Deploying Data Mining Models with 4Sight

A White Paper

Contents

Executive Summary

Examples abound of businesses employing predictive analytics to optimize operations and achieve a tangible return on investment. Although the value afforded to businesses through the insights and predictive power provided by data mining is clear, it is often the case that processes for achieving that value are less clear. Deployment of predictive analytics solutions can be a stumbling block to achieving improved decision-making afforded by data mining outcomes.

This white paper outlines flexible options for the deployment of predictive models using the 4SightBI Suite. In particular, we consider the common scenario of batch scoring to rapidly update predicted scores. 4Sight's 100% Java-based component-oriented architecture facilitates lightweight, easily customizable solutions.

Example Use Case

Consider a scenario where you have data stored in a CRM system, such as Salesforce.com for example, and you want to use a predictive model to produce a score for the likelihood of winning each of your open sales opportunities. Assuming you have a predictive model, such a scenario would require a process that could extract raw customer and opportunity data, derive the predictive fields required by the model, generate the scores, and then update the CRM system with the results. Ideally, you would want to derive such a process without having to resort to custom coding. Having a runtime environment with a small footprint, but with the ability to scale as data volumes increase, would also be advantageous. Furthermore, having the ability to "drop in" different data mining models or refresh the current model with minimal fuss would ensure that the process remains in-sync with the changing nature of your business.

The combination of Pentaho's enterprise-class data integration capabilities and powerful data mining suite makes developing, deploying and maintaining such a process simple. Pentaho's commercial open source model also ensures maximum return on investment, with no large, up-front software license fees and ongoing support and maintenance at a fraction of the cost associated with proprietary offerings.

Advantages of a 4Sight Data Mining Deployment Solution

In order to achieve a high return on investment, a data mining deployment should allow for:

- Easy and rapid model development
- Open and standards-based architecture to facilitate integration
- Deployment options that can encompass and automate the entire data mining lifecycle
- A low overall TCO

Easy and rapid model development

4Sight Data Mining, based on the Weka project, provides a modern environment for constructing analytical models. It provides a comprehensive suite of data mining tools with more than 200 state-of-the-art algorithms for classification, regression, clustering, attribute selection and data pre-processing. Furthermore, Weka's large, active community and strong ties to academia ensure that the toolkit remains up-to-date with advances in the field. Easy-to-use graphical interfaces and full support for the whole process of experimental data mining ensure the rapid development and validation of predictive models. With no

dependencies on external libraries, the entire 4Sight Data Mining toolkit can be deployed as a single 5Mb Java Jar archive.

Open and standards-based architecture

4Sight's core products are all distributed under open source licenses. This means all source code is freely available, which in turn, provides a degree of protection for your investment not necessarily afforded by proprietary alternatives. Pentaho's products also aim to support open standards. For example, 4Sight Data Mining supports import of externally created models in PMML (Predictive Modeling Markup Language) format. In situations where considerable investment resides in existing models created with other tools, support for PMML protects that investment and facilitates easy transition from one data mining runtime environment to another.

Full lifecycle deployment

4Sight Data Integration (PDI) provides an ideal platform on which to deploy predictive solutions. Its modular, 100% Java architecture facilitates natural integration with 4Sight Data Mining. PDI has built-in support for clustered and distributed execution that allows predictive scoring solutions to scale as data volumes increase. PDI's ability to not only execute a predictive scoring algorithm, but also an entire data mining process, allows the complete process of training and refreshing data mining models to be automated. This, in turn, can reduce or eliminate the costs associated with regenerating models manually.

Lower total cost of ownership

The absence of large, upfront software license fees, with low annual subscription costs for support and maintenance, makes a 4Sight predictive scoring deployment cost effective when compared with proprietary alternatives.

Data Extraction and Preparation with 4Sight Data Integration (PDI)

4Sight Data Integration's streaming, engine-driven approach not only provides an ideal environment in which to deploy predictive scoring solutions, its powerful extract and transform operations are a natural complement to 4Sight Data Mining's advanced data filters. PDI can easily export data sets in Weka's native ARFF format to be used immediately for model creation. Features of PDI include:

- Metadata-driven approach with graphical drag-and-drop development environment
- Rich transformation library with over 100 out-of-the-box mapping objects
- RDBMS-based or file-based repositories
- Extensive data source support including most proprietary and open-source databases, Excel, CSV, XML, flat files, web services etc.
- ETL engine can be clustered and scaled to large data volumes
- Integration with the 4Sight BI Platform provides scheduling, security integration, auditing and performance monitoring

For more information on the features of 4Sight Data Integration visit http://www.4SightBI.com.

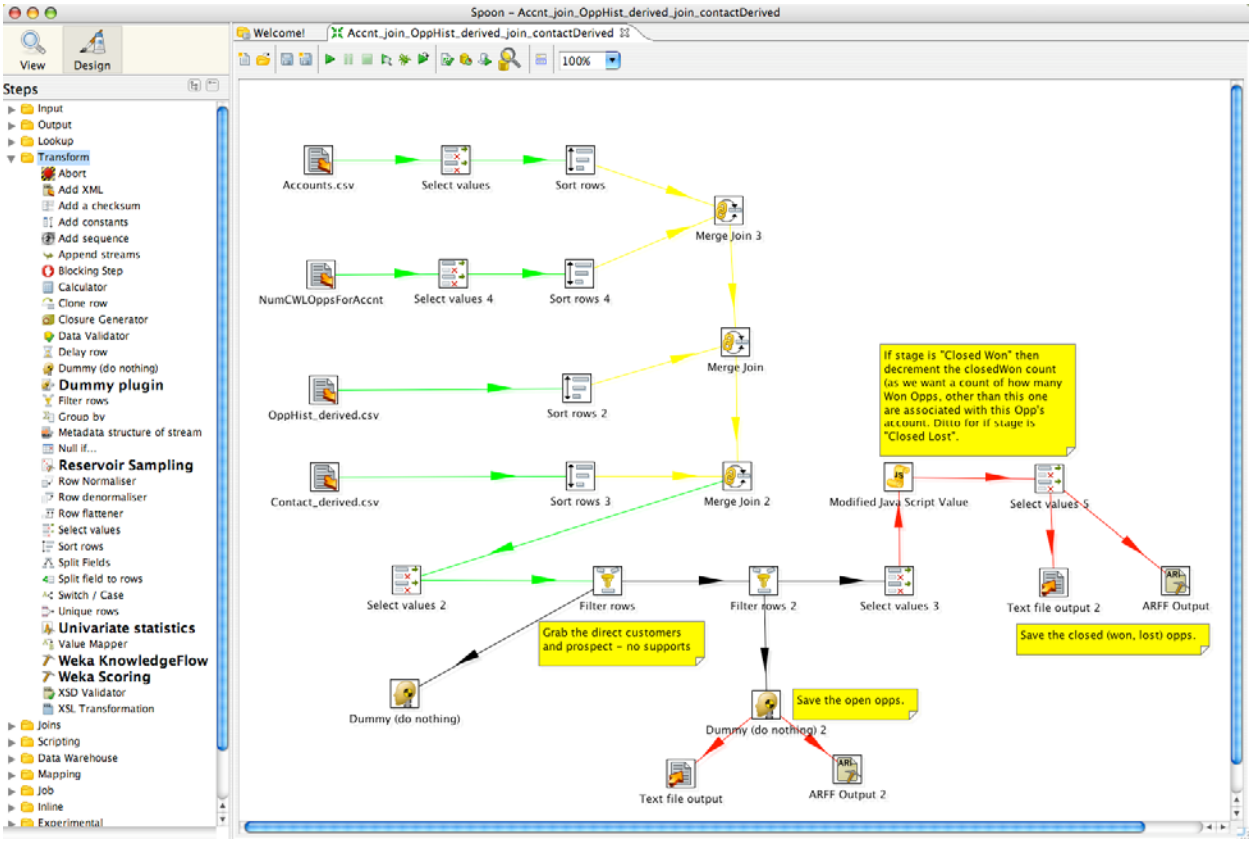

Fig. 1. PDI's graphical ETL development environment

Model Development with 4Sight Data Mining

4Sight Data Mining provides a complete environment to develop and refine predictive models according to industry standard methodologies such as CRISP-DM (CRoss Industry Standard Process for Data Mining). An extensive range of data mining algorithms combined with easy-to-use graphical front ends enable models to be developed and validated with ease. Features of 4Sight Data Mining include:

- 69 data pre-processing filters
- 116 classification/regression algorithms
- 11 clustering algorithms
- 18 attribute/subset evaluators + 12 search algorithms for feature selection
- Explorer and Knowledge Flow applications for exploratory data mining and data mining process development respectively
- Experimenter application for large-scale experimental comparison of learning algorithms using statistical techniques such as repeated cross-validation and significance testing
- Graphical visualizations such as scatter plot matrices, ROC/lift charts, tree and graph visualizations, etc.
- Export of models and data mining processes in serialized binary or XML format
- Import of externally created models in PMML format

For more information on the features of 4Sight Data Mining visit http://www.4SightBI.com.

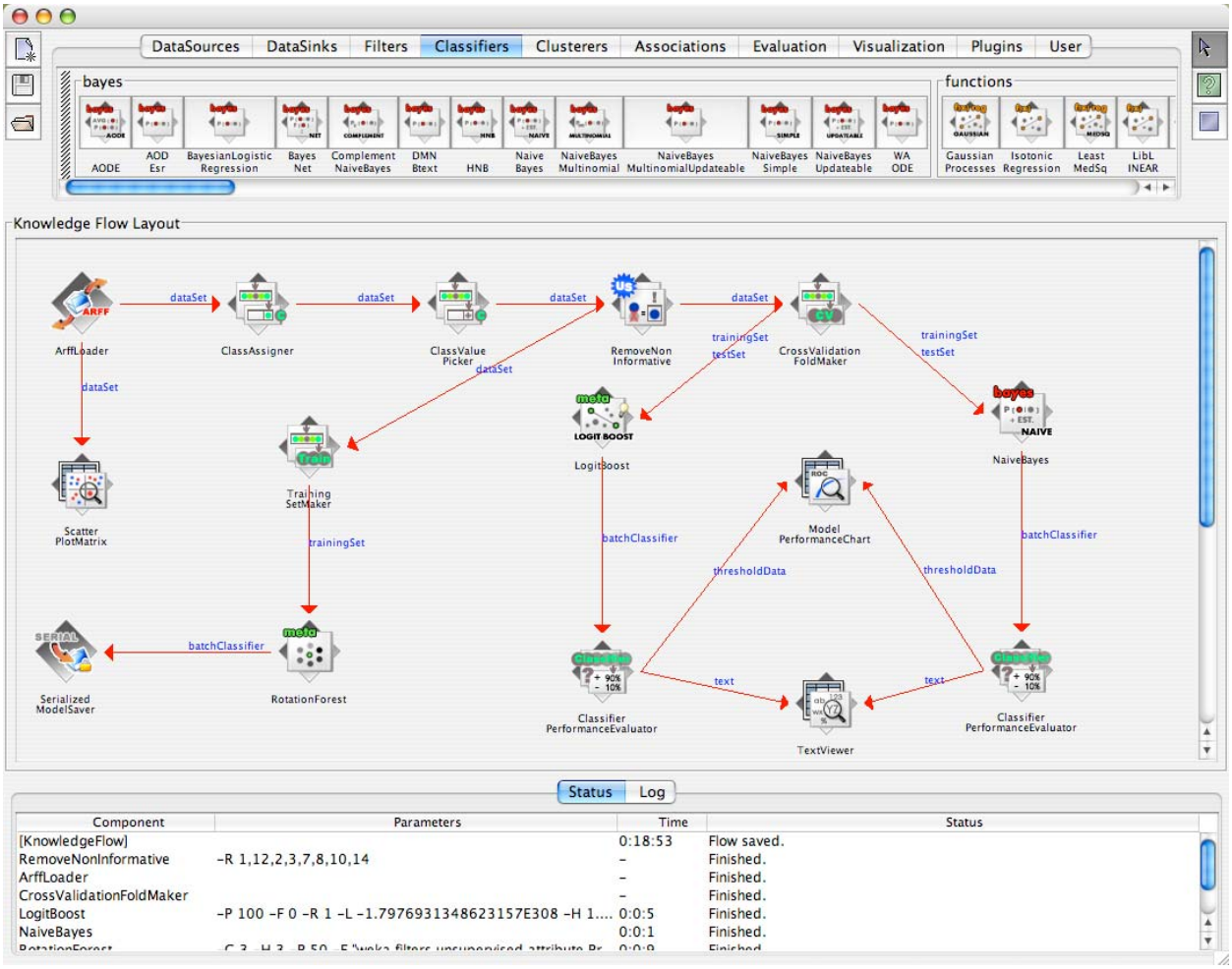

Fig. 2. Developing a data mining process with 4Sight Data Mining's Knowledge Flow

Model Deployment and Refresh with 4Sight Data Integration

Deploying a predictive model – either a native Weka model or one expressed in PMML – as part of an ETL process in PDI is easily achieved using the Weka Scoring transformation node. Features of this component include:

- Support for serialized Weka models in either binary, XML or PMML format
- Models can be sourced from the file system at runtime or imported and stored as part of PDI transformation metadata in a repository
- Automatic mapping and type checking of incoming fields against those used by the model
- Display of textual description of model
- Support for classification, regression and clustering models
- Output of predictions as labels (classification or clustering) or probability distributions
- Ability to update models that are capable of incremental training

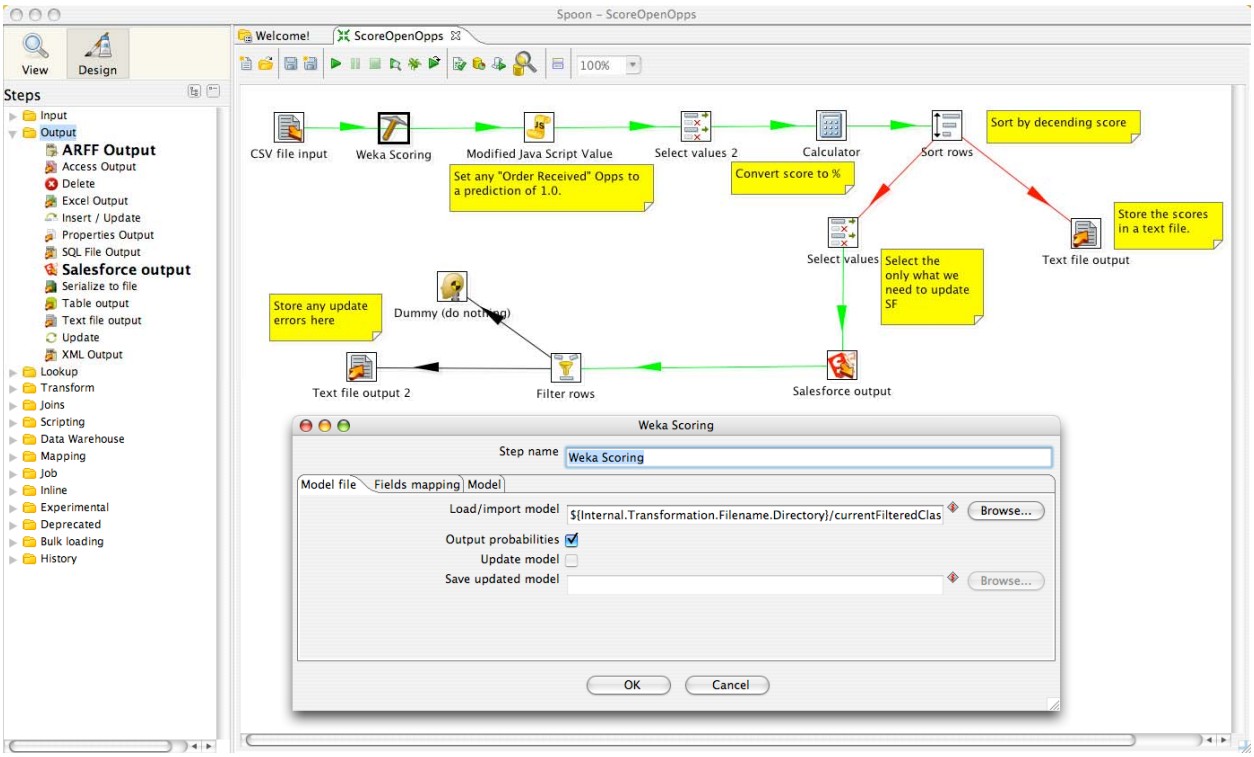

Fig. 3. A PDI transformation to score sales opportunities using a 4Sight Data Mining predictive model

Over time, as new data is collected, the predictive performance of a model may decrease. This can be caused by changes in the underlying distribution of the data and is sometimes referred to as "concept drift". For example, with our sales opportunity scoring scenario, the types of customers we sell to may gradually change over time as our business evolves. This necessitates the periodic recreation or refreshing of the predictive model using up-to-date data. Automating this process as part of a scheduled ETL job is simple using the Weka Knowledge Flow transformation node. This node allows an entire data mining process (including the export of freshly trained models) to be executed by PDI. Features of the PDI Weka Knowledge Flow component include:

- Support for serialized Weka Knowledge Flow processes in either binary or XML format
- Knowledge Flow processes can be either sourced from the file system at runtime or imported and stored as part of PDI transformation meta data in a repository
- Incoming data from the ETL transformation can be injected into the Knowledge Flow process or the Knowledge Flow process can simply be triggered (and source it's own data natively)
- Built-in support for reservoir sampling can be used to downsample incoming data from the ETL transformation to a size suitable for various batch learning algorithms
- Textual output from a model or data produced by a Weka filter can be passed on to downstream PDI transformation steps
- Embedded Knowledge Flow editor allows entire data mining processes to be created inside of PDI's graphical design environment

About 4Sight Business Intelligence, Inc.

4Sight Business Intelligence, Inc. provides business intelligence solutions designed specifically for Property & Casualty insurance industry. Our focus is providing off-the-shelf, documented, solutions with ad-hoc reporting, what-if analysis, dashboards, scorecards, standard reports, data mining and pre-built ETL processes which allow for quick configuration and implementation of typically less than 45 days. Our products are for carriers and MGAs of all sizes. Some clients include Mutual Aid Exchange, AXA Re P&C, Preferred Auto, Chautauqua Patrons Insurance, American Physicians, American Leader and the NC Farm Bureau, among many others. Our products support most platforms and your choice of database. For more information call 888-PCBI-123, visit www.4SightBI.com or email Info@4SightBI.com