A Management Guide to Data Mining
1 INTRODUCTION

The central notion behind data mining is this: we look for patterns of behavior in historical data with the aim of exploiting these patterns in the future. For example we might find patterns that identify a certain group of customers (by age, income and education say) as being more likely to respond to the promotion of a certain product. These patterns are found by analyzing historical data and a decision has to be made about the likelihood of this relationship still being true. In fact this is one of the most common applications of data mining technology and there are many technology suppliers providing solutions in this domain.

The term ‘data mining’ embraces a large number of technologies and techniques, but for our purposes we will separate out statistical methods. These are fundamentally different from many of the methods used in data mining. We are all familiar with simple statistical methods – the mean, standard deviation and regression where we fit a line to a set of data. Statistics are predefined metrics of a data set and either they make sense for the data or they do not, and it is often very difficult to establish which is the case. Data mining on the other hand often determines the metrics as a model is built. We’ll get to some of the methods used later, but this is an important point and worth remembering.

If data mining was as simple as applying the various techniques to data and exploiting the patterns that are found there would be little need to go any further. Alas this is not the case. There are two major issues that data mining throws up. The first is the fact that these methods often discover patterns that are nothing more than accidental fits with the data. It might just happen that during the time period analyzed customers with a certain profile purchased a particular product, but that the pattern has no validity outside this time period. Various names are given to this phenomenon – over-fitting being one of the most common. The algorithms will find patterns that are nothing more than a fit with data that are essentially random. There are ways to minimize this behavior, but the danger is always present.

The second major challenge presented by data mining is the problem of knowing whether the patterns that have been found are persistent. If we can establish that the patterns really do represent real behavior then is it reasonable to assume that the same behavior is still manifesting? This is where human judgment is needed, and there is no substitute for it. In some cases an extrapolation of the behavior into the future will be perfectly logical, in other cases transient fashions and biases may make us much more suspicious.

In reality data mining is as much an art as a science, even to the extent that the way data are presented to the data mining algorithms will make a huge difference in the patterns that are found. Consider an attribute such as age. If we feed the age into an algorithm simply as a number of years, many algorithms will have problems dealing with a continuous variable of this nature. It might be much more productive to categorize age as twenties/thirties/forties and so on. There really is no substitute for experience although the technology is becoming more user friendly with built in knowledge of common attributes used in a business environment.
2 COSTS, BENEFITS AND RISKS

As with all information technologies data mining offers an opportunity to increase the efficiency and effectiveness of an organisation. The core idea behind data mining is that through the use of appropriate technologies we can identify patterns of behaviour, in customers, employees, suppliers, machinery and in fact any aspect of the organisation provided data has been captured. These patterns would then allow us to improve understanding of processes and in some instances predict the outcome of a situation. Obviously this has great appeal and the suppliers of data mining technology are not shy when it comes to advertising the potential benefits. But as anyone who has been involved with information technology in a large organisation for more than two weeks will tell you – all that glitters is not gold.

The costs associated with data mining are as small or large as you want them to be. At a minimum you might download a free data mining tool (RapidMiner for example) and with the necessary skills start mining data. Patterns will be found – this is guaranteed. Whether they are useful or accurate is another matter completely, and this is where most of the risk creeps in. The same issues materialise if you spend hundreds of thousands or millions of dollars on your data mining technology. You may be able to process larger data sets with greater speed, and present very pretty visualisations – but the problems of usefulness and accuracy still remain.

If your organisation can afford to employ a room of PhD statisticians, data mining experts, mathematicians and domain experts then the risks are going to reduce. Although it should be remembered that this is precisely the sort of team employed by the large banks prior to the 2008 credit crisis – largely created by complex, but inappropriate modelling of derivatives. Even so, such a team of people should easily spot the ghosts in your data that have no reality in your business. So provided you spend wisely the risks can be reduced by hiring experts. But these people are not cheap. Their salaries are often in excess of a hundred thousand dollars a year, and they are quite difficult to find.

The opposite end of the extreme is to use a plug and play data mining tools and blindly accept its findings. Please do not do this – things will almost certainly go badly wrong. The patterns that are found may be no more than apparitions in your data with no existence in reality, and to act on them may be costly. There are ways to minimise these apparitions, but they are quite technical, and even then, not foolproof.

The key to using data mining technologies successfully is people, and particularly people who understand the domain where the technologies are being used. There is a famous story reported in the Wall Street Journal of attempts to use data mining in financial markets. It was found that US stock returns could be predicted with 99% accuracy if US cheese production and the total population of sheep in the US and Bangladesh were used as inputs. Clearly this is nonsense, but this is what data mining came up with. A domain expert (someone who knows about financial markets) knows that this is nonsense.

A strongly supervised data mining initiative (supervised by domain experts) has many benefits, but ultimately the benefits have to exceed the costs and be worth the risks. One of the most common uses of data mining is in sales and marketing. Market basket analysis is widely used to establish buying habits of customers, typically in a retail scenario. Prospecting, estimating response rates, fine tuning messages
and so on are all fair game for data mining. And just as data mining does present real risks, it also presents the opportunity to significantly improve the fortunes of an organisation.

Ultimately data mining is all about uncovering information, and someone in the organisation needs to be ensuring that the costs of unearthing this information are smaller than the benefits it delivers.

## 3 DATA MINING METHODS

A number of data mining techniques are explained which are frequently used in many types of data mining activity.

### Supervised Learning Techniques

The techniques shown below are used in a supervised learning scenario. This is where a data set is provided for the tools to learn from, so that new data can be classified or a value predicted through regression.

### Bayesian Classifiers

Bayesian classifiers use a probabilistic approach to classifying data. Unlike many data mining algorithms Bayesian classifiers often work well with a large number of input attributes, without hitting what is called the dimensionality problem. Naive Bayes is the technique most often employed – the term 'naive' coming from the fact that input attributes should be independent of each other (ie there are no correlations between them). Despite the fact that this is often not true, naive Bayes still gives good results. Unfortunately it is often overlooked for more esoteric methods, whereas it should actually be a first port-of-call if relevant to the problem and where most attributes are categorical (ie categorised).

Bayes works by combining what are called conditional probabilities into an overall probability of an event being true. Explaining Bayes is difficult (as evidenced by the large number of explanatory videos on youtube). I have made a video which can be seen here, which will hopefully explain things further for those interested.

### Decision Trees

Decision trees are a favorite tool to use in data mining simply because they are so easy to understand. A decision tree is literally a tree of decisions and it conveniently creates rules which are easy to understand and code. We start with all the data in our training data set and apply a decision. If the data contains demographics then the first decision may be to segment the data based on age. In practice the decision may contain several categories for segmentation - young/middle age/old. Having done this we might then create the next level of the tree by segmenting on salary - and so on. In the context of data mining we normally want the tree to categorise a target variable for us - whether someone is a good candidate for a loan for example.

The clever bit is how we order the decisions, or more accurately the order in which we apply attributes to create the tree. Should we use age first and then salary - or would the converse produce a better tree? To this end decision trees in data mining uses a number of algorithms to create the best tree. The most popular algorithms are Gini (which uses probability calculations to determine tree quality) and information gain (which uses entropy calculations).
When large data sets are used there is the very real possibility that the leaf nodes (the very last nodes where the target variable is categorised) become sparsely populated with just a few entries in each leaf. This is not useful because the generalisation is poor. It is also the case that the predictive capability drops off when the leaves contain only a few records. To this end most data mining tools support pruning, where we can specify a minimum number of records to be included in a leaf. There is no magical formula that will say what the level of pruning should be, it's just a matter of trial and error to see what gives the best predictive capability.

Virtually all data mining tools implement decision trees and some offer elaborations on the basic concept - regression trees for example where the tree is used to predict a value, rather than categorise.

Decision trees are often used to get a feel for data even if they are not part of the resulting model, although good results are to be got from decision trees in many business applications.

**Nearest Neighbors (k-NN)**

Entities can often be classified by the neighborhood they live in. Simply ask whether your own neighborhood gives a fair representation of you, in terms of income, education, aspiration, values and so on. It doesn’t always work – but usually it does – birds of a feather and all that. A similar mechanisms has been developed to classify data – by establishing which neighborhood a particular record lives in. The official name for this algorithm is k-Nearest Neighbor, or k-NN for short.

The essential idea is this. Imagine you are interested in buying a second hand car. Mileage, fuel efficiency, service history and a number of other attributes will typically be of interest. Someone you know has a database of used cars which includes these details and each car is categorised as a peach or a lemon. By entering the details of the car you are interested in the k-NN algorithm will find the 5 (so k=5 in this instance) cars with the closest match to yours. If more are peaches then lemons then you might have a good car – and that’s it.

Obviously it gets a bit more involved with large commercial data sets – but the idea is simple enough. It works best where most of the attributes are numbers that measure some sort of magnitude, so that the algorithm can establish where the nearest neighbors are. Attributes that represent classifications can be a problem and so k-NN may not be suitable. Even so this simple algorithm is widely used and can deliver good results.

**Neural Networks**

If decision trees represent transparency and good behaviour then neural networks epitomise opaqueness and temperamental behaviour. But what else would you expect from a sometimes brilliant and other times obstinate technology? Neural networks are used for prediction and classification, and through the development of self-organising maps (SOM), for clustering. They are called neural networks because they supposedly mimic the behaviour of neurons in the nervous system, taking inputs from the environment, processing them and creating an output. And just in the same way that neurons are linked together, so are nodes in a neural network. As with other data mining techniques neural networks demand that a good selection of relevant inputs are available, that the target output is well understood and that copious amounts of data are available for training.
The most commonly used type of neural network is called a feed forward network. As the name suggests it works by feeding the outputs from each node forward to the next node as its inputs. The flow is essentially one direction, although something called back propagation is used to tune the network by comparing the network’s estimate of a value against the actual value. Nodes in a network do two things. They combine the inputs by multiplying each input by a weight (to simulate its importance) and summing the products - this is called the combination function. Other functions are used, but this is the most common. Secondly, the output from the combination function (a single number) is fed into a transfer function which usually takes the form of a sigmoid (an S shaped curve) or a hyperbolic tangent. These curves allow the network to deal with non-linear behaviour. In essence they create a linear relationship for small values, but flatten out for large values. This form of non-linearity is an assumption - but it often works well. The output from the transfer function is then fed to the next node in the network.

Most neural networks have three layers - the input layer, a hidden layer, and the output layer. The hidden layer is so named because it is invisible, with no direct contact to inputs or outputs. Knowing how large to make the hidden layer is one of the crucial issues in using a neural network. Make it too large and the network will simply memorise the training set with absolutely no predictive capability at all. Make it too small and useful patterns will be missed.

Using a neural network requires a considerable amount of skill and the results can range from the sublime to the ridiculous simply by modifying any one of a number of parameters. The most important parameters include:

- The size of the training set.
- The number of hidden layers and the number of nodes in each hidden layer.
- Parameters affecting how quickly the network learns.
- The features to use as input.
- The combination functions and transfer functions.

This is by no means exhaustive and issues such as normalising inputs, converting categorical inputs and so on, all have a profound effect on the quality of the network produced. Some of the plug and play analytics tools omit neural networks altogether, and for good reason. Other methods produce equally good results without the temperamental behaviour. Having said this, neural networks can detect patterns that evade detection by other means, and they are very good at picking up some non-linear behaviours.

**Support Vector Machines**

Support Vector Machines (SVMs) are one of the most powerful classes of predictive analytics technologies. They work by separating out data into regions (by hyperplanes in multi-dimensional spaces for those interested), and as such classify the data. Oracle for example has a predictive analytics Excel add-on that uses SVMs exclusively. Having said this they are not relevant tool for all analytics problems and can over-fit the data in the same way as neural networks – although there are mechanisms for minimizing this effect.

SVMs are an essential component in any analytics toolkit and virtually all suppliers include an implementation.
Unsupervised Learning Techniques

These techniques are used to find relationships within data without being offered a data set to learn from. As such there is no special nominated attribute in a data set that is to be categorized or calculated (or scored in the lingo of predictive analytics). Despite this these techniques do allow new data to be allocated to a cluster or associated with a rule. The two dominant techniques here are called clustering and association.

Clustering

Clustering is very similar to the k-NN technique mentioned above but without specifying a particular attribute that is to be classified. Data are simply presented to the clustering algorithm, which then creates clusters using any one of a number of techniques. This is as much an exploratory technique as a predictive one. A typical example might be clustering patients with similar symptoms.

Association Rule Mining

Unlike the supervised learning methods association rule mining is unsupervised and is concerned with the discovery of any rules which might exist between attributes. This sounds fairly straightforward, but is riddled with potholes – the most common being the discovery of hundreds (thousands) of rules that are either trivial or spurious. However used well this technique does unearth previously unknown relationships and forms the backbone of basket analysis – a common application used in retail.

Predictive Analytics

Predictive analytics is a particular application of data mining technologies. The typical mechanism used to predict in predictive analytics is scoring. We might wish to score the credit worthiness of a new customer, or score the likelihood of machine failure in a manufacturing plant. A large number of algorithms are available to find the patterns from historical data, which are then used to score new data. The names given to these algorithms are all suitably off-putting, but in essence most of them rest on fairly simple ideas – I’ll be exploring them in other articles – there is no reason why a manager should not know when regression is used, or when it shouldn’t be used.

The overwhelming use of predictive analytics is in sales and marketing – trying to assess receptive candidates in a marketing campaign or who should be targeted for up-sell/cross-sell activity. Other uses include fraud detection, credit rating and increasingly health-care analysis. The application of predictive technologies is as broad as human activity, so these initial applications are just the early uses of the technology.

The suppliers of predictive technologies fall into three main camps:

Enterprise solutions suppliers who bolt on a predictive analytics capability. These are typically quite weak offerings, although IBM is a notable exception.

Proprietary analytics solutions aimed at large organisations. KXEN, Angoss, SAS and others come into this category.

Open Source offerings are often the most capable, but least user friendly. Revolution Analytics and Rapid-i have taken open source solutions and made them enterprise ready – these are some of the best tools available to experienced analysts.
4  Excel Data Mining Tools

Many data mining tasks can be accomplished within Excel, given a suitable add-in. The main benefit is that this is a familiar environment and is ideally suited to trying things out. The five data mining add-ins listed here differ considerably in their sophistication and user friendliness. 11Ants Model Builder hides as much of the back room activity as possible and will automatically select the most appropriate mining methods. Alyuda ForecasterXL however offers self tuning neural networks as a method of mining data. DataMinerXL is a tool for people familiar with data mining techniques and Predixion Enterprise Insight is the only solution that many organisations might need. Finally XLMiner provides a full data mining environment for people with the relevant knowledge.

All these tools can be used for predictive analytics where discovered patterns are used to score new data as it comes in.

**11Ants Model Builder** from 11Ants Analytics

This is a user friendly Microsoft Excel add-on that can be used with a minimum of training and will quickly identify predictive patterns in data. Most of the action is behind the scenes and the software will automatically home in on the most productive data mining methods. In larger organisations these models can be deployed in enterprise databases using 11Ants Predictor. This supports very high throughput scoring on Oracle, Microsoft SQL Server and Teradata databases.

11Ants Model Builder supports decision tree, Gaussian processes, logistic regression, Naïve Bayes, nearest neighbour, random forest and support vector machine – amongst others.

Marketing solutions are offered for customer churn and customer response predictive analysis. Again the primary model development is accomplished in a Microsoft Excel environment, and models can then be deployed to enterprise databases using 11Ants Predictor.

**Alyuda ForecasterXL**

This Excel add-in implements neural networks within Excel. It boasts ease-of-use with automatic neural network parameter and architecture selection. Various graphical and analytical displays are provided and the partition of data into training and test sets is straightforward.

Moderately priced at US$197 for a single user (US$997 for unlimited site), it is a low cost method of exploring the use of neural networks within an Excel setting.

**DataMinerXL**

If you have some familiarity with data mining techniques and want a low risk route then DataMinerXL is a good option. This is an Excel add-in which supports the creation of predictive models using a wide variety of techniques, including regression (linear and logistic), naive Bayes, decision trees, neural networks, support vector machines (SVM) and will even solve linear, quadratic and linear complementarity problems. Other functions are also included for those with a math bent (numerical integration, and matrix manipulation). Basic statistical functions are also included.
Clearly this is not an end-user tool. But for someone familiar with the territory it is an excellent way to build predictive models, and for all the budding information scientists out there a free version (throttled to 1000 instances) is available. The paid licence is very reasonable too at US$ 499 per year.

An excellent book has been published by the creators of DataMinerXL – Foundations of Predictive Analytics by James Wu and Stephen Coggeshall.

**Predixion Enterprise Insight**

The Excel front end is the client side to a broader data mining capability. The server side supports most data and database products including big data sources such as Hadoop and Greenplum. Collaborative capability is one of the main features of the product with full integration into the Microsoft stack. A level of end-user capability is claimed and models can be shared using SharePoint dashboards.

Predixion Enterprise Insight Developer Edition can be downloaded for free (you won’t even have to enter your details), and can be used to get a feel for the technology prior to commitment.

**XLMiner from FrontlineSolvers**

This add-in for Excel provides a full-blown data mining capability with data preparation tools, support for times series analysis and visualisation tools. The techniques used by the add-in include regression (logistic and linear), Bayes classifier, association rules, neural nets, classification and regression trees, clustering, principal components and discriminant analysis.

The data sources supported include Microsoft’s PowerPivot, Microsoft/IBM/Oracle databases and of course simple spreadsheets.
5 5 FREE DATA MINING TOOLS

The five free data mining tools listed below are equally as capable as many products that have high price tags attached to them. They are in no way inferior, and most are Open Source with a large community of knowledgeable developers.

Knime

Knime is a widely used open source data mining, visualisation and reporting graphical workbench used by over 3000 organisations. Knime desktop is the entry open source version of Knime (other paid for versions are for organisations that need support and additional features). It is based on the well regarded and widely used Eclipse IDE platform, making it as much a development platform (for bespoke extensions) as a data mining platform.

It incorporates hundreds of different nodes for data I/O, preprocessing, cleansing, modelling, analysis and data mining. WEKA analysis modules are also incorporated and an additional plugin allows R scripts to be run.

Knime runs on Windows, Mac OS X and Linux.
**Orange**

This is a very capable open source visualisation and analysis tool with an easy to use interface. Most analysis can be achieved through its visual programming interface (drag and drop of widgets) and most visual tools are supported including scatterplots, bar charts, trees, dendograms and heatmaps.

A large number (over 100) of widgets are supported. These cover data transformation, classification, regression, association, visualisation and unsupervised learning methods. There are also some specialised add-ons covering bioinformatics, text mining and other specialist requirements. The environment is extendible through Python scripting and this includes creating new widgets if needed.

The documentation is good too and includes first steps, detailed widget descriptions and scripting. It runs on Windows, Mac OS X and Linux.
R

Strictly speaking R is a programming language, but there are literally thousands of libraries that can be incorporated into the R environment making it a powerful data mining environment. In reality R is probably the most flexible and powerful data mining environment available, but it does require high levels of skill.

From a career perspective learning R is a good investment. Many enterprise tools support R (SAP Predictive Analysis, Tibco Spotfire for example) and it addresses much more than data mining. RevolutionAnalytics has based its products on R and have added a graphical front-end. They also offer a free version of R that is calimed to be faster than the general distribution.
Rapid Miner

This is perhaps the most widely used open source data mining platform (with over 3 million downloads). It incorporates analytical ETL (Extract, Transform and Load), data mining and predictive reporting. The graphical user interface and visualisation tools are excellent, with considerable intelligence built into the workflow construction process. This provides on-the-fly error recognition and suggested quick fixes. Its meta data transformation capability is unique among tools of this nature allowing results to be inspected at design time.

It incorporates over 500 operators and includes the WEKA machine learning library. Many extensions are available for analysis of time series and text and other specialised processes. Most data sources are supported including Excel, Access, Oracle, IBM DB2, Microsoft SQL Server, Sybase, Ingres, My SQL, text files and others.

Rapid-i provides support and training services for organisations that want a supported product.
WEKA

This set of data mining tools is incorporated into many other products (Knime and Rapid Miner for example), but it also a stand-alone platform for many data mining tasks including preprocessing, clustering, regression, classification and visualisation. The support for data sources is extended through Java Database Connectivity, but the default format for data is the flat file.

WEKA comes from the highly respected machine learning group at the University of Waikato, New Zealand (same origin as the 11AntsAnalytics Excel data mining tool).

Models can be built using a graphical user interface or a command line input.
6 ENTERPRISE DATA MINING TOOLS

Tibco

Tibco has got a real winner with Spotfire. It provides an easy to use interface for data visualisation, analytics and the creation of dashboards. Most of the slicing and dicing can be done through drag and drop and a multitude of intelligent functions (eg scaling the time axis on charts automatically) make light work of many analysis tasks. The lightening fast execution speeds are also a great advantage, particularly on large data sets.

More complex analytics can be accomplished through the R programming language, and the R runtime engine has been embedded into the Spotfire statistical server. This allows R based analysis to be fed out to as many users as required (typically through its WebPlayer web client).

Version 5.0 of Spotfire has embraced big data and particularly in-database analysis, with support for Oracle, SQL Server and Teradata - others will follow.

A developer licence costs around US$4000 and the server licence per user is US$500 - with volume discounts.

www.tibco.com

Statsoft

Statsoft is best known as the supplier of Statistica. This comprises a large set of statistics and data mining tools with over thirty separate products within the Statistica portfolio. The company sees itself as a somewhat less costly, but equally capable alternative to SAS and isn't shy about telling the world when a customer moves over from the SAS camp.

Statistica offers a very broad capability. There would be little point providing a detailed analysis, since it does pretty well everything an analyst might require. There are some vertical solutions for pharmaceuticals, credit scoring and quality control. For analysts familiar with R, Statistica offers a high level of integration with its own tools.

The Enterprise versions of Statistica support access to Big Data sources (eg Hadoop) with associated techniques such as multi-threading.

www.statsoft.com

Oracle

Oracle’s positioning of predictive analytics is unusual. While it supports an arsenal of data mining techniques and technologies with no pretense of ‘plug and play’ functionality, predictive analytics has been encapsulated into a spreadsheet add-on with the Oracle’s signature ‘it’s so easy’ marketing. This is despite the fact that support vector machines (SVM) form the backbone of their predictive analytics
offering (one of the most powerful analytical techniques and worthy of some respect). It's all very
strange, and potentially misleading.

In fairness to Oracle the data mining capability (and predictive analytics is just a particular usage of data
mining technologies) is on a par with many other offerings, and as such Oracle has a competitive
offering. It is to be hoped that not too many people see predictive analytics as simply a plug and play
spreadsheets add-on utilising SVMs - because it isn't.

**SAP**

Organisations wedded to the SAP way of doing things will probably choose SAP Predictive Analytics. The
consolation is that SAP has very sensibly integrated R into its analytics offering. The front-end is a
Windows client that makes the modeling process more user friendly and is integrated into SAP's Visual
Intelligence offering.  

This is designed to be an extension of the overall SAP architecture incorporating in-memory processing
and connection to a variety of data sources (including Hadoop for big data applications).

Various algorithms are used - many of which are based on R, and some others developed by SAP. The
usual candidates are present - decision trees, neural networks, regression, clustering - and so on.
Resulting predictive models can be exported as PMML (Predictive Model Markup Language) for
deployment in a production environment.

There is nothing particularly interesting here, but for SAP users looking to extend into the predictive
analytics space this is an obvious candidate.

**Salford Systems**

Salford Systems delivers a portfolio of products capable of traditional descriptive analytics (what has
happened or is happening) and predictive analytics. What distinguishes this company is the lack of hype
around the technology it offers and a willingness to discuss the pitfalls and traps associated with
predictive analytics - which ironically is a prerequisite for successful analytics. The SPM Salford
Predictive Modeler supports both traditional descriptive and predictive analytics. CART supports
classification and the discovery of hidden relationships between attributes. It embodies a number of
proprietary methods and patented extensions to the original work done in the eighties.

TreeNet is something of a black box and generates a plethora a reports for analysts to decipher the
results of analysis. MARS (Multivariate Adaptive Regression Splines) produces regression models and is
seen as a complement to CART. Finally Random Forests is best used on small(ish) data sets which might
have many attributes, and includes prediction clusters and segment discoveries.

It should be clear that this is not an end-user tool, although the company does claim some end-user
functionality in CART. What is offered here is innovative and different and will quite possibly reveal new
insights and models for many organisations. [www.salford-systems.com](http://www.salford-systems.com)

Copyright Butler Analytics 2013
**Revolution Analytics**

Real analysts use R - well it sounds a bit macho, but actually there is some truth in it. R is the most widely used, and arguably the most powerful analysis software on the planet. Revolution Analytics has taken this Open Source wild child and turned it into something the enterprise can use with relevant support, training and enhanced productivity.

Revolution R Enterprise is built on open source R and has been enhanced for performance, productivity (through visual tools), and integration with enterprise data sources - and particularly Apache Hadoop for big data applications. Support and training services are bundled on top of the technology - something most organisations will require.

A community edition of Revolution R is available for free. It doesn't come with the visual tools or database interoperability, but it is faster than the Open Source version.

Many large corporations use Revolution R Enterprise - it's a bedrock of their analytical activity. Unlike more proprietary offerings you are unlikely to reach any dead ends using R - but you have to know what you are doing.

**FICO**

FICO provides a broad range of technologies and services to support business optimisation and the embedding of intelligence into applications of various kinds. Under the hood there is a lot going on - from linear (and non-linear) programming through to predictive analytics and other extremely powerful methods of supporting business decisions through applying intelligence to a variety of applications. It is clear that the experience and know-how of FICO in the industries it serves is probably unique. Financial services, retail, government and healthcare are its main markets, but the technologies and methods they employ have broad applicability.

Fraud detection and customer credit worthiness are two of the primary themes in the application of its technology, but the portfolio is so broad that most business problems will be addressable. Perhaps most interesting is the capability to combine different model types into a cohesive whole. Business rules (which are deterministic) can be combined with predictive analytics (which is usually probabilistic) to create very accurate models for dealing with customers and detecting anomalies.

[www.fico.com](http://www.fico.com)

**SAS**

In a sense there is almost nothing to say about SAS and the analytics space - it does everything. What is probably unique to SAS is the speed at which new techniques are adopted. Singular spectrum analysis is a glaring omission in many analytics packages, but it was incorporated into SAS with lightning speed (SSA is useful for trend and cycle analysis).
To complement this very broad range of capabilities SAS provides a number of vertical solutions. These address financial services (particularly fraud and financial crime), customer analytics, governance and compliance, supply chain intelligence, and several others.

It has to be said that SAS focuses on the analytical tools part of the equation and other technologies will be needed for 'big data' and deployment into production systems. There has also been some controversy over the high cost of SAS technology and World Programming provides a SAS type environment at a lower cost. This of course was challenged by SAS, but the EU Court of Justice ruled in favour World Programming.

That said, SAS is the foundation stone of analytics in many of the world’s large corporations - and it will probably stay that way.

IBM

For IBM predictive analytics is largely a data management and infrastructure issue. In my conversations with them they stress the data management aspect particularly, and with good reason. The application of algorithms to data and the building of models, which is primarily accomplished with SPSS, is really just a small part of the story. The management of large data volumes and the deployment of models into the production environment is the more challenging aspect of analytics, and it is something IBM does very well.

The IBM analytics solution will primarily be of interest to large organisations looking for more than a point solution, and wanting to create a viable, long term analytics infrastructure and capability. To this end IBM offers its InfoSphere data management and infrastructure products, and the SPSS suite of analytical tools for both analysts and end users. The combination represents the premier analytical solution currently available, and of course IBM has a number of vertical solutions to offer also. It is of course a fairly expensive solution, but in many ways is unchallenged.

SPSS

Data Collection Family

This suite of products from IBM is primarily aimed at the design, creation, deployment, analysis and reporting of surveys. They provide a top-to-tail capability that supports various means of survey distribution (web, paper, phone, in-person) and the supporting technology to capture the results, including scanning of documents and text processing.

The SamplePower utility provides a means of establishing survey sample size - something that would normally require a skilled statistician. This sets the tone for the whole Data Collection product set, since virtually all elements of the process can be handled by users. This does not however include the analytics used to draw conclusions from the data, and is the domain of the statistics and Modeler packages.
IBM SPSS Statistics

This perhaps the most widely used set of statistical products in the world. The capability ranges from end user marketing tools through to specialised statistical analysis, and of course the very well respected SPSS analyst workbench. There isn't much utility in detailing the features of the statistics capability because it does pretty well everything. A few things are also available that are not really statistical in nature such as neural networks.

IBM SPSS Modeler

This employs data mining techniques to find relationships within data. The professional version supports the creation of predictive models using classification, association and segmentation techniques. Modeler Premium adds the ability to process unstructured data from the web, text, email, social data and so on. Again there is little point listing all the techniques supported by Modeler since most conceivable options are present (Bayes, SVM, K-means etc).

Deployment Family

IBM SPSS Decision Management allows predictive models to be integrated with business rules for deployment into production systems. The Collaboration and Deployment option supports the sharing of analytical assets and provides an environment to automate the analytical process.

InfoSphere

InfoSphere addresses more than predictive analytics requirements and is fully addressed in a separate paper. However the broad capability of the product suite includes InfoSphere Warehouse for traditional data warehousing, InfoSphere Information Server, DataStage and Data Replication to support integration and data staging, Master Data Management and Big Data analytics, which is based on the Apache Hadoop technology.

Big Data analytics not only supports large data sets, but provides sufficient performance for real-time analytics and accommodation of very high volume streaming data. This will become more important as information sources from various sensors (eg RFID) and real-time market information becomes more widely used.

KXEN

KXEN is one of the leaders in the world of predictive analytics, and with good reason. Although a recent marketing makeover seems to have deprived prospects of learning what is under the hood, we can tell you that there are some heavy duty algorithms working to make sure that predictive models are valid (Structural Risk Minimisation techniques are used). This is a heavy duty product suitable for large organisations in the main, although recent cloud based offerings make it accessible to smaller businesses.

There are six elements to the product range:

Explorer streamlines the data preparation process which is usually the lengthiest part of any model building exercise.

Copyright Butler Analytics 2013
Modeller provides a high level modelling capability that avoids a great deal of experimenting.

Scorer allows predictive models to be deployed within a database for production use.

Factory is a mechanism to production-line analytic tasks.

Social Network Analysis is a specific application of the KXEN technology.

Genius provides marketers with a user friendly interface to conduct their own analytics.

More recently the company has started to offer cloud based analytics targeted at the sales and marketing functions and specifically targets Salesforce users.

KXEN is particularly experienced in the communications, financial services and retail industries with customers such as Barclays, Vodafone and Sears.

Shame about the patronising marketing makeover, but this is good technology and will address the needs of many large businesses.

www.kxen.com