for one-off research projects to answer a specific question or with the construction of statistical models per se. Only models that can be applied to a specific transaction or item to classify it or make a prediction about it are included. Other forms of data mining and predictive analytics can have tremendous value to an organization but they are not relevant to this discussion of Decision Management Systems.

The predictive analytic models created can predict a binary outcome (yes or no), provide a number (often representing a probability or ranking of likelihood) or a selection from a list (of products for instance). They might also cluster or group based on likelihoods or identify which items are associated with which other items.

Data mining and predictive analytics allow organizations to turn historical data into useful, actionable analytic insight.

Data mining and predictive analytic models are often grouped with business intelligence, reporting, and visualization under the general term “analytics.” Data mining and predictive analytics differ from business intelligence capabilities in several ways:

- They are focused on extracting meaning about the likely future rather than summarizing or understanding the past—they use historical data to make predictions about what is likely in the future.
- They are probabilistic rather than definitive in that they rarely if ever make a prediction that something concrete is definitely going to happen. Generally, they say how likely something is, make a prediction with a certain degree of confidence, or rank order a set of possible outcomes from most to least likely.
- They rely on mathematical algorithms rather than the visual processing power of humans to see patterns in data.

This last point has an important consequence for products being used to develop Decision Management Systems. These products must do more than define the right mathematical model and present it as mathematics, a visualization or a report. It must be possible to both produce a predictive analytic model and embed it into an operational Decision Management System.

© 2017 Decision Management Solutions

Architecture

A predictive analytic workbench needs to support a range of activities that are generally performed in a highly iterative way:

- Integration with a wide range of data sources so that data can be brought into a modeling environment for analysis.
- Cleaning, integration, summarization, and exploration of this data.
- The creation of an analytical dataset suitable for analysis.
- Automated or mostly automated analysis of very large numbers of records using a variety of algorithms.
- Creation of analytic representations based on this analysis.
- Validation of these models to prove they will be predictive with data not used to build them, as well as assessment of their effectiveness in making predictions.
- Deployment of these models into an execution environment or as code that can be independently executed.
- The definition and management of repeatable processes or workflows to handle all these steps so that they can be repeated with new data.

One of the most important facets of these kinds of workbenches is their support for an industrial scale process for building predictive analytic models. With organizations increasingly needing dozens or hundreds of models, a more industrial process is called for. This does not eliminate the skill of a modeler, but it does require more repeatability, automation, and scalability in the way predictive analytic models are built and managed. This is where a predictive analytic workbench is essential.

A predictive analytics workbench gives data miners and possibly business analysts the ability to derive useful probabilities about the future from potentially large amounts of data about the past. These probabilities may group or segment customers or other records, identify the propensity of someone to do something (e.g., buy, churn, respond, visit), determine the strength of an association between two records, or identify what is likely to be the best combination among many possible ones.

Capabilities

Data Management

Predictive analytic models are typically built from a large amount of data, often pulled from multiple data sources. A predictive analytic workbench must be able to

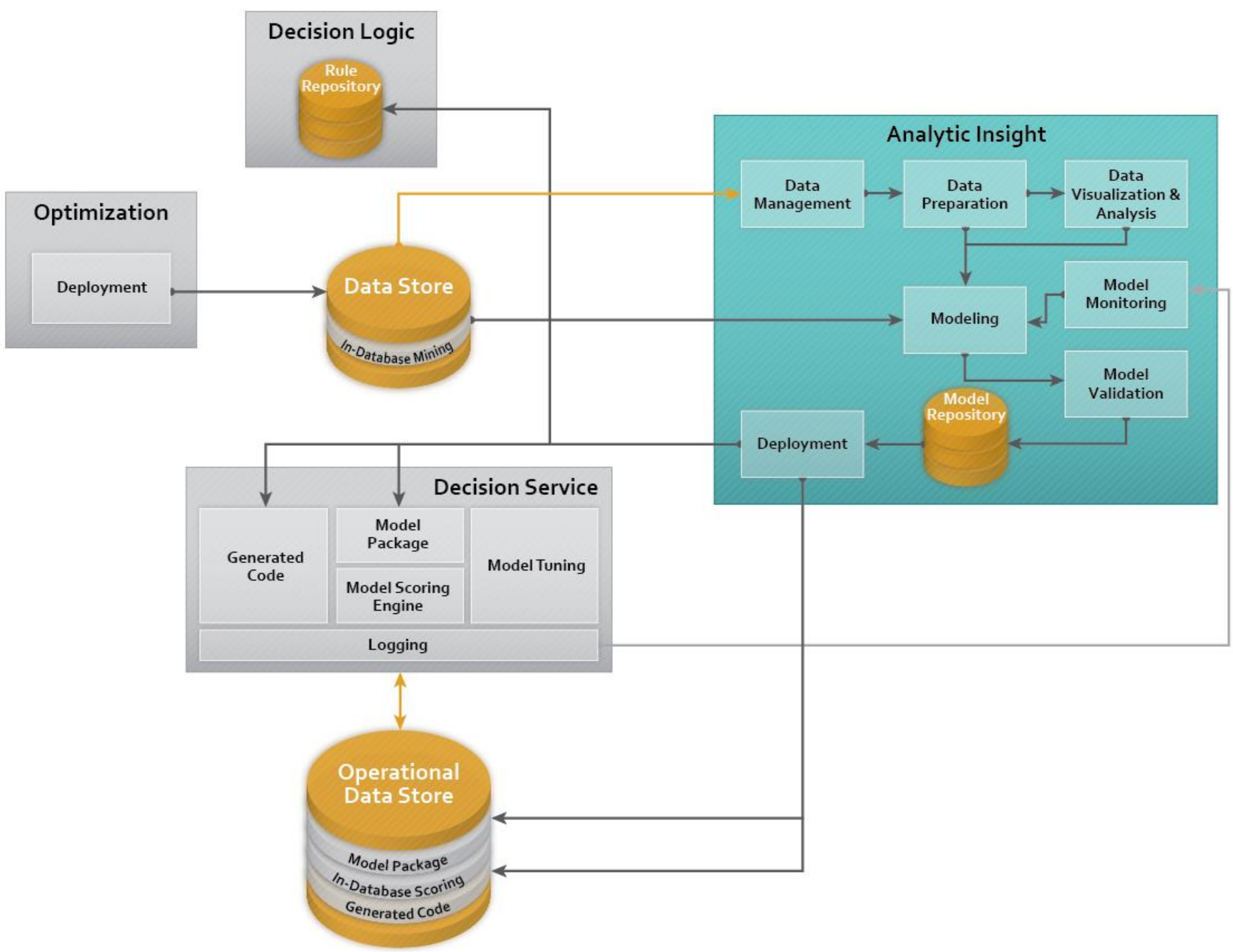
© 2017 Decision Management Solutions

connect to and retrieve information from a variety of structured and unstructured data sources as well as flat files of various kinds.

Data Preparation

The data available is often not immediately suitable for the construction of predictive analytic models. A predictive analytic workbench provides a variety of tools to allow the cleanup and integration of data prior to modeling. These tools include renaming and re-categorizing data fields, imputing missing values, filtering outliers, extracting samples and transforming data to make it more suitable for modeling. The end result of this data preparation work is what is often called an analytical dataset—a large set of data attributes (some original, some derived) with any hierarchical structure “flattened” into a single list of attributes.

Figure 1. Capabilities for Embedding Predictive Analytics



© 2017 Decision Management Solutions

Data Visualization and Analysis

Modeling efforts typically begin with exploration of the data available to develop some understanding of the data and of the patterns in that data. A rich set of visualization and graphical tools as well as statistical analysis routines help find the hidden patterns and relationships that might drive an effective model. These tools are often used in conjunction with the data preparation tools so that problems found in graphing the data, for instance, can be corrected in a data preparation routine. The same visualization and analysis tools will also be used to assess model outcomes once models have been developed.

Predictive Modeling

At the core of a predictive analytics workbench is a model creation environment suitable at least for data miners and other analytic users. The modeling environment might also allow business analysts to create and manage the modeling process—typically through a combination of automation and simplified interfaces.

Some predictive analytic workbenches are designed for expert users. Some are primarily aimed at these experts but provide simplified interfaces that aim at a broader audience. Some are designed with a single environment that works for both expert and less expert users. While the style of interface and its expectations can vary, all these workbenches create predictive analytic models and related resources in some form of shared repository.

The modeling environment typically involves laying out a series of steps that will result in the construction of a model or models that can be evaluated for performance. Steps will include data preparation and analysis as well as the execution of one or more algorithms from an extensive set. Algorithms supported include clustering, association, linear and logistic regression, decision trees, support vector machines, Bayesian modeling and nearest neighbor techniques to name a few. It is increasingly common to find ensemble models where several techniques are applied, or one technique is applied with different parameters, and the results aggregated in some fashion to create a single, overall ensemble model[1].

Some predictive analytic workbenches can take advantage of in-database modeling engines that can handle some of the data preparation tasks as well as execute the modeling algorithms themselves on the database server that contains the data being analyzed. This improves performance by eliminating the need to move data from the database to a separate analytic server and takes advantage of the increasingly powerful servers supporting data infrastructure.

[1] Ensemble models are explored extensively in "Handbook of Statistical Analysis and Data Mining Applications" (Elsevier, 2009) by Nisbet, Elder and Miner.

© 2017 Decision Management Solutions

Model Validation

Regardless of which technique or set of techniques is used, model performance assessment and comparison tools are used to see how well a model performs. Different models can be compared and tools such as lift curves (comparing selection using a model to a random distribution) used to see how effective the model would be in production. These tools typically use new data (data that was not used to build the model) to see how predictive the model would be once deployed.

Deployment and Scoring

Once the final model or models have been identified they must be deployed. A predictive analytic workbench may allow multiple approaches to deployment:

- Models can be used to score data in a batch mode, applying the results back to the database that contained the data from which the model is built.
- Some predictive analytic workbenches can act as a real-time scoring server using their own scoring engine and providing a web services or other API to allow it be called during decision-making.
- Scoring code can also be generated (as C or Java, as SQL or as business rules) so that it can be deployed to a Decision Service for real-time scoring.
- In-database scoring is also available, with the definition of the model being pushed to the analytic infrastructure where the scoring engine is running.
- Several predictive analytic workbenches also allow models to be generated using the Predictive Model Markup Language[2] (PMML), allowing the model to be executed by any business rules or scoring engine that supports this standard.

Model Monitoring

Models are built from a snapshot of data. As such they “age”—as time passes the data being fed into the deployed model may look less and less like the data from which it was built. A predictive analytic workbench needs tools to monitor deployed models to see how their performance is varying over time and to identify variations in performance or in data distributions. Many new models are initially deployed to challenge an existing model and the performance of both the original “champion” model and the new “challenger” model need to be compared to see if the challenger is good enough to replace the champion. Model monitoring tools need to identify opportunities to refresh and retrain models and to provide tools to make it easy for users to rebuild models to take advantage of new data.

[2] PMML is an open standard managed by the Data Mining Group that provides a standard XML representation of predictive analytic models so that they can be exchanged between multiple products.

© 2017 Decision Management Solutions

Model Tuning

Some predictive analytic workbenches provide components for automated model tuning and updating. These machine learning techniques monitor the performance of a model as it is used in deployment and automatically adjust its underlying equation based on that performance. Some of these environments can start with no model and gradually build a predictive model based on the results of random experiments while others are designed to be used with pre-defined models. Model Tuning can be left to run forever or it can tune the model within defined boundaries and flag a model for re-building if its performance starts to drift outside those boundaries. Model Tuning capabilities are often deployed in a Decision Service if that is where the model is being executed.

Repository

A predictive analytics workbench should offer an enterprise-class repository for storing and managing predictive analytic models. This repository may be a complete decision management repository that also stores business rules and optimization models. It should provide access control and security, audit trails for changes made to models and versioning.

There is a growing category of software products that allow business rules to be specified and managed alongside predictive analytic models built in the same product. The degree to which large numbers of business rules can be managed and the range of predictive analytic models that can be built varies and such a combined product may not therefore support the complexity required for a specific Decision Management System. These products typically allow models built in other predictive analytic workbenches to be integrated also.

In-database Analytics

In-database analytics can mean exactly that—analytic capabilities embedded in a relational or columnar database—though the phrase is also used to describe analytic capabilities embedded in data warehouse software, in data appliances and increasingly in Hadoop clusters.

In-database analytic capability is delivered as a set of libraries, User Defined Functions, that deliver analytic or data mining functions such that they can:

- Access the data in the database, data warehouse, appliance or Hadoop file system in situ, without needing to extract it to some interim format.
- Directly use the memory, parallel processing capabilities and load balancing/processor management of the data infrastructure.
- Be accessed both from specialist analytic tools (for model creation or data quality tasks for instance) and from operational systems.

© 2017 Decision Management Solutions

In-database analytic capabilities are specific to a particular database, data warehouse, data appliance, or Hadoop distribution. Many vendors offer support for multiple data infrastructure platforms. Some capabilities are provided by the data infrastructure vendors, some by specialty analytic vendors, and some through partnerships between analytic and data infrastructure vendors.

For more information on In-Database Analytics, download our white paper sponsored by SAS.

For Decision Management Systems, the core capabilities to look for today in an in-database analytic product are:

In-Database Data Preparation and Quality

Data preparation, integration, and cleaning often consumes 60-70% of the time and effort on an analytic project. In a traditional approach, data is extracted from the data infrastructure in which it is stored, processed through various preparation steps, and then presented to the analytic modeling algorithms that need it.

With in-database capabilities, however, these steps all execute in-database. This means the original data is not extracted from the database but is processed in situ. The resulting cleaned and transformed data may be stored in the data infrastructure or passed out to a predictive analytic workbench for further processing. The net is that data required for analytic modeling is transformed in-database.

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, while the ability to apply the schema of data only when it is being read means Hadoop is very flexible for a wide variety of data types. Storage and processing are streaming-centric, and this enables the environment to handle fast moving data.

Hadoop has a lot of potential for companies adopting predictive analytics, but it must be applied in context. Beginning with a business problem—a decision that must be made—determines the analytics that will be required and thus what kind of data will be required. This creates a use case for Hadoop by identifying a business problem that requires data not already available in existing infrastructure.

In-Database Predictive Model Development

In-database model development allows predictive analytic models to be developed using algorithms embedded in the data infrastructure. These algorithms access tables and views directly to get the data they need, process the data using the data infrastructure's processing capabilities, and create a predictive analytic model. This model may be stored in the data infrastructure for in-database scoring or it may be passed out for use elsewhere. These capabilities may be integrated with an external predictive analytic workbench.

© 2017 Decision Management Solutions

R is fundamentally an interpreted language for developing these models, known for statistical computing and the ability to graphically display results. 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.

In-Database Model Deployment and Scoring

In-database model deployment and scoring infrastructure takes models developed using some combination of in-database modeling infrastructure and a predictive analytic workbench and executes them in an operational datastore so they are available to operational systems accessing that datastore. This generally involves turning models into UDFs or stored procedures that can be called using SQL and that take database fields as input.

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 data dictionary, data transformations, and models.

PMML offers an open, standards-based approach to operationalizing predictive analytics. Support for PMML is increasingly broad-based with analytic tools, databases, data warehouses, and server deployments. Business rules and other development environments also increasingly support it.

In the future, more extensive support for analytic model management and for wrapping analytics in business rules for in-database decision-making will become increasingly important.

For more information on Standards in Predictive Analytics, download our white paper sponsored by Revolution Analytics, Software AG, and the Data Mining Group

© 2017 Decision Management Solutions

Next Steps

An optimization suite is an environment for defining and solving mathematical models and for simulating the differences between multiple similar mathematical models. An optimization suite allows a modeler or business analyst to define a business objective and a set of constraints and then "solve" this problem to see how best to run the business. Optimization suites support what is sometimes called Operations Research or Management Science.

Continuing reading The Decision Management Systems Platform Technologies Report

Learn More:

We have extensive experience helping organizations like yours define, configure and implement Decision Management Systems that deliver on the value propositions described in this Report. Our clients are leading companies in insurance, banking, manufacturing, telecommunications, travel and leisure, health management, and retail.

- Client Case Studies
- Our Services

Contact Us Today for a Free Consultation

© 2017 Decision Management Solutions

About The Decision Management Systems Platform Technologies Report

This report is focused on platform technologies used to build custom Decision Management Systems and our goal is to be comprehensive within this scope. Many vendors have developed powerful pre-configured Decision Management Systems focused on solving specific decision problems such as loan underwriting, claims handling or cross-channel marketing. For many organizations these solutions are ideal but they are not the focus of this report. Similarly, there are vendors that build custom Decision Management Systems for their customers and that have developed powerful platforms for doing so. If such a platform is not for sale to those building their own solutions, then it is out of scope for this report.

In both these scenarios the report's discussions of what kinds of functionality is useful, best practices and characteristics for suitable products may well be useful in the selection of vendors but some interpretation will be necessary.

Vendors and products in scope for the report are added continually. First Looks are also posted to www.JTonEDM.com as they are completed. Each new version of the report will be made available at decisionmanagementsolutions.com/decision-management-platform-technology/.

This report can be freely circulated, printed and reproduced in its entirety provided no edits are made to it. Please email info@decisionmanagementsolutions.com if you would like to publish an extract. Quotes from this report should be correctly attributed and identified as © 2017, Decision Management Solutions.

While every care has been taken to validate the information in this report, Decision Management Solutions accepts no liability for the content of this report, or for the consequences of any actions taken on the basis of the information provided.

If you know of products you believe should be included or have other feedback, please let us know by sending us an email info@decisionmanagementsolutions.com.

CONTACT US

Decision Management Solutions specializes in helping organizations build decision-centric, action-oriented systems and processes using decision management, business rules and advanced analytic technologies.

www.decisionmanagementsolutions.com Email: info@decisionmanagementsolutions.com