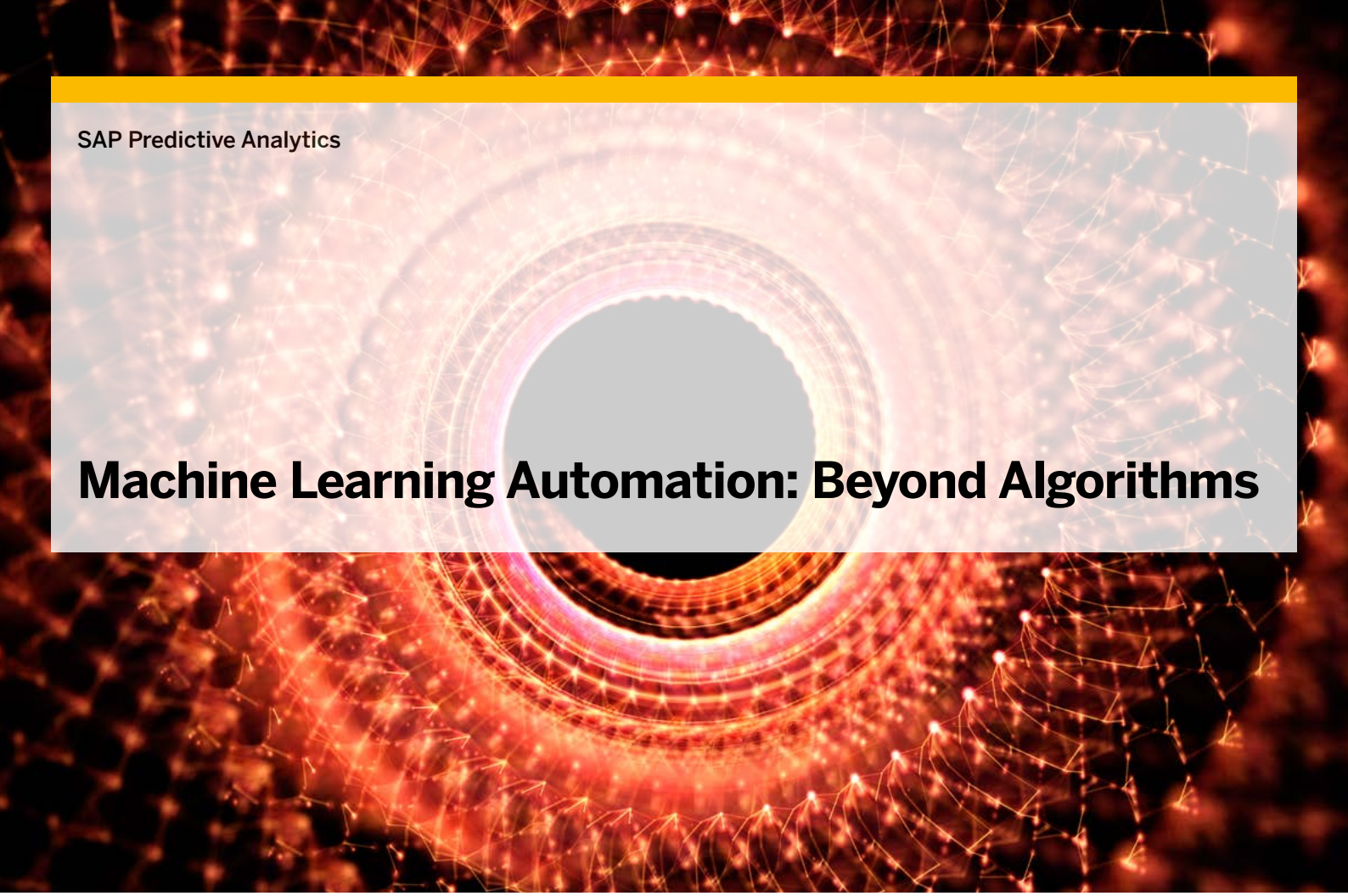


SAP Predictive Analytics

# Machine Learning Automation: Beyond Algorithms

## Table of Contents

Machine learning is going mainstream – it has already penetrated the customer relationship management and risk sectors, and it is a cornerstone of Big Data, the Internet of Things, and Industry 4.0. A new generation of tools, including SAP® Predictive Analytics software, focuses on the automation of a massive number of models built and applied on large data sets, and uses the existing skills of the enterprise. These tools are succeeding where most previous attempts have failed.

## DATA MINING FUNCTIONS

What happens if automated techniques that provide results as good as models handcrafted by data scientists come to the market? As we've seen with the industrial revolution and the advent of the assembly line, when disruptive technologies enter the market, productivity rises dramatically and pent-up demand drives rapid expansion. We are seeing this with our customers, who are using automation techniques to answer many more business questions than before with the same amount of staff and skills.

Another consequence of employing automated analytic techniques is that the doors are now open for nonexpert users to access business analytics. What skills are required? The average user just needs to be able to express a business question in terms of data mining functions. So then, what are these data mining functions?

### Classification

Classification serves to generate scores or probabilities at a fine-grain entity level that can be associated with an object of interest, such as a customer. It can be used to predict, for example, the probability of said customer buying a specific product.

The doors are now open to nonexpert users to access business analytics.

**Regression**
Regression is used to predict a value that can be associated with an object of interest, such as the amount a customer will buy over the next three months. Regression allows you to generate value estimates at a fine-grain entity level.

**Clustering**
Clustering is the grouping together of objects of interest, such as customers, that are similar. Clustering allows you to generate either prototypes or segments.

**Time-Series Forecasting**
Time-series forecasting is the high-level extraction of information in past data values to predict future values, such as forecasting the number of units sold per month for the next 12 months.

**Association Rules**
Association rules are used to detect associations between fine-grain events to detect correlations or simple interactions between data elements. For example, data on past purchases can be used to establish rules for proposing products to customers.

**Attribute Importance**
Attribute importance means identifying the key influencers that explain whether a customer will buy a product or how much they will buy. It can generally be a side effect of good techniques to perform classification, regression, clustering, or even time-series forecasting.

**SAP PREDICTIVE ANALYTICS**
The good news is that SAP Predictive Analytics has a solution for each of these data mining functions. The software is fully automated, highly scalable, nonparametric, and fully understandable by a business analyst and business user.

All processing stages in the analytics framework in SAP Predictive Analytics use this type of technique. In fact, designers for SAP Predictive Analytics have used these concepts on real-world problems since 1992. SAP Predictive Analytics focuses on implementations that are quick and interpretable at the same time. The design team made use of our experience to determine what information to provide and how to provide it so results are interpretable by line-of-business users as well as statisticians.

SAP Predictive Analytics focuses on implementations that are quick and interpretable at the same time.

## COMBINING ALGORITHMS WITH BEST PRACTICES

The mathematics that are employed in the framework of SAP Predictive Analytics can be used to build several models that are made to compete. It does not simply do this at random by changing the modeling technique, but it uses Vapnik's structured risk minimization (SRM) to scan through different model sets. SAP Predictive Analytics compares models before presenting the one with the best compromise between fit and robustness.

SRM provides a mathematical framework that allows the writing of algorithms that have the same objectives as best practices developed over the years by experts. This principle opened the door for an automated process that was not possible before.

SRM is in the public domain and can be implemented by any vendor. So what makes SAP Predictive Analytics different? SAP Predictive Analytics was built with several goals in mind. We believe that machine learning services must be scalable and interpretable and provide results with a quality that can be compared to models that are handcrafted by specialists using first-generation tools.

Algorithms fulfilling these constraints were not available. So we looked to develop a service with the scalability of logistic regression, the quality of a neural network, and the interpretability of a decision tree. The services needed to support any data situation, with no underlying assumptions about the statistical distribution of either the input or the target.

These requirements were not only for classification but also for regression, clustering and segmentation, attribute importance, and time-series forecasting. This is why we developed our own technology.

With SAP Predictive Analytics, interpretability is an essential requirement to spreading advanced analytics across the enterprise. It is designed to present to end users meaningful results that can be displayed in a report or chart, such as the notion of key drivers, category importance, quality indicators, and robustness indicators.


SRM opened the door for an automated process that was not possible before. SAP Predictive Analytics uses SRM to bring this process to your company.

# Benefits of SRM

There is a significant difference between the algorithmic automation framework used in SAP Predictive Analytics and what is called automation by other vendors. Specialists of the domain often associate automation with engineering techniques that fire a lot of different algorithms on the same data set, and compare results to select the best one on the given data set. There are several important limitations to this approach:

- It performs poorly on a wide range of analytics data sets, such as classification cases with uneven target distribution or rare case events.
- It fails in very high dimensional spaces or provides very unstable results in such situations. Most classical techniques tend to overfit when the number of inputs is increasing, leading to less predictive power when used on new or unseen data.
- It is parametric in nature, meaning that several parameters need to be tuned for each algorithm, either requiring a skilled person (which is against the automation principle) or requiring a test for an even greater number of challengers. This solution is not practical when dealing with several tens of thousands of models in production, as some customers of SAP customers are currently managing.
- It automatically updates and selects models based on the results, so that the January version of a model could be a decision tree and the February version of the model could be a neural network. This prohibits the use analytics on analytics, which customers of SAP are already using.

Vladimir Vapnik's SRM theory offers a mathematical foundation explaining why using several modeling functions can be useful; but because it is done in a proper theoretical framework, it can be controlled and used to overcome the previous four limitations.

With SAP Predictive Analytics, interpretability is an essential requirement to spreading advanced analytics across the enterprise. It is designed to present meaningful results to end users.

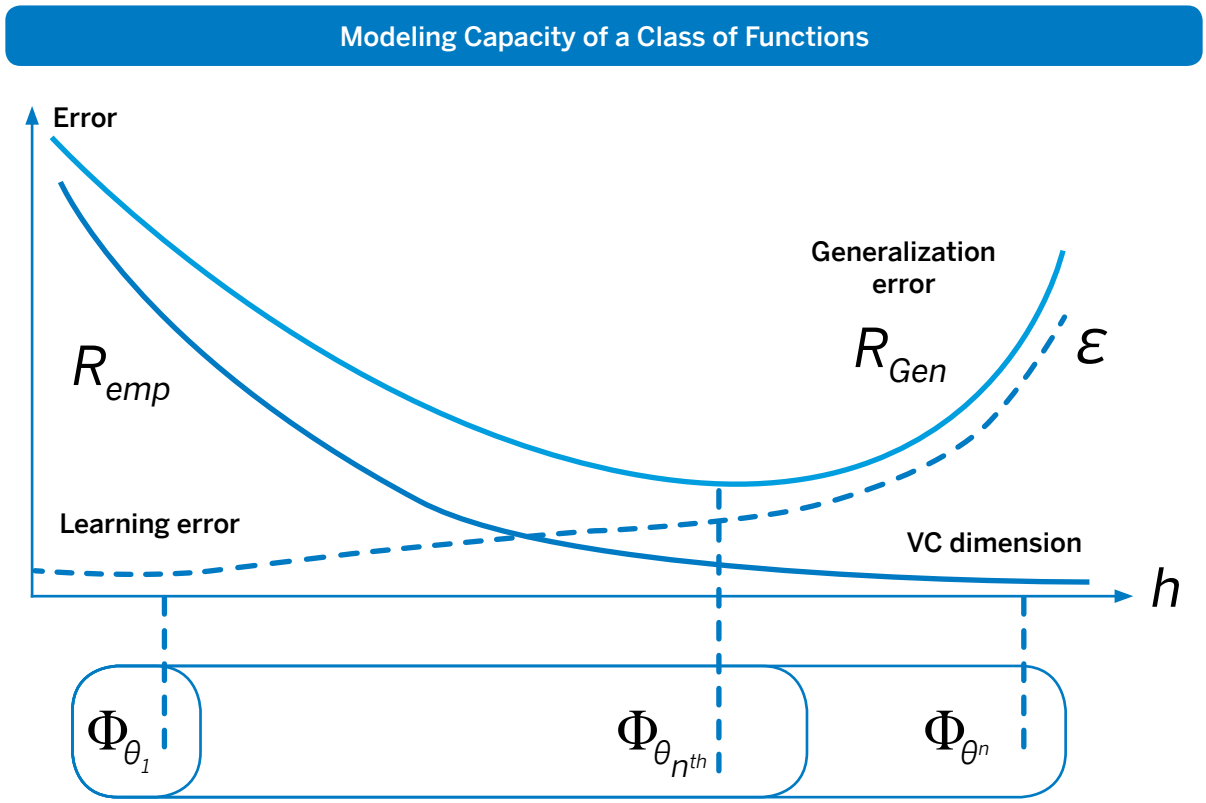
## BUILT-IN ROBUSTNESS

SRM provides a way to select the model based on a decomposition of the error – one part is attributed to the error on the training data set and the other one is attributed to the confidence interval on the extra error that will be made on any new data set. Building an algorithm following this method will provide models to represent the best compromise between fit and robustness. In other words, models are selected not based on their errors on the training data sets but on their expected errors on any new data sets, which is much closer to an operational view of the modeling activity.

Legacy tools do not integrate such built-in mechanisms and require modelers to do this validation themselves, using validation and test samples, for example. The only algorithm found in legacy tools that has built-in robustness features is called support vector machines (SVM) invented by Vapnik and I. Guyon when they were both working for AT&T Labs. This algorithm does not offer scalability with a number of examples, cannot provide results interpretable by business users, and is sensitive to noise near the decision border, requiring careful tuning or a lot of CPU power.

The figure represents the Vapnik–Chervonenkis dimension (VC dimension), which measures the modeling capacity of a class of functions. The models in a class of functions with a high VC dimension will have the "capacity" to represent very complex relations between inputs and outcomes. The horizontal axis of the graph represents the VC dimension, so the left part of the graph points to simple models and the right part of the graph represents more complex models. The schema shows the evolution of the error made by models with respect to the evolution of their capacity or complexity.

Figure: The VC Dimension Modeled Against Empirical Risk and Generalization Risk

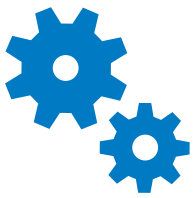



## CORRECTING THE ERROR

SRM decomposes this error in two parts. The first part comes from the error made on a training data set, also called empirical risk, which is the error the system is making on the data that is available to train the model. The second part is the extra error (or its upper bound) the system will make on any new data set, also called generalization risk. For the sake of simplicity, the training data will be called the past data (usually models are trained on data coming from the past or data that was collected through an experiment) and the data on which to apply the forecasts will be called the future data.

SRM finds an upper bound of the generalization risk, which makes almost no hypothesis on the statistical distribution of the data. What does a business user want to minimize? Would it be the error made on the past data or would it be the estimation of the error made on future data? Everyone with at least one practical experience in data mining is aware of a well-known problem called overfitting. Vapnik explains this phenomenon very well – it occurs when using a class of functions with a too-large capacity. In this case, the error on the past data will be very small, but the value derived by Vapnik's theory for the generalization error will be very large.

From this theoretical notion, Vapnik derived a way to build robust algorithms that would look at different classes of functions with different capacity. This is done in order to find the optimal class of functions in which a model will minimize the expected error made on new data – this is the foundation of SRM. Algorithms obtained this way are different from all previous data-mining algorithms. Algorithms used in SAP Predictive Analytics, thanks to SRM, are minimizing the expected error on new data. This is where the robustness of our techniques comes from; algorithms used in SAP Predictive Analytics do not overfit as much as the first-generation tool algorithms.



Modeling techniques in SAP Predictive Analytics can be used in very high dimensions. SAP Predictive Analytics has been used on customer tables for risk projects with more than 15,000 attributes.

# Modeling in High Dimensions

Another side effect of using SRM is that the algorithms may be built in a way that is independent of the number of attributes provided to the modeling technique. Most first-generation algorithms tend to base their implementation on classes of functions that have a capacity linked to the number of dimensions – the higher dimension, the higher the capacity. This explains why most first-generation tools do not handle high-dimensional spaces well – they tend to overfit very easily. Algorithms derived from SRM may counter this fact by controlling their capacity in a way that is loosely linked with the number of input dimensions. This is the foundation for the explanation of why algorithms in SAP Predictive Analytics tend to perform well in very high dimensional spaces.

SAP Predictive Analytics is not focused on finding perfect algorithms but instead on finding algorithms and techniques that can automatically handle any kind of data and return excellent results.

# Quality and Robustness Metrics

Classically, predictive models can be characterized in terms of quality, which is the easy part if taken as the performance on the training data set. They can also be characterized on robustness, which is related to the expected extra error that will be made on new data sets.

Scientific surveys and extensive bibliographies can be found about quality metrics, discussing pros and cons. Knowing that, it can come as a surprise that scientific papers evaluating new algorithms are still using metrics known to be very bad in general cases such as "classification rates" for classification tasks or "mean square error" for regression tasks.

A model quality that could be forwarded to non-statisticians should be compliant with some basic requirements, independent of the modeling task:

- The model quality should be expressed as a value between 0 and 1, or as a percent.
- The model quality of a perfect system should be 100%.
- The model quality of a random (or constant) system should be 0%.
- The model quality should be computed without any assumption about the underlying algorithm or the target distribution.
- The model quality should relate to an intuitive notion of quality in low-dimensional space.

It is difficult to cope with the two last points. In the last 10 years, there has been a great normalization effort on the quality metrics used for binary classification. Data miners agreed not to use the classification rate anymore (the ratio of correctly classified examples) but instead use a metric known as Somers' D. In SAP Predictive Analytics, this metric is called the information indicator (KI). This metric is directly linked with some others, such as the area under the receiver operating characteristic (ROC) curve and Kolmogorov-Smirnov test.

## METRICS FOR REGRESSION

The same level of normalization is not achieved for metrics on regression tasks. One of the main problems in finding a high-quality metric for regression is linked to the need for "no assumption" about the target distribution. A lot of scientists are still using mean square errors and derivatives; this family of metrics is known to induce a lot of problems because it is rooted in a Gaussian framework that has shown a lot of limitations in practice.

An effective way to approach this problem is to take metrics that are based on the relative order of the estimates with respect to the order of actual values. In particular, such metrics are invariant with all monotonic transformations of the target; forecasting a continuous value or the logarithm of this value has no impact on the metric that is based on relative order, for example. But in practice this would not be enough, because in some cases a decision cannot be optimized only based on the relative order of estimates but also must be based on the value of estimates. A single estimate of a very large value may be the only one of interest, such as a single expected profit of US$1 million. With SAP Predictive Analytics, the proposal for continuous targets is to extend the notion of KI, which is based on the notion of order, and to propose final estimates obtained through monotonic recalibration. This is based on the validation data sets, thus providing estimate values in the correct range.

There is no consensus on the notion of robustness metrics, as such metrics are not even provided in most cases by first-generation tools. SAP Predictive Analytics has the only framework to systematically provide such value as robustness metrics (KR). SAP Predictive Analytics computes this value by looking at the variability of the quality metrics on parts of the training data set. This value can be considered as a good "proxy" of the intuitive notion of robustness and will certainly be improved as time and research go by.

# Exploratory Analytics

One of the side effects of being able to perform modeling in high dimensions is that it changes the way analysts work. With first-generation tools, there are no viable techniques that can be used to directly perform modeling in high dimensions. Specialists are required to perform a priori variable selection, which, in turn, requires performing data exploration through descriptive statistics (in particular, cross statistics with the target variables). Sometimes, the tool may provide an approximate version of variable selection based on the same cross statistics.

When using SAP Predictive Analytics, everything is reversed, because there is no need to perform a priori variable selection. Modeling techniques in SAP Predictive Analytics can be used without worry in very high dimensions; SAP Predictive Analytics has been used on customer tables for risk projects with more than 15,000 attributes. There is no need to perform a priori variable selection, which, in turn, removes most of the need for data exploration. This may dramatically reduce the time needed to perform the data preparation phase, but this does not obviate the need for descriptive statistics. Instead of trying to show descriptive statistics on all variables, it can just show them on the variables that are the most important for the current project. In SAP Predictive Analytics, these are called key drivers, which is part of a bigger domain called exploratory analytics.

# Key Drivers

When using SAP Predictive Analytics, the major predictive and descriptive functions (classification, regression, segmentation) are able to provide the user with the notion of key drivers. Key drivers are the most important attributes explaining the outcome; models from SAP Predictive Analytics provide the sorted list of the most important attributes to predict or describe the outcomes.

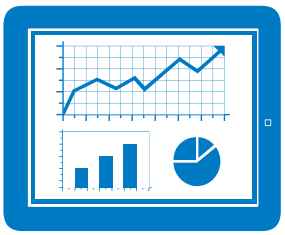
Algorithms in SAP Predictive Analytics do not need to cope with correlated variables, which could cause numerical problems for other algorithms, as is the case with algorithms such as linear or logistic regression. SAP Predictive Analytics handles them in order to present the results to the user, because when two variables are highly correlated, a robust algorithm will tend to equalize the weight on both variables. This causes the user to miss that there is a concept, which is encoded using two variables, that has a weight that can be the sum of the weights of both variables. If the user wants to perform variable selection, it is important to keep the variables that are most representative of each concept. While this might seem to be just a matter of presentation, it is also very important.

The other nice thing about the way that SAP Predictive Analytics computes attribute importance is that it will not lose information when the difference between two variables has more weight than each of the original variables. This can be pictured with two date variables – a duration, which is the difference between two dates – and can automatically be found by the system to be more important than each of the dates taken separately.

# The Consensus for Data Mining Workbench Users

In the data mining literature, there are many proponents of using a modeling toolkit or workbench and combining as many well-established methods (algorithms) as possible. The core message is that there is no single algorithm that will optimally solve problems, so you must have them all. We agree with this statement, but it might be too strong. We prefer another version – automated analytics is a fully automated modeling process that will get excellent results almost every time. With SAP Predictive Analytics, automated analytics design is not looking for perfect algorithms but instead for algorithms and techniques that will automatically handle any kind of data and return excellent results. This is a very different quest, and with SAP Predictive Analytics, we have achieved something remarkable in this direction. To do this, we designed our own algorithms because they did not exist at the time, and still do not in any other software product.

SAP Predictive Analytics provides expert analytics and automated analytics. With expert analytics, data scientists have the ability to pick and choose algorithms.



# The Pros and Cons of Multiple Algorithms

Proponents of the have-it-all approach do not usually explain why the algorithms present in their toolboxes cannot do it all. It turns out that one modeling technique does not make the same errors as another, either because it has been optimized for a specific situation or because of its capacity to represent one class of situations and not another. So with luck, or if data happens to have the right characteristics, good results will be obtained. If not, time will be wasted. Is that the end of the story?

The answer is yes for data scientists that can afford and master all these algorithms, and for the companies that can afford bigger machines with more CPU and memory. They just say, "I have tried them all and picked the one I prefer." Picking the best model is not always an easy task, but we do have a solution – expert analytics in SAP Predictive Analytics. This solution can be used to implement "classical" automation techniques associated with the four problems that we covered earlier.

Automated analytics algorithms in SAP Predictive Analytics, based on SRM, formalize this search. Instead of randomly searching and trying all the algorithms, they give a direction for the search and compare the methods.

Another way to look at this situation is that, instead of focusing on the competition between algorithms on a single data set, automated analytics in SAP Predictive Analytics focuses on the competition between the data elements. This means that the algorithms proprietary to SAP Predictive Analytics are scalable in very high dimensions, which is not the case for almost all of the classical algorithms found in the workbenches of legacy tools.

# What is the Remaining Value of Classical Algorithms?

After all this demonstration, the consequence is clear – the remaining value of classical algorithms is minimal. That means that it is perfectly okay to use these classical algorithms, but they should be provided for free. There are plenty of open source environments providing the well-known algorithms. Microsoft Excel can handle large data sets and provides simple and almost free implementation of what was common knowledge in data mining in 2008.

SAP Predictive Analytics provides expert analytics and automated analytics. With expert analytics, data scientists have the ability to pick and choose algorithms.

### AUTOMATION IS MORE THAN ALGORITHMS

Automated modeling environments for operational deployment cannot be used without an automated environment to create analytical data for certain uses. These uses include retraining predictive models, applying these models, testing for deviations, and even automating back testing for models used in operations.

Very often model management tools only focus on the fact that predictive algorithms or predictive models will be used on the freshest version of the data. This is not enough for managing automation for the full lifecycle of data mining projects. Such massive automation of modeling environments should propose an easy way to create the analytical data set that should be used to train a predictive model. An example of this would be to put the system into the situation it was in three months ago and test it over the last three months before deciding to put this model or this modeling technique into operation. This is why there is a data manager integrated into SAP Predictive Analytics, creating a way to manage analytical data sets through time.

# Automated Data Encoding

One of the key steps in a predictive analytics project is making data compatible with the modeling algorithms the analyst wants to employ. Some algorithms only accept symbols, others only accept numbers. Specialists with hands-on experience spend much of their time on data manipulation and data encoding.

In practice, data encoding means a lot of tricky operations designed by data mining experts, such as processing missing values, processing out-of-range values, and encoding the data with respect to the algorithm applied. It also means having robust algorithms that deliver consistently good results.

SAP Predictive Analytics helps solve this problem by using our academic experience as well as our real-world business experience for how to put predictive and descriptive analytics into operation. Our design goal was to build integrated, automated ways of processing missing values and out-of-range values.

The objective of automated analytics in SAP Predictive Analytics is to make sure that once the user has chosen the problem description, the maximum amount of interpretable information for the business problem is extracted.

**LEARN MORE**
Discover more about predictive analytics solutions from SAP.

**ABOUT THE AUTHOR**
Erik Marcade is the vice president of advanced analytics products for SAP SE, heading the advanced analytics development team within the product and innovation division. Erik came to SAP through the acquisition of KXEN, where he was the founder and chief technical officer. With over 30 years of experience in the machine-learning industry, Erik was responsible for software development and information technologies at KXEN. KXEN sold advanced analytics automation tools mainly used in the customer relationship management space for verticals such as telecommunications, bank and finance, retail, and Internet-based businesses.

Erik holds an engineering degree from Ecole de l'Aeronautique et de l'Espace, specializing in process control, signal processing, computer science, and artificial intelligence.

The Best-Run Businesses Run SAP®