Knowledge Integrity, Inc.

Business Intelligence Solutions

**Democratizing Data Mining – A Radical Approach to Pervasive Predictive Analytics**

White Paper

Written by:

**David Loshon**, Knowledge Integrity, Inc.

Sponsored by Microsoft

Date of Publication: 12.12.08

Version: #1



Contents

[Introduction 1](#_Toc217106416)

[Moving Predictive Analytics out of the Ivory Tower 1](#_Toc217106417)

[Predictive Analytics, Business Value, and Data Mining 2](#_Toc217106418)

[Clustering 2](#_Toc217106419)

[Classification 3](#_Toc217106420)

[Market Basket Analysis 4](#_Toc217106421)

[Using Predictive Techniques: Profiling, Estimation, Forecasting 5](#_Toc217106422)

[Applying Data Mining Techniques 5](#_Toc217106423)

[Clustering 5](#_Toc217106424)

[Case- or Memory-Based Reasoning 5](#_Toc217106425)

[Association Rules 6](#_Toc217106426)

[Decision Trees 6](#_Toc217106427)

[Neural Network 6](#_Toc217106428)

[Technology Expectations 6](#_Toc217106429)

[Summary 7](#_Toc217106430)

# Introduction

The area of incorporating predictive analytics into operational processes has, for the most part, relied on “industrial strength” data mining and knowledge discovery products employed by expert analysts with deep backgrounds in statistics and probability. These analysts not only understand the approaches employed by the analysis tools, but are also experts in designing, building, and refining analytical models. As subject matter experts, these professional data miners have been able to add significant value to business operations, but only as long as they are fully integrated in the loop. While this top-down approach yields valuable insight, it is limited in scalability due to the costs of the industrial strength tools as well as the bandwidth of the experts available to perform the analysis. These constraints have prevented widespread organizational scalability, limiting the use of advanced analytics to mostly strategic scenarios.

But consider this: a person doesn’t need to understand how a radio works in order to listen to music, nor does a person need to know how to calculate torque in order to use a power drill. In fact, in most situations people are able to use tools even without knowing how to build them. So why do we expect a similar level of expertise with advanced analysis? In actuality, it is no longer the case that expertise is required in order to make use of data mining, nor is advanced analysis confined to strategic use.

Anyone should be able to exploit the benefits of data mining in improving the results everyday activities, whether that involves strategic decision-making or guiding immediate operational choices. For example, a marketing manager can employ advanced analytics to help analyze lead generation and subsequent pipeline management, sales representatives can employ analyses to better qualify leads to prioritize prospects for contact, and sales managers can evaluate whether successful sales patterns can be used to help generate more qualified leads.

But this all involves putting the power of analytics in the hands of different individuals across the organization, not just the senior managers. And it implies a radically different approach to developing analytic models: instead of using a top-down approach with *a priori* knowledge of statistics and probability to develop the “right model,” employ a bottom-up approach trying many different analysis techniques to see which models provide the most valuable answers.

In this paper we will discuss the business value of predictive analytics and look at the way that data mining techniques are employed in developing analyses. We then provide enough of a working overview of data mining techniques to get the developer started. This is followed by a deeper dive into the bottom-up approach to show that one does not need a deep understanding of predictive analytic algorithms to be able to take advantage of their power. In the bottom-up approach, the developer uses available data mining techniques in an iterative manner to see which ones work best to meet business needs. The paper then reviews techniques that use knowledge discovery algorithms to achieve desired predictive capabilities, as well as looking at some expectations of the environment, technology, and people to make this feasible. Finally, we summarize the approach to delivering the value of predictive analytics to a broader audience.

# Moving Predictive Analytics out of the Ivory Tower

Although common use of the term “data mining” has diluted some of the phrase’s original meaning, there has been a maturing body of tools intended to apply statistics, probability, and computational algorithms to seek out patterns in data sets that upon evaluation add insight into ways to improve and optimize business processes, especially when providing predictive power. Also referred to as “knowledge discovery,” algorithms such as clustering, association rule analysis, and neural networks al apply statistical techniques to find those critical nuggets of knowledge that can be generalized into predictive models. A predictive model uses knowledge and patterns inherent in existing data to be used to predict future actions or behavior associated with new data.

The common methods employed by expert analysts to develop predictive models essentially compose a top-down approach. The analyst evaluates the business problem, assesses the data, and then considers the options for using high-priced software tools in developing the “right model” to meet the business needs. Yet achieving the benefits and business value of predictive analytics should not be restricted to those organizations able to invest significant budget allocations to licensing and maintaining high-end analysis tools as well as the expert staff to run them.

There is, however, an alternate path that does not require knowledge of statistics or expertise in data mining to be able to develop predictive models. Instead of developing the right model from the top-down, even non-experts can address the prediction and forecasting challenges using a bottom-up approach that organically grows good models from among a library of many.

The bottom-up method allows a broader constituency to take advantage of data mining techniques. With an understanding of what the techniques are intended to achieve, almost anyone can use and refine many models and then iteratively see how well each model works and which ones provide the most valuable insights.

This approach brings predictive analytics out of the ivory tower and into the hands of people who can immediately benefit. By broadening the community that can take advantage of knowledge discovery utilities, an organization is no longer constrained to apply its limited resources to a small number of strategic problems. Application developers, mid-level managers, and business analysts can demonstrate that testing out many models and seeing which ones work best will eventually lead to convergence on the right model. This approach does not require users to have deep knowledge in statistics or data mining, but rather let them develop many models and then iteratively see which ones are best at meeting business user needs.

# Predictive Analytics, Business Value, and Data Mining

Business analysts have applied data mining techniques to enhance and optimize business processes in different ways. For example, data sets can be analyzed with no preexisting expectations as a way to find interesting patterns that warrant further investigation. This *undirected analysis* relies on no preconditions, and can be a good way to start developing predictive models. Alternatively, once a model has been developed, it can be used in a *directed* manner, using the patterns that have been revealed to predict future events or help in achieving a particular goal. Subjecting a set of customer records to clustering is an undirected process that will group customers based on discovered similarities and differences. Evaluating the dependent variables resulting from clustering customer data enables directed classification of new customer records into the discovered clusters. There are a number of undirected and directed data mining techniques that can be combined to develop predictive models, and here we look at a few:

## Clustering

Clustering is a process that collects records into groups such that each group is clearly different from all others, and that the members of each group are recognizably similar. The records in the set are organized based on similarity; since there are no predefined specifications for classification, the algorithms essentially “select” the attributes (or “variables”) used to determine similarity. Someone with business context knowledge might be called upon to interpret the results to determine if there is any specific meaning associated with the clustering, and sometimes this may result in culling out variables that do not carry meaning or may not have any relevance, in which case the clustering can be repeated in the absence of the culled variables.



Figure 1: Example of clustering

The results of clustering might be used for another data mining technique, *classification*. Clustering can help in highlighting those attributes that characterize similarity of segments within a population that can be used for subsequent evaluation, categorization, and segmentation.

## Classification

Classification is the process of organizing data into predefined classes. Those classes may be described using attributes selected by the analyst, or may actually be based on the results of a clustering model. During a classification process, the class definitions and a training data set of previously classified objects is presented to the application, which then attempts to build a model that can be used to accurately classify new records. For example, a classification model can be used to evaluate public companies into good, medium, and poor investments, assign meta-tags to news articles based on their content, or assign customers into defined market segments.



Figure 2: Classification of records based on defined characteristics.

## Market Basket Analysis

Market basket analysis is a process that looks for relationships of objects that “go together” within the business context, and its name is derived from the concept of analyzing the contents of a supermarket shopper’s cart to see if there are any “naturally occurring” affinities that can be exploited in some other manner. Some examples of the use of market basket analysis include:

* Product placement – identifying products that have may often be purchased together and arranging the placement of those items (such as in a catalog or on a website) close by to encourage the purchaser to buy both items.
* Physical shelf arrangement – an alternate use for physical product placement in a store is to separate items that are often purchased at the same time to encourage individuals to wander through the store to find what they are looking for to potentially increase the probability of additional impulse purchases.
* Up-sell, cross-sell, and bundling opportunities – Companies may use the affinity grouping of multiple products as an indication that customers may be predisposed to buying the grouped products at the same time. This enables the presentation of items for cross-selling, or may suggest that customers may be willing to buy more items when certain products are bundled together.
* Customer retention – When customers contact a business to sever a relationship, a company representative may use market basket analysis to determine the right incentives to offer in order to retain the customer’s business.

## Using Predictive Techniques: Profiling, Estimation, Forecasting

Each of these concepts implies the use of a model for prediction. Once a model has been developed, it can be used to answer specific questions within a business process; in fact, some models combine descriptive techniques with predictive techniques to suggest ways to improve operational processes. Some examples include:

* Churn Analysis, in which models of customer turnover and attrition are used to help in predicting situations where customer relationships are at risk, evaluating customer lifetime values, and then providing suggested approaches that have high probability of retaining high value customers;
* Fraud Detection, where anomalous events are detected and highlighted for review to determine if fraudulent activity is taking place;
* Customer Response Prediction, where market segmentation, marketing campaign effectiveness, and customer lifetime value are compared to rate the expectation that a specific customer will respond to a particular campaign.

# Applying Data Mining Techniques

There is a difference between knowing *how* a data mining technique works and knowing *what* it is supposed to do. Noting this distinction and then training organizational staff on the “what” allows an organization to broaden its staff’s capabilities in developing predictive models that be applied in everyday activities. This section describes some commonly used data mining techniques to provide the necessary groundwork and details about data mining so that almost anyone to take on the challenge of predictive analysis.

## Clustering

Clustering algorithms are usually unsupervised methods intended to segment instances in a data set into groups in which the members of a group are similar and the groups are distinct from each other. There must be some way to determine “similarity,” and this introduces a little more complexity. Each data instance represents an entity with specific characteristics, which are reflected as the values in each data attribute. The measure of similarity of two data instances corresponds to how close the corresponding attribute values are to each other. For some types of attributes, the measure of similarity is straightforward. For example, for continuous values (such as “person age” or “person height”), similarity is based on proximity of the values, and for categorical values (such as color or car make/model), similarity is based on whether the values are the same or not. The clustering algorithms can be provided with similarity measures, but in some cases they attempt to infer that as well.

There are two common approaches to the clustering process. One starts out by placing all the data instances into one bucket and then attempt to break them out into some number of groups. Another starts by placing each data instance into its own group, then attempt to merge groups together until there is only one cluster left. This second approach documents the clusters after each round of analysis, and lets the analyst decide which level best addresses the analytic needs.

## Case- or Memory-Based Reasoning

A sick visit to a doctor typically involves the patient’s describing a collection of symptoms and the doctor’s review of documented history to match the symptoms to known illnesses, come up with a diagnosis, and then recommend ways to treat the patient. This is a real-world example of case-based reasoning (CBR, sometimes also called memory-based reasoning or instance-based reasoning), in which known situations are employed to form a model for analysis. New situations are compared against the model to find the closest matches, which can then be reviewed to inform decisions about classification or for prediction.

A CBR algorithm will build a model using a training set of data and specific outcomes. The model is then used for comparison with new data instances, resulting in an entity’s classification or with suggestions for actions to take based on the new data instance’s value. Again, the model must rely on some mechanism for computing similarity to match instances within the model as well as matching new instances against those within the model.

## Association Rules

An association rule describes a relationship between sets of values occurring with enough frequency to signify an interesting pattern. A rule usually takes the form of “If {X} then {Y},” where X is a set of conditions regarding a set of variables upon which the values of set Y depend. An example is the canonical data mining story regarding the frequency of the purchase of diapers and beer together. The co-occurrence of the variable values must have *support*, which means that those values occur together with a reasonable frequency, and the apparent co-dependent variable(s) must show a degree of *confidence* indicating that of the times that the {X} values appear, so do the values for {Y}. Market basket analysis depends on the discovery of association rules

## Decision Trees

Many situations can be addressed by answering a series of questions that increasingly narrow down the possibilities of a solution, much like the game “twenty questions.” A decision tree is a decision-support model that encapsulates the questions and the possible answers and guides the analyst towards the appropriate result, and can be used for operational processes as well as classification.

Decision tree analysis looks at a collection of data instances and given outcomes, evaluates the frequency and distribution of values across the set of variables, and constructs a decision model in the form of a tree. The nodes at each level of this tree each represent a question, and each possible answer to the question is represented as a branch that points to another node at the next level. The analyst uses the model to seek a desired result as the decision support process traverses the tree and stops when the traversal reaches the leaf of the tree. A nice aspect of decision tree models is that the “thought process” used by the model is transparent, and it is clear to the analyst how the model reached a particular conclusion.

## Neural Network

A neural network is a data mining model that is used for prediction. The neural network model is trained using data instances and desired outcomes, and the algorithms for building neural networks encapsulate statistical artifacts of the training data to create a “black box” process that takes some number of inputs and produces some predictive output. Originally envisioned as a way of modeling human thought, neural network models are based on statistics and probability, and once trained are very good for prediction problems. However, the knowledge embedded in the training set becomes integrated into the neural network in a way that is not transparent – the neural network model is very good at prediction, but can’t tell you why it came up with a particular answer.

# Technology Expectations

When data mining first emerged as a viable technology, the computational demands of the algorithms coupled with the necessary requirements for expertise kept the barrier to entry high for most organizations. The most significant challenges lie in reducing the barrier to entry in two different ways: reducing the computational strain and eliminating the need for data mining algorithm expertise. The first challenge has essentially gone away, as the availability of computing power and the necessary storage systems have become increasingly affordable to almost every kind of business.

The second challenge is a technical one in incorporating the right kind of “business smarts” into the predictive analysis utilities to make the processes pervasive in a way that supports the bottom-up approach. Achieving this goal leads to three expectations that business and data analyst would have of technology suppliers:

* **Availability of the utilities** – Availability of the tool set is critical, and it involves providing predictive analysis services in a way that is easily accessed across the analyst spectrum (senior managers, mid-level managers, business analysts, superusers, as well as developers).
* **Education and training** – Availability is of limited value if the target audience is not aware of the tools or how those tools can be employed to build predictive models. Any technology innovations must be accompanied by the corresponding educational collateral to operationalize the value proposition of bottom-up analytics.
* **Ease of use** – Yet another barrier to entry has been the difficulty in developing models and applying them in the proper context. Pervasive predictive analysis is dependent on reducing the complexity of analyzing data sets, building models, and employing those predictive models in tactical and operational contexts.
* **Embedding and integration –** The best way to meet all of these expectations is through full integration of the technology directly into the productivity tools used on a daily basis. Many analysts interact with databases that can be subjected to analysis in order to build predictive models. Embedding the capability to build the model into the database fabric effectively eliminates the middleman, enabling the analyst to invoke the right services at his or her own convenience to build predictive models for daily use.

This last point defines the technical opportunity: embedding data mining capabilities and packaging their use in business analytic packages within the existing infrastructure. This includes incorporation of data mining services directly integrated at various levels in applications, either at the data level (e.g., integrated within the query environment), at the management level, at the developer level (embedded APIs within development platforms), or at the desktop level (e.g., as part of desktop tools such as spreadsheets, database tools, project planning tools, presentation development tools, etc.). This allows for data preparation, model training, evaluation of discovered knowledge, and integration of pattern-based business rules or predictions back into operational applications while staying within an unified information architecture.

# Summary

The emergence of embedded predictive analytics may be a significant boon to anyone interested in exploiting the benefits of data mining to improve the results of everyday activities. Anyone across the organization willing to invest some time to learn the basics of predictive analysis may be able to improve business processes, whether those are strategic decision-making processes or decision support to supplement ongoing operational work flows. Improved marketing, sales, customer service, reduction in attrition, fraud, and improved satisfaction for customers and staff members can result from fine-tuning business processes through the filter of advanced analytics.

Putting the power of analytics in the hands of different individuals across the organization is the key driver to success in this endeavor. Instead of relying on a select few individuals in the organization entrusted with this predictive power, everyone in the organization can exploit these tools by employing a radically different approach to developing analytic models. Instead of using a top-down approach with *a priori* knowledge of statistics and probability to develop the “right model,” let anyone in the organization use embedded data mining capabilities using a bottom-up approach to use many different analysis techniques to see how different kinds of models ultimately provide the most value to the enterprise.