

Predictive Analytics in Business

Strategy, methods, technology

Contents

Page	Topic
3	Management Summary
4	The Business Case for Data Mining
6	Enterprise Predictive Analytics Platforms
13	Excel Data Mining Add-Ins
15	Open Source and Free Data Mining Tools
20	4 Machine Learning Startups
23	A Brief Introduction To Analytics Techniques

Please note: This is an independently produced, unsponsored report. Please do not distribute or reproduce this document without permission (one hard copy may be produced per downloaded document).

Butler Analytics focuses exclusively on the use of analytics technology in business. We can be contacted at info@butleranalytics.com.

Management Summary

Used with care and skill predictive analytics will bring significant benefits to your business. The central idea underpinning predictive analytics is that we process historical data with the aim of finding patterns of behaviour (in customers, suppliers, machinery, employees etc.) that can be exploited in future activities.

The term predictive analytics simply refers to a particular use of data mining technologies where data are scored. A new loan applicant may be scored based on the patterns that have emerged while processing historical data. The score may simply put the new prospects into the category of 'unsuitable' or 'suitable' – or the categories may be more finely grained.

At the present time predictive analytics, and its sibling big data, are terribly over-hyped. This is a classic tech industry bubble with massive over-expectation, accompanied by chronic over-confidence. There are dangers associated with this technology which management should be aware of.

The most serious of these is to place too much confidence in the technology and not enough in people. Domain experts are a crucial ingredient in the effective use of predictive analytics technologies. Many of the data mining techniques will identify patterns that are either trivial, unusable or simply unrelated to reality. Domain experts are needed to sanitize the results of data mining. Some suppliers of these technologies are fond of making statements such as 'trust the math' – please do not trust the math, it is just a sausage machine and prone to obey the law of rubbish-in, rubbish-out.

The application of predictive analytics will be as broad as the activities of the organization. Domains such as marketing and sales can be addressed through a variety of packed analytics solutions, and there is little need to reinvent the wheel. Obviously predictive capabilities will become more prevalent in mainstream business applications, but a watchful eye is always needed. Bespoke applications are likely to present most opportunity for a decisive advantage, by definition. A wide variety of tools are available, ranging from Excel spreadsheet add-ins to complex, large scale analytical tools that support big data.

Obviously the safest route into predictive analytics is through a modest beginning. There are issues relevant to organization, culture, politics, skills and confidence that must be addressed. It is best to meet these issues through prototypes, rather than all-at-once during a big-bang approach.

Whether predictive analytics and big data technologies are adopted depends on perception of need. If there is no pressing need it is always best to wait for the dust to settle, for the cost of technology and skills to fall, and for lessons to have been learned (at other peoples' expense). Where there is a pressing need then caution is the keyword, and it is much wiser to trust people than technology. We don't want the predictive models suggesting that we turn customers away, when in reality they are perfectly good candidates.

This is something of a brave new world, and reports are already appearing (probably sponsored by the tech industry) suggesting that analytics is increasingly sorting out the winners from the losers. Used with skill, predictive analytics undoubtedly has the ability to transform the fortunes of an organisation. Used recklessly and you might just end up with the dumbest smart applications in your industry.

The Business Case for Data Mining

The overwhelming use of IT in most organizations centres around the filing and retrieval of data. While the filing may be electronic, it is nonetheless simply a more efficient means of acquiring, sorting and retrieving data. This is simply a cost associated with running a business, there is very little value in it. Data mining is a determined attempt to extract value from the terabytes of data most organizations acquire on an ongoing basis. Without such technology these data are simply wasted assets and the organization misses opportunities to gain insights into customers, operational activities, suppliers and even employees. And it should be added that these insights are often not trivial in nature.

What is Data Mining?

Data mining is the act of looking for patterns of behaviour that are reliable enough to be used in current operational activity, to improve both efficiency and efficacy. The process typically involves analysis of historical data, the discovery of patterns, validation of patterns, sanitisation by domain experts, and if accepted, the implementation of newly discovered patterns in business processes. Perhaps one of the best known examples is that of deciding whether a candidate is a good risk for a loan. By mining through historical data it can usually be established which customers have proved to be bad risks, and which proved to be good risks. Attributes such as salary, age and so on will often be used by the data mining algorithms to determine the behaviour of customers. Once the patterns have been discovered they can be implemented in the systems that support loan approval. The use of data mining in marketing is also very common, and patterns are often discovered which indicate the best candidates to be approached in a marketing campaign, or the best customers for up-sell or cross-sell opportunities. The younger brother of data mining is predictive analytics. This is simply a particular use of data mining technologies to predict behaviour based on scoring. The loan approval process detailed above is one such application. Predictive analytics has assumed a high profile, probably because the name is more appealing than 'data mining' – but it uses exactly the same technologies.

More than business intelligence, data mining and predictive analytics provide mechanisms to extract value from otherwise wasted information. While none of these terms has a rigorous definition, most of us identify manual data exploration and reporting activities with BI, and automated processes with data mining and predictive analytics. As data sources become greater in number, and each provides larger data volumes and more diversity in the type of data, so the need for automated analysis increases. Of course data mining does not replace BI activities, it is simply an essential activity that complements it.

Another IT Fad?

The IT industry has no problem with hype and inflated expectations. We've seen it with customer relationship management (CRM), Service Oriented Architectures (SOA) and any number of other three letter acronyms. However data mining has been around for well over a decade and some of the methods it uses originate with the renaissance – so it is here to stay. It may be polished up and given go-faster stripes (as is really the case with predictive analytics) now and then, but the underlying technology and philosophy is not going to go away. The departure from how we have traditionally used IT needs to be emphasized. Until recently IT has been almost exclusively targeted at cost reduction through labour displacement. This has involved the use of computers as glorified filing cabinets and calculators. Data mining is something quite different. With data mining we are actually looking for the technology to add value – to make our operations more effective. This may mean targeting customers more accurately, predicting machine failure on a production line,

analysing which suppliers are most responsive – and so on. In a nutshell we are talking about the introduction of intelligence into the systems we use. In many ways it emphasizes the distinction between efficiency and efficacy. While efficiency is necessary for business success it is not sufficient, the extra ingredient needed is efficacy. We may be able to produce green mugs with pink handles more efficiently than anyone else, but does anyone want them? Where is the fashion in mugs moving? What associated products could be sold (mats for example)? Data mining provides the means to answer these questions (through targeted market research).

Costs and Risks

Since data mining is a relatively new notion for many organizations it is a good idea to phase implementation so risks are reduced. The best place to start is small, and embrace larger projects as confidence grows. Of course this goes against many human biases, but experience shows over and over again that an incremental approach is best.

Starting small may mean prototyping with an Excel add-in, or utilizing one of several open source data mining tools. Organizations with an existing commitment to a particular data mining tools supplier should also start small – with a number of pilot projects. Not only should the initial projects be brief and well defined, but they should also address areas of business functionality where mistakes can easily be spotted and where impact is measurable but minimal. Data mining technologies are complex and validating the patterns they find is a subtle process. If an organization employs experienced professionals then the risk of curve fitting, over-fitting, data mining bias and so on should be minimized. These things happen when the patterns that are found are no more than accidental fits, and when we mine data until it is forced to show us what we want to see. Even with the various statistical checks and cross checking methods that data mining employs it is best to involve domain experts. They will have a good feel for whether a pattern represents something real or not. Initial costs will be modest and will relate more to skills than technology. However as confidence builds there may be a decision to embrace big data, and for this considerable infrastructure will be required. Eventually a full team of analysts, statisticians, programmers and domain experts may be needed, although clearly someone has to be measuring delivered benefits against ongoing costs.

Benefits

Assuming we know how to deal with the risks associated with data mining the benefits are potentially as broad as the operations of the organization. At the time of writing it has been revealed that HP used data mining to identify employees who might leave. So it has application in every part of the business and the benefits will range from the trivial (predicting the demand for stationery in various departments) through to the profound (discovering a simple adjustment of marketing message that produces a significant boost in sales). Analytics activities will become the primary information processing activity in most organizations, demanding the most resources and delivering the most benefits. Traditional transaction based systems exist for administrative purposes and as fodder for analytics.

Summary

The business case for data mining is very simple really. Every aspect of the systems used in organizations should be complemented with intelligence. Much of this will eventually happen by default – suppliers of solutions will embed data mining type processes into their products. But as always it will be the bespoke analytics that deliver the most benefit and for this skills and resources will be needed. In a nutshell this is a move from the efficient enterprise to the intelligent enterprise – the latter encompassing the former.

Enterprise Predictive Analytics Platforms

The ten predictive analytics offerings listed below vary enormously in functionality and applicability. Where a product is exceptional it is given a five star rating – although this obviously does not mean it is the best solution for your organization. Tibco has been added because it embraces R for predictive analytics, but is also capable of BI and visualization – an interesting mix.

FICO ★★★★★

FICO provides a complete analytics environment, not only for model development and deployment, but also for monitoring and management. This latter capability is quite unique and is key to management confidence and the adoption of new technologies. The analytic capability comes under the broad heading of decision management, since ultimately all analytics technologies have as their single aim better decisions under uncertainty.

The technology is not limited to data mining or predictive analytics, but embraces many other regimes which support decision making by managers and in the production environment. This includes business optimization, business rules and business simulation – and many others too numerous to list here.

It is worth remembering that FICO has been developing and deploying these solutions in large organizations for over three decades and has a mature understanding of what works, what doesn't, and where the risks lay. Hence the emphasis on the ability of managers to monitor the performance of models and applications. It is clear that the experience and know-how of FICO in the industries it serves is probably unique.

More recently FICO has decided to launch its FICO Analytic Cloud. This is a full suite of analytics capability, and big data storage resources, that are offered as a cloud based service to organizations wishing to avoid the cost, disruption and complexities associated with an in-house facility. Most importantly it gives customers access to solutions, frameworks and design patterns which encapsulate decades of experience. It also allows organizations to create their own apps or even create a service. This will undoubtedly be of great interest to many large organizations wanting to avoid the teething troubles associated with the adoption of these advanced technologies.

Many of the algorithms and capabilities offered by FICO are specific to the types of business problems addressed. Access to other algorithms is made available through R integration, and text analytics is implemented through integration with Lucene.

FICO's traditional customer base is found in banking, retail, government, health-care and insurance, but the technology and services have broad applicability and FICO also has customers in virtually every other industry.

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:

1. Explorer streamlines the data preparation process which is usually the lengthiest part of any model building exercise.
2. Modeller provides a high level modelling capability that avoids a great deal of experimenting.
3. Scorer allows predictive models to be deployed within a database for production use.
4. Factory is a mechanism to production-line analytic tasks.
5. Social Network Analysis is a specific application of the KXEN technology.
6. 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.

Oracle Advanced Analytics (Overall ★★★★★) (Existing Oracle Users ★★★★★)

Not surprisingly Oracle provides a full repertoire of technologies to handle data mining and statistical analysis, with or without big data. This is mostly put under the Oracle Advanced Analytics umbrella, encompassing predictive analytics, text mining, statistical analysis, data mining, mathematical computation and visualisation. Key to the approach taken by Oracle is the notion of in-database processing – they were originally a database company after all. This means that processing happens in the database environment and that data extraction is unnecessary. However it is unlikely that any organisation would want to run its data mining activities in the same environment as the transactional systems, and so there is an implication at least that data is extracted to a data warehousing, or other environment.

Oracle provides two routes to analysis:

1. Oracle Data Mining revolves around SQL and the actual modelling environment, Oracle Data Miner, comes as an extension to Oracle SQL Developer. It supports most of the usual data mining algorithms with the exception of neural networks – oddly enough. Support vector machines feature strongly, and the range of algorithms is adequate but not as extensive as some other offerings.
2. Oracle R Enterprise extends the database with a library of R functions and makes database tables and views available to the R environment as native R objects. Oracle positions this as

addressing statistical analysis, but in reality R encompasses many data mining algorithms – more than Oracle Data Mining.

For organisations wanting to move into big data Oracle provides two hardware/software solutions:

1. Oracle Big Data Appliance is a relatively low cost platform for running big data software, and specifically the Cloudera distribution of Hadoop and Oracle NoSQL Database Community Edition – take your pick or use both.
2. The Exalytics platform features in-memory processing for very high throughput of analysis tasks, and data visualisation and exploration tools.

Oracle does provide a Predictive Analytics add-in for Microsoft Excel. This utilises support vector machines and is something of an oddity – but an interesting end-user tool all the same.

A natural route for existing Oracle users with full capability to move into big data. Probably of interest to very large organisations only, and for those without an existing Oracle commitment there are alternative solutions to data mining needs.

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.

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 of 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.

SAP (Overall ★★★★★) (Existing SAP Users ★★★★★)

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.

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.

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.

TIBCO ★★★★★

Unlike many of its competitors Spotfire provides a full arsenal of visual and computational analytics tools. These deliver powerful analytical capabilities ranging from the preparation and distribution of data visualisations, to the development and implementation of sophisticated data mining models. Users take whatever they need. If they just need tools that provide access to various data sources and allow that data to be graphed, then Spotfire will oblige. If on the other hand a predictive sales model needs to be developed and embedded into a dashboard it will do this too. The bottom line with Spotfire is that there should never be any dead ends.

It should also be remembered that TIBCO is not a one trick pony. For over twenty years it has provided leading process and data integration technologies and Spotfire benefits from this infrastructure capability immensely.

Like: A full range of capabilities – from simple graphing to real-time analytics.

Dislike: The silly marketing messages – ‘two second advantage’ – really!

You need only download the Spotfire free trial to see that it will admirably serve the data visualisation needs of any reasonable user. All the usual suspects are available including tables, cross tables, graphical tables (tables with graphs in them), bar charts, line charts, combination line and bar charts, pie charts, scatter plots, 3D scatter plots, map charts, treeMaps, heatmaps, parallel coordinate plots and box plots. Interacting with these various graphical types is a very dynamic experience with some useful rescaling features, highlighting, creation of legends and so on. Hierarchies can also be created to add additional information to a chart. We might graph sales of our top twenty customers over the last year. By adding the industry sector to the x-axis the customers will be grouped accordingly. Obviously filters play an important part in all of this and when Spotfire loads data it provides a filter panel for each attribute with many, many configurable options.

Ad-hoc analysis, interactive reporting and dashboards are all supported and Spotfire provides various mechanisms for collaboration – both free form and structured. Spotfire WebPlayer provides a zero footprint web based environment to access published documents, and effectively makes the environment scalable to the most demanding of needs.

Data Sources

If the data storage mechanisms you use are not some home-grown oddity then Spotfire will be able to get at them. You have to remember that TIBCO has been in the data and process integration game a long time, and so they know how to access data. There is no point providing a list, it would be two pages long.

Analysis

So, you have created some graphs, charts, tables and dashboards, but now you want to get a little bit fancy. Maybe you want to do some correlation analysis or create some predictive models to embed into charts. This is where Spotfire really excels, and none of the newer generation of BI/analysis tools equal it in this respect. There are several components including Spotfire Statistical Services to execute R, S+, SAS and MATLAB routines, and Spotfire Miner, a graphical environment for

data mining and creating predictive models. And as you may or may not know, R is a free, Open Source statistics and analysis package that is very widely used, and so it makes sense to be able to embed R based analysis into your analytics environment.

Performance

Spotfire employs a hybrid of in-database (processing carried out in the database system environment) and in-memory (data held and processed in local memory – i.e. on your PC). In the main it works very well, although conducting analysis on 100GB of data will mean waiting a few minutes for it to be loaded into memory (if you have that much). This is the reality of processing large data sets, and you certainly wouldn't want to load a central server with dozens of such requests – you might be waiting more than a few minutes.

TIBCO

TIBCO's roots go way back to 1985 and for many years it majored on it's middleware products so that applications and data could be integrated. The actual name TIBCO (The Information Bus Company) was adopted in 1997.

The company is headquartered in Palo Alto US and has offices around the world. It is a publicly quoted company with revenues of around US\$1 billion.

Excel Data Mining Add-Ins

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.

It should be added that if your organization uses Microsoft SQL Server, and specifically SQL Server Analysis Services (SSAS), then an add-in is available for Excel called the Data Mining Add-in which supports the creation of data mining models with Excel – it's very powerful too.

11Ants Model Builder

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

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.

Open Source and Free Data Mining Tools

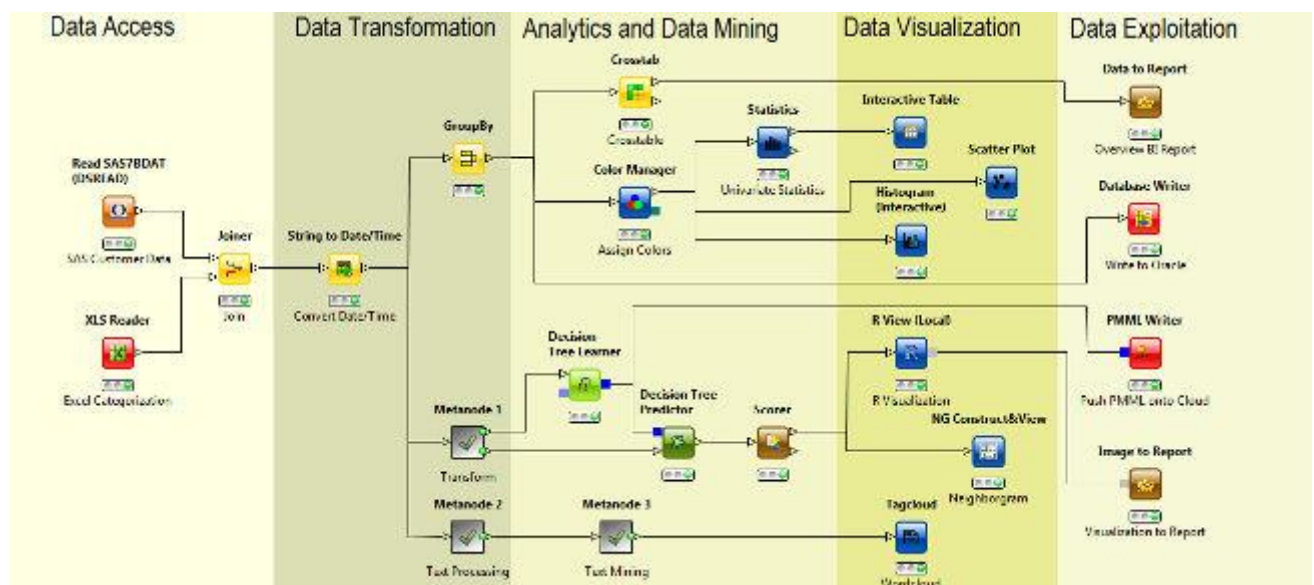
The five free data mining tools listed below are equally as capable as many products that have high price tags attached to them. They are in no way inferior, and most are Open Source with a large community of knowledgeable developers.

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.

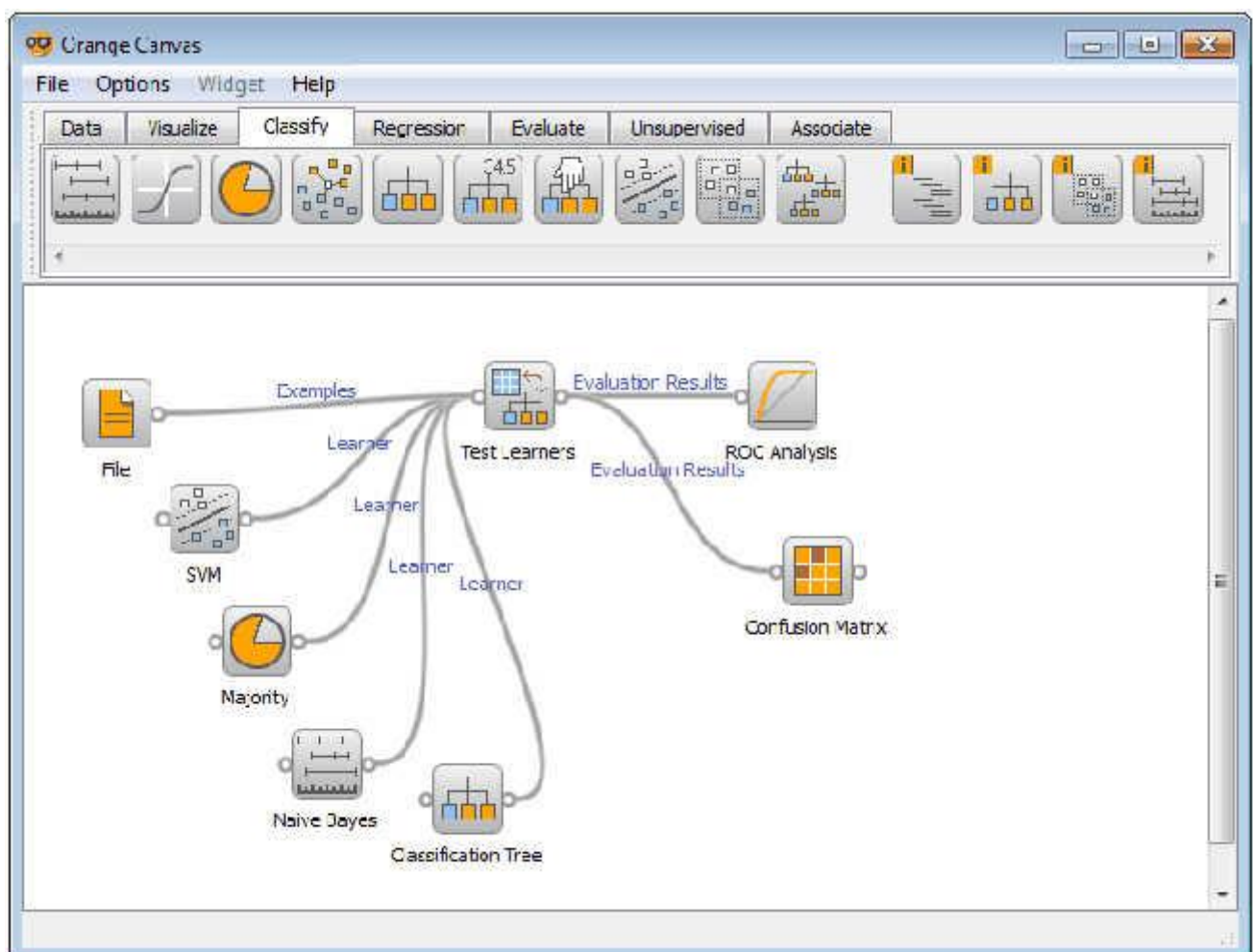


Orange

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.

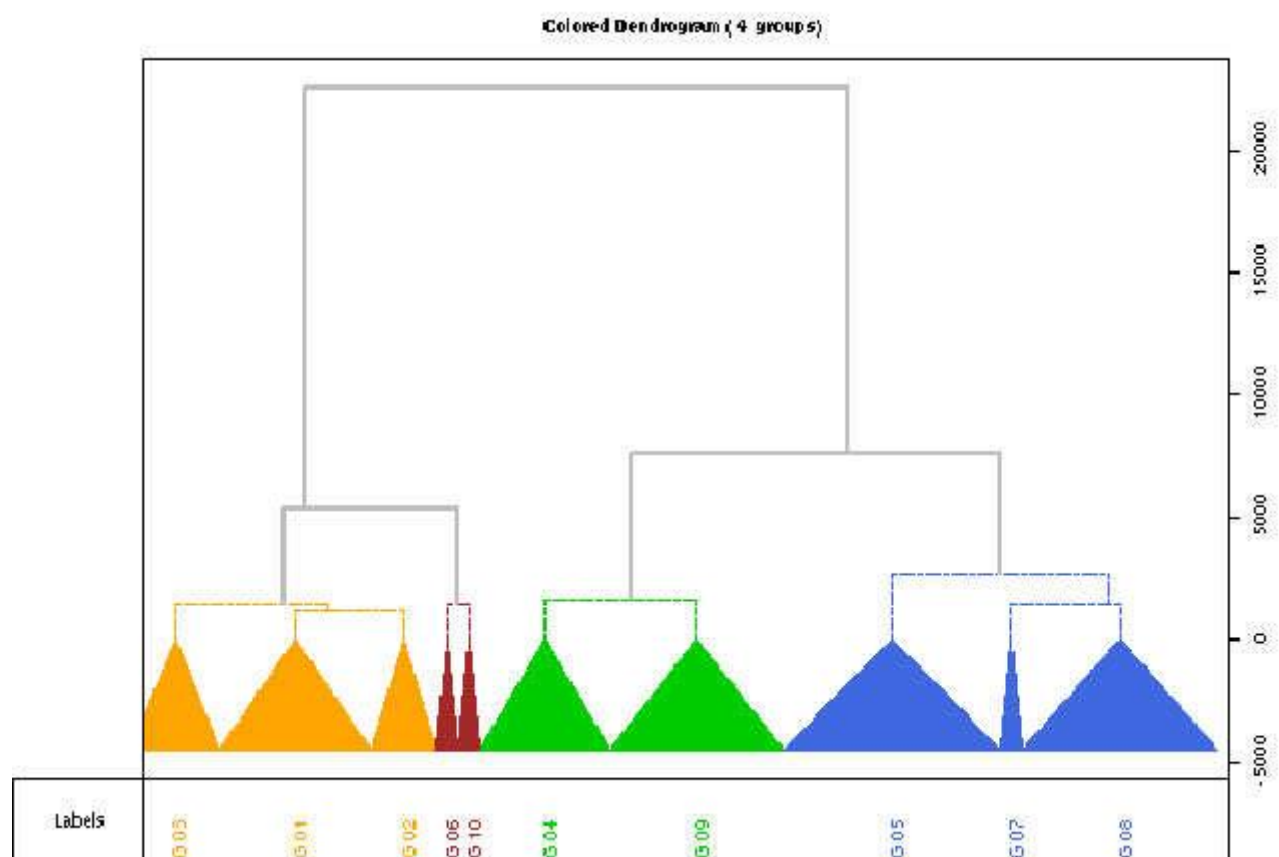
The documentation is good too and includes first steps, detailed widget descriptions and scripting. It runs on Windows, Mac OS X and Linux.



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.

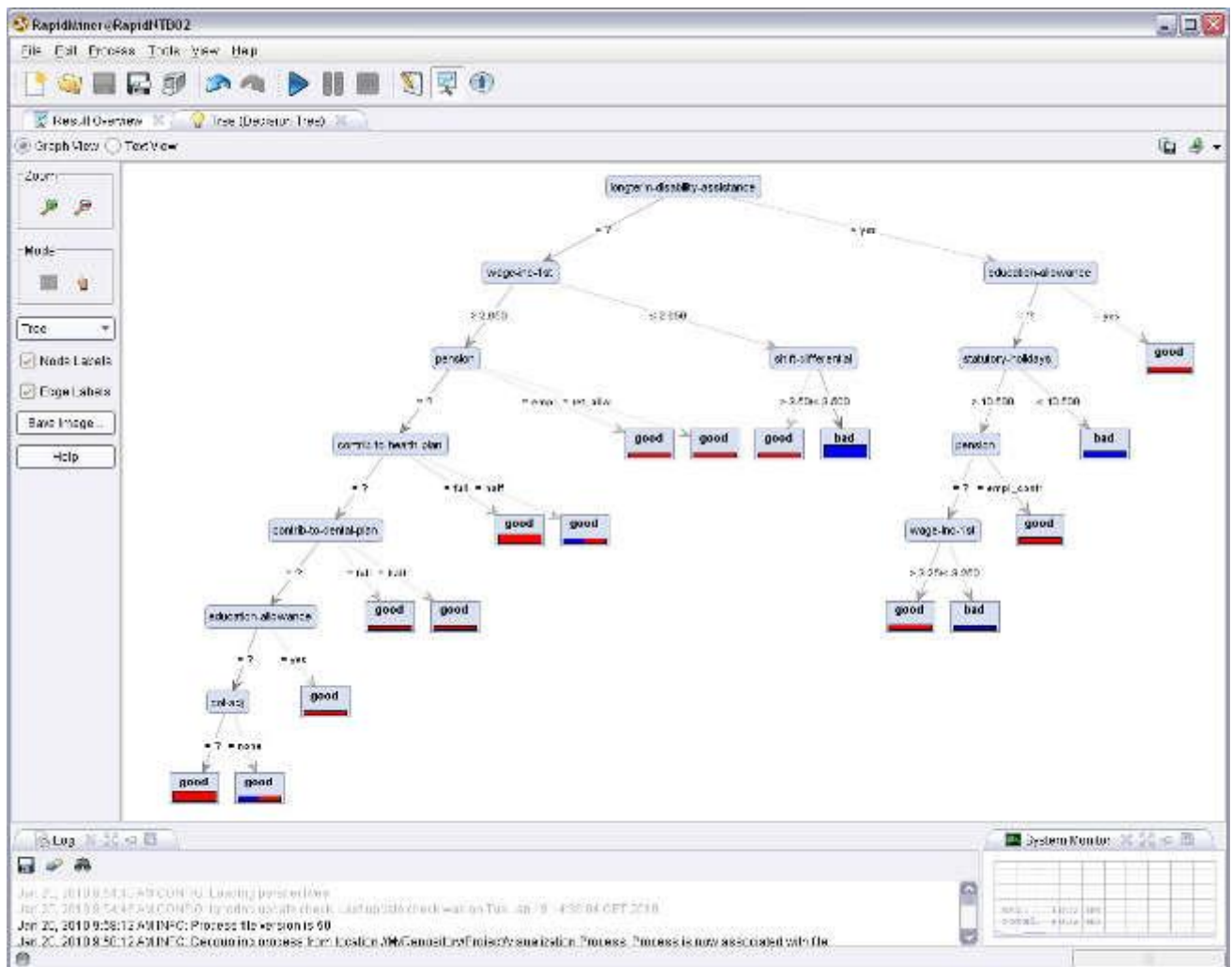


Rapid Miner

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.

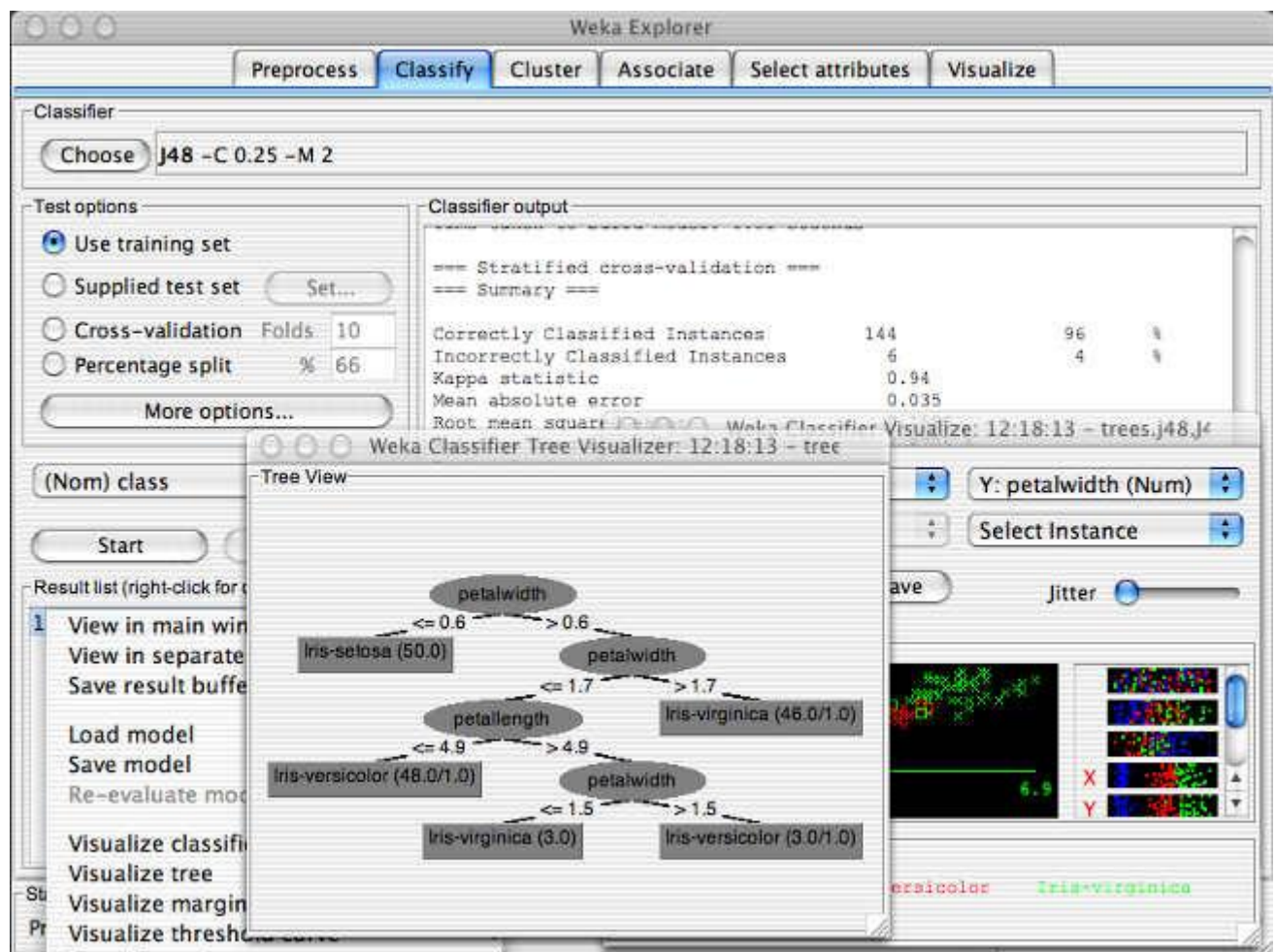


WEKA

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.



You might also want to look at [GGobi](#) if visualisation and exploratory data analysis is of interest, and [Tangara](#) for state-of-the-art algorithms but a somewhat less sophisticated user interface. [Predixion Software](#) also has a free developer edition of its excellent data mining tools with an Excel add-in.

4 Machine Learning Startups

It is clear from the positioning of these startups that machine learning may soon be wrested from the hands of machine learning experts, and put into the hands of the people who most need it – namely business users. All five of these startups are targeting the business user, and although this potentially has significant dangers associated with it, the advantages are obvious. This is after all the way of all technology – moving from the esoteric to the everyday, and it is very encouraging to see.

Alpine Data Labs

Alpine Data Labs provides an enterprise solution to the problems of data access, model building, model deployment and model performance monitoring for predictive analytics applications. As data mining and machine learning technologies move out of the realm of esoterica, so organizations will need a full implementation and management environment. I spoke with Steven Hillion, the Chief Product Officer at Alpine Data Labs and he was very aware of the need to satisfy the requirements of business management, as well as those of IT and data scientists. Many analytics technologies focus on the technical aspects with scant regard for the monitoring of model performance and the sharing of information in a collaborative environment. Although this is one of the less glamorous aspects of predictive technologies, in many ways it is one of the most important. Without the means to establish confidence in predictive models the technology will always be underexploited and untrusted.

One of the central philosophies of the Alpine approach concerns the movement, or rather lack of it, of data. Most data mining technologies require that data be extracted from the production or data warehousing environment and set aside for model development. Alpine takes a different approach and executes data exploration, data transformation, modeling, testing and deployment in the native database environment. In practice this means that models are created and execute in the same operating environment and on the same hardware as the database. Hadoop, Greenplum and most relational database systems are supported.

One of Alpine's central messages is a direct outcome of its ability to access data sources directly. Cross functional analytics is certainly possible with Alpine, and in some organizations will be possible. In reality however, highly stovepiped businesses will have a hard time exploiting this capability.

The user interface to Alpine's capabilities is through a the web browser, accessing a server that simply functions as an interface to data resources and as the broadcaster of the web based environment. At the data exploration phase it provides a plethora of visualization tools – frequency diagrams, box plots, scatter charts and so on. It provides the tools for data transformation modeling, testing and scoring.

All the popular mining algorithms are supported and there is ongoing activity aimed at creating fast implementations of these. Logistic regression and support vector machine are two methods which have been subjected to this treatment. Random Forests comes in two flavors. One is essentially exploratory in nature and executes faster than the full blown implementation, which is more likely to be used for creating a final model.

A new feature called Chorus is one of the stronger differentiators. This supports strong collaboration between data scientists, IT and management, and allows information pertaining to models to be freely shared. It's an effective way of building up a knowledge base for those working on and using models, and will eventually give management the visibility they need. It seems likely that Alpine will

partner with BI and data visualization tools providers to open up the analysis of data, and performance monitoring to a wide audience.

Alpine is one of a new generation of predictive analytics solution, providing a platform that satisfies all those with a stake in the exploitation of data mining technologies, and specifically IT, data scientists and management.

BigML

BigML offers a cloud based predictive analytics capability that is both refreshingly straightforward and extremely powerful. These two qualities are usually mutually exclusive, but by using decision trees and decision tree ensembles in conjunction with some very useful visualizations, the whole process of building and testing predictive models becomes so much easier.

The steps required to build, test and use a model are simple enough. Upload data to the BigML SaaS platform, build a model (which might be just a single operation), test it on test data and if all is well download it as Java, Python, PMML or any one of several formats. Then plug it in to your production systems.

The users of BigML will range from skilled business users through to data scientists and consultants. Obviously some level of knowledge and training is necessary, but a savvy business user should get the hang of things very quickly.

While BigML will produce a decision tree with considerable speed (typically in seconds or minutes), the real power is to be found in the decision tree ensembles where many trees are created and an 'average' created. A technique known as bagging is used where the data are randomly sampled multiple times (with replacement) and a tree created from each sample. It emulates having a much larger data set and nearly always produces much more accurate models.

The decision tree graphics are not only very visually appealing, but contains a great deal information and are interactive. A Sunburst visualization shows which classifications have most support and confidence in a highly graphic manner, allowing users to quickly home in on the the most useful classifications.

In my opinion the focus on decision tree ensembles is very appropriate. Various ensemble methods have won the vast majority of machine learning competitions in recent years, and have been called the most significant development in machine learning over the last decade. This is a good strategic decision by the founders of bigML.

The technology has found a broad range of applications including predictive marketing, fraud detection, recommendation systems, image analysis, pricing optimization and many others that satisfy very specific needs.

BigML went live to the public less than a year ago and obviously it plans the roll out of further capability and product. These include additional learning methods (k-means for example), non-linear decision trees and time series analysis. It will also be beefing up its cloud offerings to include virtual private clouds (VPC), and multi-cloud (for other cloud platforms – eg Azure).

BigML satisfies the requirement that 'Things should be made as simple as possible, but not any simpler.' very well, and is worthy of investigation by any organization (of any size) that needs to employ predictive tools.

SkyTree

SkyTree is primarily a server based approach to implementing data mining and machine learning techniques. An end-user tool (Adviser) is also being introduced to give power users access to machine learning techniques. Looks promising, but more information is needed. It isn't clear how the resulting models are put into a production environment, or how the end-user tool protects users from invalid models.

SkyTree Server connects with most commonly used data sources and executes many data mining methods to address categorization, clustering, association and regression. It incorporates advanced versions of many data mining and machine learning algorithms that both speed up execution and enhance capability.

The spec of SkyTree can be enhanced through PowerPack plugins. These include a Nonparametric Power Pack which includes specialized nearest neighbor algorithms, Prediction Power Pack for testing the validity of models and the Ensemble Power Pack for utilizing ensemble based methods (Random Forest and Gradient Boosted methods).

SkyTree Adviser is in Beta at the time of writing and is targeted at power users who want to do their own analysis. This is fraught with dangers, but it is to be hoped that the product makes users aware of these and offers mechanisms for addressing them. Advisor will handle up to 100,000 rows and will connect to databases, local data, and the web.

A variety of services are offered to provide initial support and ongoing training and support.

Wise.io

wise.io provides an extremely fast implementation of the Random Forest machine learning technique that is suitable for classifying complex, high dimensional data orders of magnitude more quickly than most alternatives. The technology was originally developed to search data generated by astronomical observations. The power of the technology is well expressed by its ability to learn and categorize handwriting within just a few minutes, compared with a week of learning using the favorite technology for this problem, the support vector machine.

I spoke with Joshua Bloom, the CEO of the company and UC Berkeley Associate Professor Astrophysics. He was keen to emphasize that the learning algorithm is just a small part of the overall model production and deployment cycle, although having a resource that executes at this speed makes many otherwise difficult problems amenable to a solution.

The company offers a SaaS facility called Machine Intelligence Engine where users upload their data, build a model and then either download it, or have it execute in the SaaS environment. Fees are levied according to the level of usage, and users typically require a period of hand-holding, which may range from a few hours to a few days. WiseRF on the other hand is downloadable and allows models to be built in-house. It comes in three flavors – Pine, Oak and Sequoia with increasing scalability and capability. A 15 day trial can be downloaded.

Applications range from OTC trading through to industrial safety, and while the accuracy of Random Forest is widely appreciated, having these very high levels of performance means that 'real-time' problems can be addressed.

wise.io is supported by its customers and is cash flow positive (a rare state for a startup) and will undoubtedly make a nice acquisition should the firm wish to go that way.

A Brief Introduction To Analytics Techniques

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.

Bayes 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). But if you want to learn more an introductory article can be found [here](#).

Decision Trees

Decision trees are a favourite 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 Neighbours (k-NN)

Entities can often be classified by the neighbourhood they live in. Simply ask whether your own neighbourhood 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 neighbourhood a particular record lives in. The official name for this algorithm is k-Nearest Neighbour, 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 neighbours 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.

Self-organising maps (SOM) represent a particular use of neural networks to perform cluster analysis. The major difference is that instead of a single output a SOM may have hundreds, traditionally arranged in a grid, although any topology can be used. Neighbouring nodes are connected through a neighbourliness parameter which affects the way the weights on the nodes are adjusted as the network is trained. Since each output node is connected to every input node it is possible to find average values for all the instances in a cluster, and as such characterise the cluster.

SOMs are unsupervised, discovering patterns through clusters, and are used in many areas of business including marketing, finance and credit risk analysis.

Association Rule Mining

Summary – Use Association Rule Mining (ARM) when you want to discover rules relating attribute values to each other. The general form of a rule is ‘IF this THEN that’, and in a supermarket shopping habits analysis an example might be IF milk THEN bread. Establishing the usefulness of rules is a major part of most ARM projects.

On the face of it association rule mining (ARM) is a simple enough activity. We simply let loose the ARM algorithms on the data of our choice and a plethora of rules will be found. The general format for these rules is as follows:

IF this THEN that

To add some meat to the bones let’s consider the habits of shoppers at a supermarket. A specific rule might say:

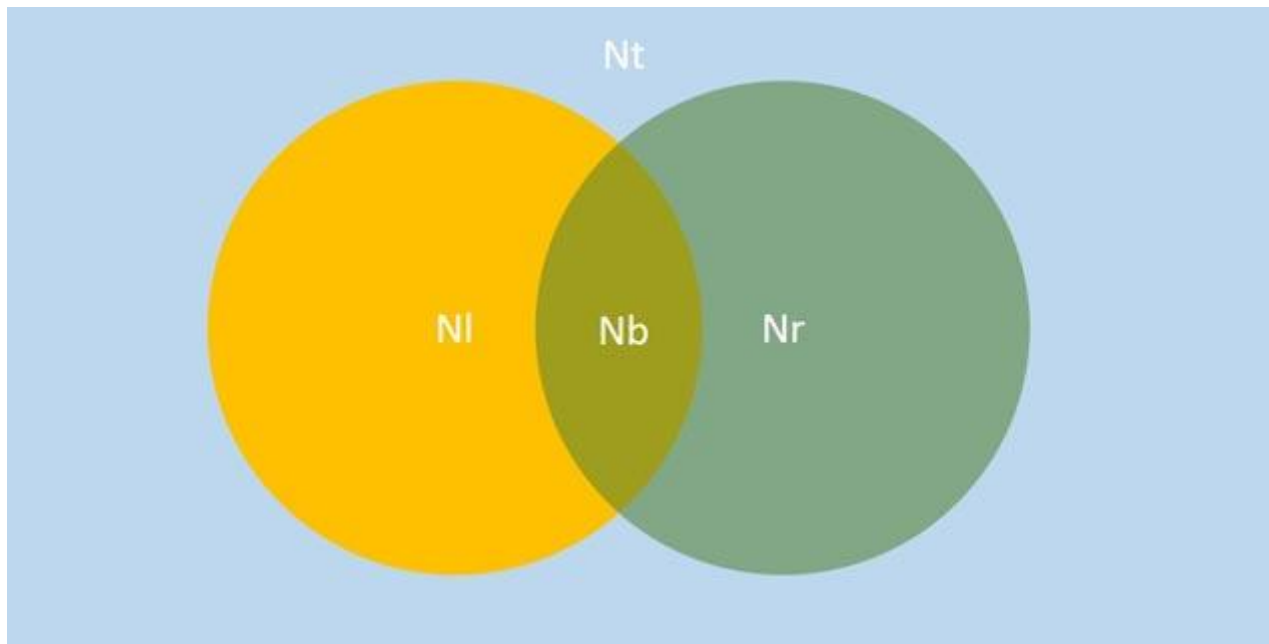
IF (bread and milk) THEN (Jam)

The ARM algorithm will have ploughed through millions of combinations of products to identify this particular rule and thousands of others. In fact this is one of the problems associated with ARM, it finds an excess of rules, many of which are simply not useful.

ARM is generally categorized as an unsupervised learning method. Supervised learning involves giving the data mining algorithms a target variable to classify (a good or bad loan prospect for example), and a data set to learn from. Unsupervised learning does not involve a target variable and is more exploratory in nature.

In the supermarket example the ARM algorithm will typically produce thousands of rules of the form mentioned above. Another one might be IF (cheese and apples) THEN (beef and oranges). When we look at the data we might find that the number of people buying cheese and apples together is quite small, and that only 100 cases out of millions of shopping baskets obeyed this rule – clearly it doesn’t mean very much.

So that we are delivered a meaningful set of rules, there are a number of measures that can be used to filter out the more trivial ones. The use of such measures is a major part of using ARM and so we will give it some attention. This involves the use of a Venn diagram, just to make things easier to understand.



In the diagram above we have three basic shapes. The blue rectangle represents all the shopping baskets we wish to consider, and this has a total of N_t shopping basket details. The yellow circle represents all the shopping baskets with the 'this' combination of goods. The green circle represents all the 'that' combination of goods, and the intersection of the two circles represents out rule. Going back to our basic rule formulation we can rephrase it as:

IF left THEN right

Both 'left' and 'right' represent combinations of items, and are labelled such because they appear on the left and right hand side of the rule. So NI is the number of instances (shopping baskets) that satisfy the 'left' combination of items (cheese and apples say). N_r is the number of instances that satisfy the 'right' combination of items (beef and oranges). The items which satisfy the actual rule are represented by the intersection of the two circles and number N_b . So having got the terminology out the way we can define three central measures used to filter useful rules.

The first is **Confidence** and this is calculated as N_b/NI . This is a measure of the number of baskets that satisfy the rule as a fraction of baskets that satisfy the left hand combination. To make this meaningful we will go back to our first rule:

IF (bread and milk) THEN (Jam)

Let's say there are 500,000 shopping baskets that contained bread and milk (NI) out of a total of 1,000,000 shopping baskets (N_t). And there are 300,000 shopping baskets that contained jam (N_r). The number of shopping baskets which contained bread, milk and jam is 200,000 (N_b). Confidence for this rule would be $200,000 / 500,000 = 0.4$ or 40%. Confidence gives a measure of how exclusive a rule is. In this case 60% of the bread and milk purchases were associated with other items. Confidence should be much higher than this – around 75%. So knowing someone purchased bread and milk does not really tell us all that much.

Support is specified as N_b/N_t . This gives us some measure of how 'big' this rule is. Typically users of ARM methods want a support of greater than 1%. A rule really doesn't mean very much if there are only five instances of it out of 1,000,000. In our example support is calculated as $200,000 / 1,000,000 = 0.2$ or 20% – a very respectable number.

Completeness is specified by **Nb/Nr**. This tells us whether the rule predicts a substantial number of the instances where the 'right' combination is present. If it only predicts a small fraction then it would not be much use in promoting those products on the 'right'. The completeness of our rule is $200,000 / 300,000$ or 0.67 or 67% – once again very acceptable. There are many other measures of a rule, but these are the most common.

ARM is used for many purposes other than analysing shopping baskets. It can be used to analyse credit card purchases, medical symptoms and other uses. In the next article we will get more into the algorithms used for ARM with their relative strengths and weaknesses.