Predicting the future, Part 4: Put a predictive solution to work

Alex Guazzelli July 10, 2012

This is the last article of a four-part series focusing on the important aspects of predictive analytics. Part 1 offered a general overview of predictive analytics. Part 2 focused on predictive modeling techniques, the mathematical algorithms that make up the core of predictive analytics. Part 3 put those techniques to use and described the making of a predictive solution. This final article focuses on the deployment of predictive analytics, or the process of putting predictive solutions to work.

View more content in this series

Introduction

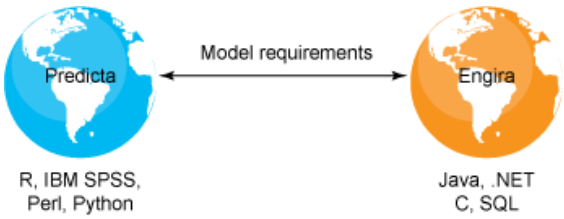
Putting a predictive solution to work has traditionally been a laborious process, involving quite a bit of time and resources. The advent of the Predictive Model Markup Language (PMML) has completely changed this scenario. The same people that build it can now put a predictive solution to work in a matter of minutes. As the de facto standard for predictive analytics, PMML is supported by all top data mining vendors, both commercial and open-source. After building a predictive model, it's easy to export it into a PMML file. Then you can directly deploy that file into a PMML-based scoring engine where it is available for execution. Given the big data era we live in, predictive models should benefit from fast deployment and execution. The availability of scoring solutions in the cloud and in-database make it possible for predictive analytics to fulfill its promise and crack the big data code.

This four part series has covered many predictive analytics topics, mostly related to model development. I have described many applications of predictive analytics. For a predictive solution to be applied to new data though, it needs to be operationally deployed. That is, it needs to be moved from the scientist's desktop, where it was built, to the operational environment, where it will be put to work.

As it turns out, putting a model to work is not an easy feat. To be operationally deployed, a predictive solution needs to successfully bridge the gap between two very different worlds. I call these, planets *Predicta* and *Engira*. Planet Predicta is populated by data scientists with expertise

in statistics, data mining, and language skills such as Perl and Python. Planet Engira, on the other hand, is populated by IT engineers with expertise in Java™, .NET, C, SQL. Without a common language, moving a predictive solution from Predicta to Engira can get lost in translation (see Figure 1). After a model is developed, a data scientist writes a word document describing the model then sends it to an IT engineer who will start coding the model into the operational environment. Questions may arise and diplomatic cables sent back and forth between the two. However, more often than not, when the model is considered ready to be put to work, the engineer that recoded it realizes that the scores it produces in production do not match the scores the data scientist got with the same model in development. The engineer must communicate with the scientist to resolve discrepancies. This process may take anywhere from three months to a year.

Figure 1. Model deployment process can get lost in translation between planets Predicta and Engira without a common language between the two.



Thankfully, an event occurred that shortened model deployment from months to minutes. This event was the advent of a common language that data scientists and IT engineers alike could understand. This language is called PMML.

PMML

PMML is the brainchild of the Data Mining Group (DMG), a consortium of companies that works together to define it. All the top statistical and data mining tools, commercial and open-source, support PMML. In this way, a model may be developed on the planet Predicta and directly sent to planet Engira for instantaneous deployment. No translation, no recoding, no custom code is necessary. With PMML, the moving of a predictive solution from the scientist's desktop to the operational environment becomes a very simple task.

PMML allows companies and individuals to use a single language to represent a predictive solution in its entirety, regardless of the environment in which it was built. Part 3 of this series covered all the phases involved in the making of a predictive solution, from data pre-preprocessing and model building all the way to post-processing of model scores. PMML is able to represent all these phases in a single file. It can also represent solutions that include multiple models or a model ensemble.

PMML is XML-based. Its schema follows a well-defined structure in which elements and attributes are used to define:

1. The input data through element `DataDictionary`
2. Invalid, missing and outlier value handling strategies through element `MiningSchema`
3. Data pre-processing via element `TransformationsDictionary`

4. A myriad of modeling techniques via specific model elements such as: `NeuralNetwork`, `TreeModel`, `SupportVectorMachineModel`, `Scorecard`, and `RegressionModel`
5. Post-processing of model outputs via element `Output`

PMML also contains other language structures including specific elements for model verification and model explanation and evaluation. Given that PMML can represent a predictive solution in its entirety in a clear and structured way, we can use it to unveil the secrecy and the black box feeling many people have when it comes to predictive analytics. A company can us PMML as the lingua franca not only between Predicta and Engira, but also between service providers and external vendors. In this scenario, it defines a single and clear process for the exchange of predictive solutions. It becomes the bridge not only between data analysis, model building, and deployment systems, but also between all the people and teams involved in the analytical process. This is extremely important, because we can use it to disseminate knowledge and best practices, and ensure transparency.

The latest version of PMML, version 4.1, was released in December 2011. As a language, however, it has been around for more than 10 years and for this reason has achieved a great level of maturity and refinement. As a DMG representative introduced to PMML many years ago, I was taken aback by its range and power as well as all the benefits it brings to any organization that wants to profit from the predictive value inherent to its historical data.

Representing a predictive solution in PMML

All the top statistical tools currently available on the market for model development export models in PMML. Some of these also provide import functionality so that a model can be visualized and further refined. An open-source environment worth mentioning is KNIME (see Resources), which imports and exports many PMML models. Another is the R project for statistical computing (see Resources). A myriad of commercial products also support PMML. In this article, I focus on IBM SPSS Statistics, which is able to export PMML for a variety of predictive techniques. IBM SPSS Statistics can export PMML for data pre-processing too, which is an important piece of the predictive puzzle.

Part 3 of this series described how to use IBM SPSS Statistics to automatically perform data pre-processing. The goal is to augment the predictive value of the raw input data in order to improve the accuracy of the resulting model. For that, choose **Transform menu > Prepare Data for Modeling** and click on **Automatic**. On the tab-delimited window called "Automatic Data Preparation", click on the **Settings** tab and select item **Apply and Save**. Check the box "Save transformations as XML" and enter a file name. This is the file that will contain the transformations in PMML format. After you select the option **Prepare Data for Modeling**, you can also choose **Interactive**. This option will bring up a tab-delimited window called "Interactive Data Preparation". In this case, you will need to click on the **Analyze** button first before saving the PMML file containing the resulting transformations. In both cases, you will end up with a PMML file that fully describes the data pre-processing steps taken by IBM SPSS Statistics in preparation for modeling. To understand how PMML represents data transformations, Zementis released an interactive tool called the *Transformations Generator*. This tool allows users to graphically represent a variety of

transformations and automatically export them in PMML, which can subsequently be merged with a model file (see Resources).

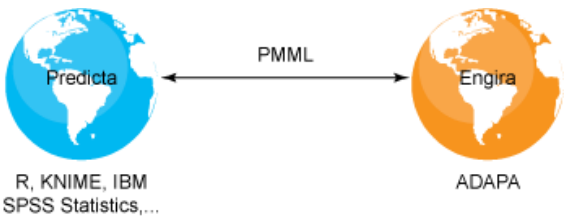
The data preparation process will also create additional data fields in the IBM SPSS Statistics data editor, which can be used for model training in conjunction with any other raw input fields. As described in Part 3, to train a neural network model, for example, simply choose **Analyze menu > Neural Networks** and select **Multilayer Perceptron**. After making all the appropriate selections in the multi-tab window, select the **Export** tab, check "Export synaptic weight estimates to XML file" and enter the name of the PMML file to which you want your neural network model to be saved. In a similar fashion, IBM SPSS Statistics also allows for many other predictive techniques to be exported in PMML.

The PMML file for the neural network model will also contain scaling of numerical inputs and discretization of categorical variables. However, if your model used any of the variables automatically discovered by IBM SPSS Statistics, you will need to merge the two previously described PMML files (automatic data pre-processing and model into a single file). For that, choose **Utilities menu > Merge Model XML**. You will be faced with a window in which you can enter the name of the "Model XML File" (the PMML file containing the predictive model) and the "Transformation XML File" (the PMML file containing the automatic or interactive data pre-processing steps). Also enter a name for the "Saved Merged XML File". Click **OK**. Now your predictive solution is completely represented in PMML.

Operational deployment with PMML

Once represented in PMML, a predictive solution can be deployed in minutes. At Zementis, we have created a PMML-based predictive analytics decision management platform called ADAPA. It is able to consume predictive solutions expressed in PMML and execute them in real-time. Since ADAPA lives on the operational side, it frees Engira resources from the burden of custom coding and it allows Predicta resources the opportunity to deploy predictive solutions on their own. As shown in Figure 2, after a predictive model is exported from IBM SPSS Statistics, or any other PMML-compliant tool such as R and KNIME, the data scientist can directly upload it into ADAPA where it is ready to be used.

Figure 2. With PMML, predictive solutions built by Predicta resources can be deployed in minutes



Once uploaded in ADAPA, predictive models can be managed and executed directly through a web console or through web services. In the latter case, scores and predictions can be directly embedded into any application throughout the enterprise.

ADAPA is available as a traditional license for on-site deployment. It is also available as a service on the IBM SmartCloud. Once in the cloud, model execution benefits from a cost-efficient and scalable structure for computing via the Internet. Zementis also offers the Universal PMML Plug-in (UPPI) for in-database scoring and for Hadoop. UPPI is currently available for the EMC Greenplum database, SAPSybase IQ, and IBM Netezza. In this way, models expressed in PMML can be easily deployed inside the database and reside next to the data itself. Applications that require in-database scoring usually involve big data. IBM estimates that 90 percent of the data that exists today was generated in the past two years alone. That gives you an idea of how much data we as a society are generating and collecting on a daily basis. Big data means lots of data. To benefit from all the secrets lurking underneath this ever-expanding sea of data, fast deployment and execution of predictive solutions is paramount. Luckily, PMML combined with powerful databases, Hadoop, and cloud computing makes unlocking the value in big data possible.

Model execution in the IBM SmartCloud

At its core, cloud computing is a set of services that provide computing resources through the Internet. Large data centers deliver scalable, on-demand, and often-virtualized resources as a service, eliminating the need for investments in specific hardware, software, or on your own data center infrastructure. The term *cloud* is used as a metaphor for the Internet. Cloud computing allows for a variety of services, including storage capacity, processing power, and business applications. Accessing services on the cloud only recently became available as a secure and reliable infrastructure. The IBM SmartCloud is a prime example of a generic cloud infrastructure. Powered by IBM, it provides dynamic compute capacity in the cloud through several data centers spread throughout the world.

Software as a service (SaaS) is a software license model in which a business or a user may access software through the Internet and pay for the right to use the software for a certain time period rather than acquiring a perpetual software license to be installed in-house. This is extremely advantageous for customers, since there are no upfront costs in setting up servers or licensing software and it minimizes the risk of purchasing costly software that may not provide adequate return of investment. Since the SaaS license model and cloud computing are both Internet-centric, more vendors combine them to deliver novel software solutions.

As mentioned before, ADAPA uses web service calls to allow automatic decisions to be virtually embedded into systems and applications throughout the enterprise. To minimize total cost of ownership, model execution in ADAPA is available as a service through the IBM SmartCloud (see Resources). This partnership between Zementis and IBM allows companies to deploy and execute predictive models and easily use produced scores and predictions to influence their day-to-day operations.

The process of launching a virtual ADAPA server in the IBM SmartCloud corresponds to the traditional scenario of buying hardware and installing it in a server room. The only difference is that the server in this case sits in the cloud, comes with a pre-installed version of ADAPA, and launches in just a few minutes, on-demand, and ready to be used.

Conclusion

Traditionally, the process of deploying a predictive solution and putting it to work was a lengthy one, taking months and draining invaluable resources from both the people responsible for building the models and those responsible for recoding it into production. The advent of PMML dramatically changed that. With PMML, the same people responsible for building the model can now deploy it in minutes.

As the de facto standard to represent predictive models, PMML can represent a predictive solution in its entirety, in a clear and structured way. PMML is supported by all the top statistical packages, including commercial and open-source. IBM SPSS Statistics, for example, is able to export data transformations as well as several predictive modeling techniques in PMML. A PMML file can then be easily deployed in a system such as ADAPA, the PMML-based Zementis scoring engine. Since ADAPA is available on the IBM SmartCloud, it benefits from a scalable and reliable infrastructure powered by IBM. Once in the cloud, predictive models can be accessed from anywhere at anytime and their scores and predictions embedded in any application throughout the enterprise through web services.

After reading this four part series, you should have a good understanding of predictive analytics and its applications. I started by telling you what it is. I described many of the predictive techniques that are at the core of predictive analytics. These techniques can learn patterns from historical data and detect them when presented with new data. I then described the making of an entire predictive solution. It all starts with a well-defined problem, followed by data analysis and pre-processing. Data is then presented to a predictive technique for model building, which is evaluated for accuracy. Discrimination thresholds are then set depending on the model accuracy and the cost associated with any prediction errors. Business decisions are then tied to different thresholds. Finally, when exported as a PMML file, a predictive solution is ready to be deployed and put to use. When that happens, predictive analytics has truly fulfilled its promise to learn valuable patterns from historical data and use them to predict the future.