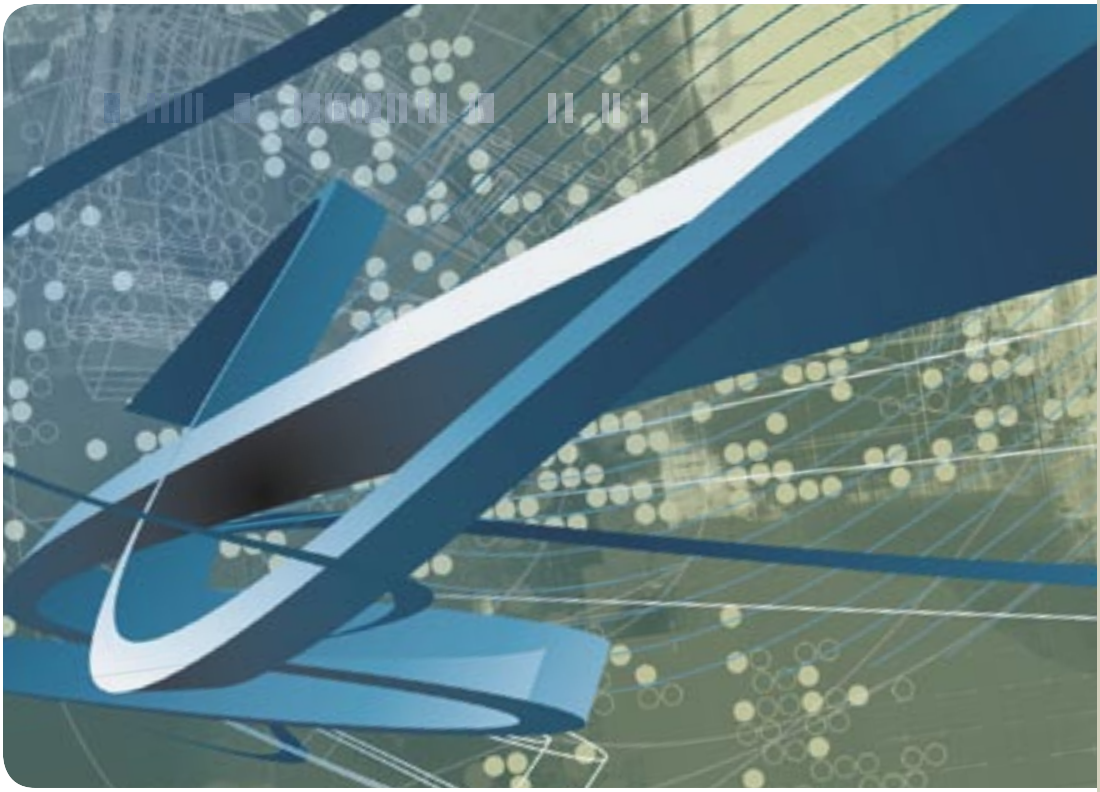
FIRST QUARTER 2007
TDWI BEST PRACTICES REPORT

PREDICTIVE ANALYTICS

Extending the Value of Your Data Warehousing Investment

By Wayne W. Eckerson



Sponsored by

sas

tdwi
THE DATA WAREHOUSING INSTITUTE

PREDICTIVE ANALYTICS

Extending the Value of Your Data Warehousing Investment

By Wayne W. Eckerson

Table of Contents

About the Author

WAYNE W. ECKERSON is director of research and services for The Data Warehousing Institute (TDWI), a worldwide association of business intelligence and data warehousing professionals that provides education, training, research, and certification. Eckerson has 17 years of industry experience and has covered data warehousing and business intelligence since 1995.

Eckerson is the author of many in-depth reports, a columnist for several business and technology magazines, and a noted speaker and consultant. He authored the book *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*, published by John Wiley & Sons in October 2005. He can be reached at weckerson@tdwi.org.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. Starting in 1995 with a single conference, TDWI is now a comprehensive resource for industry information and professional development opportunities. TDWI sponsors and promotes quarterly World Conferences, regional seminars, onsite courses, a worldwide Membership program, business intelligence certification, resourceful publications, industry news, an in-depth research program, and a comprehensive Web site: www.tdwi.org.

About TDWI Research

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, as well as those who responded to our requests for phone interviews. We would also like to recognize TDWI's account and production team: Jennifer Agee, Bill Grimmer, Denelle Hanlon, Deirdre Hoffman, and Marie McFarland.

Sponsors

MicroStrategy, Inc., OutlookSoft Corporation, SAS, SPSS Inc, Sybase, Inc., and Teradata, a division of NCR, sponsored the research for this report.

Research Methodology

Focus. This report is designed for the business or technical manager who oversees a BI environment and wishes to learn the best practices and pitfalls of implementing a predictive analytics capability. While it discusses some technical issues, it is designed to educate business managers about how to drive greater value from their existing investments in data warehousing and information delivery systems.

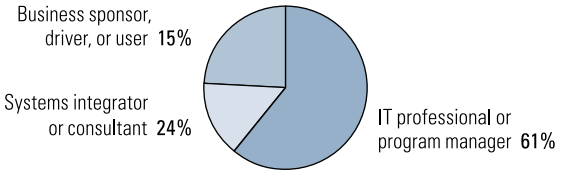
Methodology. The research for this report is based on a survey that TDWI conducted in August of 2006, as well as interviews with BI and analytics practitioners, consultants, and solution providers. To conduct the survey, TDWI sent e-mail messages to IT professionals in TDWI's and 1105 Media's databases. (TDWI is a business unit of 1105 Media.) A total of 888 people responded to the survey, including 55 people whose responses we did not count since they work for a BI vendor in a sales or marketing capacity, or are professors or students. Thus, our analysis was based on responses from 833 people. Of this group, 168 had either partially or fully implemented a predictive analytics solution. Most of our survey analysis is based on the answers provided by these 168 respondents. Percentages may not always add up to 100% due to rounding or questions that allow respondents to select more than one answer.

Respondent Profile. A majority of the 833 qualified survey respondents (61%) are corporate IT professionals who serve as mid-level managers in the United States and who work for large organizations. (See charts.)

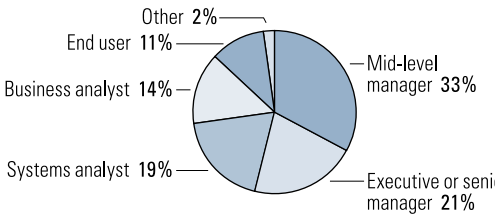
Company Profile. A majority (58%) work in groups that support the entire enterprise, while 20% support a business unit, and 16% support multiple departments. The industries with the highest percentage are consulting and professional services (13%), financial services (12%), software/internet (9%), and insurance (8%.) Respondents work for companies of various sizes. One-fifth of respondents (21%) hail from companies with less than $100 million in revenue a year, while another 26% of respondents come from companies that earn less than $1 billion, while 15% come from companies with annual revenues of between $1 billion and $5 billion.

Demographics

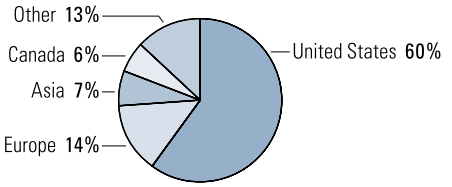
Position



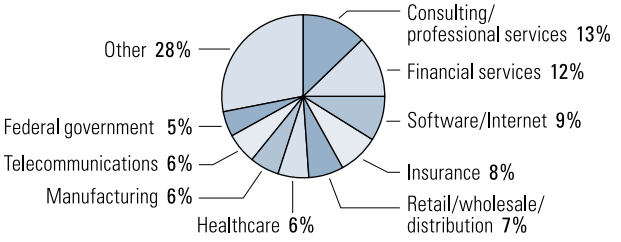
Role



Location



Industry



Based on 750 qualified respondents.

What Is Predictive Analytics?

Consider the power of predictive analytics:

- A Canadian bank uses predictive analytics to increase campaign response rates by 600%, cut customer acquisition costs in half, and boost campaign ROI by 100%.
- A large state university predicts whether a student will choose to enroll by applying predictive models to applicant data and admissions history.
- A research group at a leading hospital combined predictive and text analytics to improve its ability to classify and treat pediatric brain tumors.
- An airline increased revenue and customer satisfaction by better estimating the number of passengers who won't show up for a flight. This reduces the number of overbooked flights that require re-accommodating passengers as well as the number of empty seats.

As these examples attest, predictive analytics can yield a substantial ROI. Predictive analytics can help companies optimize existing processes, better understand customer behavior, identify unexpected opportunities, and anticipate problems before they happen. Almost all of TDWI's Leadership Award[1] winners in the past six years have applied predictive analytics in some form or another to achieve breakthrough business results.

High Value, Low Penetration. With such stellar credentials, the perplexing thing about predictive analytics is why so many organizations have yet to employ it. According to our research, only 21% of organizations have "fully" or "partially" implemented predictive analytics, while 19% have a project "under development" and a whopping 61% are still "exploring" the issue or have "no plans." (See Figure 1.)

Status of Predictive Analytics

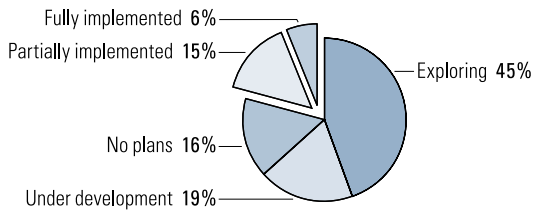


Figure 1. Predictive analytics is still in an early-adopter phase. Based on 833 respondents to a TDWI survey conducted August 2006.

Predictive analytics is also an arcane set of techniques and technologies that bewilder many business and IT managers. It stirs together statistics, advanced mathematics, and artificial intelligence and adds a heavy dose of data management to create a potent brew that many would rather not drink! They don't know if predictive analytics is a legitimate business endeavor or an ivory tower science experiment run wild.

[1]For many years, TDWI recognized the top overall applicant to its Best Practices Awards program with the TDWI Leadership Award for excellence in data warehousing and business intelligence. For more information on this program, visit www.tdwi.org/Education and click on Best Practices.

Where Do You Start? But once managers overcome their initial trepidation, they encounter another obstacle: how to apply predictive analytics optimally in their company. Most have only a vague notion about the business areas or applications that can benefit from predictive analytics. Second, most don't know how to get started: whom to hire, how to organize the project, or how to architect the environment.

Definitions

Before we address those questions, it's important to define what predictive analytics is *and* is not. Predictive analytics is a set of business intelligence (BI) technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events.[2] Unlike other BI technologies, predictive analytics is forward-looking, using past events to anticipate the future. (See Figure 2.)

Applications. Predictive analytics can identify the customers most likely to churn next month or to respond to next week's direct mail piece. It can also anticipate when factory floor machines are likely to break down or figure out which customers are likely to default on a bank loan. Today, marketing is the biggest user of predictive analytics with cross-selling, campaign management, customer acquisition, and budgeting and forecasting models top of the list, followed by attrition and loyalty applications. (See Figure 3.)

The Spectrum of BI Technologies

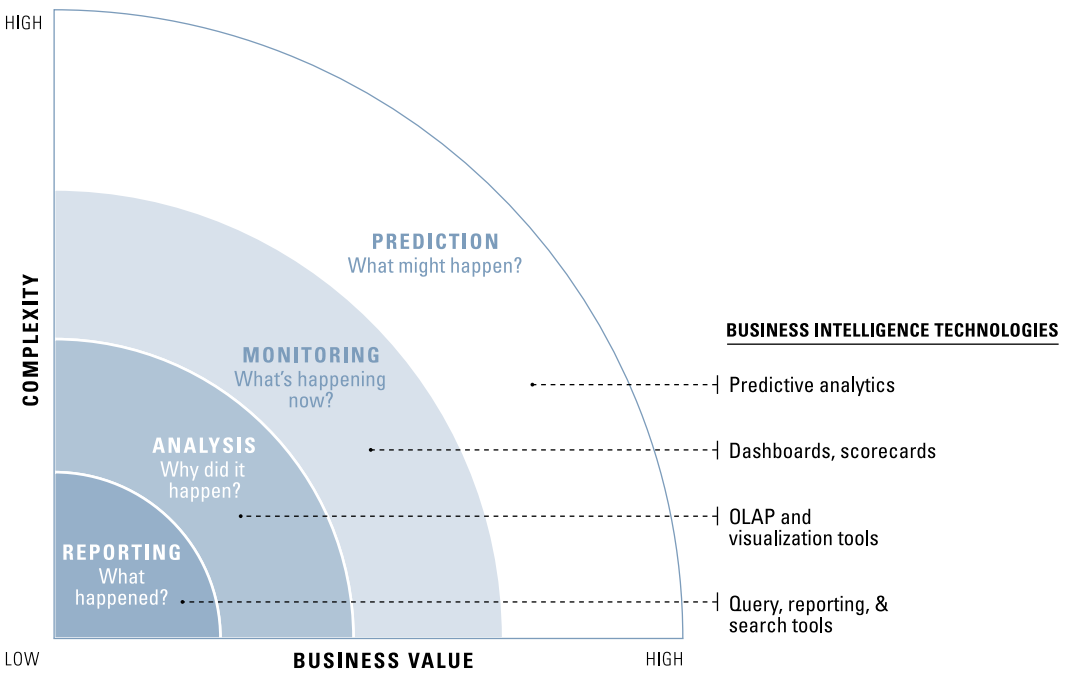


Figure 2. Among business intelligence disciplines, prediction provides the most business value but is also the most complex. Each discipline builds on the one below it—these are additive, not exclusive, in practice

[2] TDWI defines business intelligence as the tools, technologies, and processes required to turn data into information and information into knowledge and plans that optimize business actions. In short, business intelligence makes the businesses run more intelligently. It encompasses data integration, data warehousing, and reporting and analysis tools. Colloquially, most people use the term "BI tools" to refer to reporting and OLAP tools, not the full spectrum of BI capabilities.

Applications for Predictive Analytics

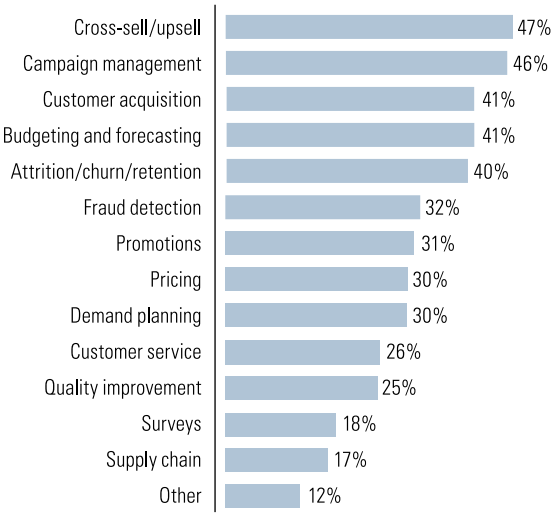


Figure 3. Based on 167 respondents who have implemented predictive analytics. Respondents could select multiple answers.

Versus BI Tools. In contrast, other BI technologies—such as query and reporting tools, online analytical processing (OLAP), dashboards, and scorecards—examine what happened in the past. They are ***deductive*** in nature—that is, business users must have some sense of the patterns and relationships that exist within the data based on their personal experience. They use query, reporting, and OLAP tools to explore the data and validate their hypotheses. Dashboards and scorecards take deductive reasoning a step further: they present users with a de facto set of hypotheses in the form of metrics and KPIs that users examine on a regular basis.

Predictive analytics lets data lead the way.

Predictive analytics works the opposite way: it is ***inductive***. It doesn't presume anything about the data. Rather, predictive analytics lets data lead the way. Predictive analytics employs statistics, machine learning, neural computing, robotics, computational mathematics, and artificial intelligence techniques to explore *all* the data, instead of a narrow subset of it, to ferret out meaningful relationships and patterns. Predictive analytics is like an "intelligent" robot that rummages through all your data until it finds something interesting to show you.

No Silver Bullet. However, it's important to note that predictive analytics is not a silver bullet. Practitioners have learned that most of the "intelligence" in these so-called decision automation systems comes from humans who have a deep understanding of the business and know where to point the tools, how to prepare the data, and how to interpret the results. Creating predictive models requires hard work, and the results are not guaranteed to provide any business value. For example, a model may predict that 75% of potential buyers of a new product are male, but if 75% of your existing customers are male, then this prediction doesn't help the business. A marketing program targeting male shoppers will not yield any additional value or lift over a more generalized marketing program.

Predictive analytics is statistics on steroids.

More Than Statistics. It's also important to note that predictive analytics is more than statistics. Some even call it statistics on steroids. Linear and logistic regressions—classic statistical

techniques—are still the workhorse of predictive models today, and nearly all analytical modelers use descriptive statistics (e.g., mean, mode, median, standard deviation, histograms) to understand the nature of the data they want to analyze.

However, advances in computer processing power and database technology have made it possible to employ a broader class of predictive techniques, such as decision trees, neural networks, genetic algorithms, support vector machines, and other mathematical algorithms. These new techniques take advantage of increased computing horsepower to perform complex calculations that often require multiple passes through the data. They are designed to run against large volumes of data with lots of variables (i.e., fields or columns.) They also are equipped to handle "noisy" data with various anomalies that may wreak havoc on traditional models.

Predictive analytics versus data mining.

Terminology. Predictive analytics has been around for a long time but has been known by other names. For much of the past 10 years, most people in commercial industry have used the term "data mining" to describe the techniques and processes involved in creating predictive models. However, some software companies—in particular, OLAP vendors—began co-opting the term in the late 1990s, claiming their tools allow users to "mine" nuggets of valuable information within dimensional databases. To stay above the fray, academics and researchers have used the term "knowledge discovery."

Today, the term data mining has been watered down so much that vendors and consultants now embrace the term "predictive analytics" or "advanced analytics" or just "analytics" to describe the nature of the tools or services they offer. But even here the terminology can get fuzzy. Not all analytics are predictive. In fact, there are two major types of predictive analytics, (1) supervised learning and (2) unsupervised learning.

Training Models. Supervised learning is the process of creating predictive models using a set of historical data that contains the results you are trying to predict. For example, if you want to predict which customers are likely to respond to a new direct mail campaign, you use the results of past campaigns to "train" a model to identify the characteristics of individuals who responded to that campaign. Supervised learning approaches include classification, regression, and time-series analysis. Classification techniques identify which group a new record belongs to (i.e., customer or event) based on its inherent characteristics. For example, classification is used to identify individuals on a mailing list that are likely to respond to an offer. Regression uses past values to predict future values and is used in forecasting and variance analysis. Time-series analysis is similar to regression analysis but understands the unique properties of time and calendars and is used to predict seasonal variances, among other things.

Unsupervised Learning. In contrast, unsupervised learning does not use previously known results to train its models. Rather, it uses descriptive statistics to examine the natural patterns and relationships that occur within the data and does not predict a target value. For example, unsupervised learning techniques can identify clusters or groups of similar records within a database (i.e., clustering) or relationships among values in a database (i.e., association.) Market basket analysis is a well-known example of an association technique, while customer segmentation is an example of a clustering technique.

Whether the business uses supervised or unsupervised learning, the result is an analytic model. Analysts build models using a variety of techniques, some of which we have already mentioned: neural networks, decision trees, linear and logistic regression, naive Bayes, clustering, association,

and so on. Each type of model can be implemented using a variety of algorithms with unique characteristics that are suited to different types of data and problems. Part of the skill in creating effective analytic models is knowing which models and algorithms to use. Fortunately, many leading analytic workbenches now automatically apply multiple models and algorithms to a problem to find the combination that works best. This advance alone has made it possible for non-specialists to create fairly effective analytical models using today's workbenches.

The Business Value of Predictive Analytics

Organizations with a "strike-it-rich" mentality are likely to get frustrated and give up.

Incremental Improvement. Although organizations occasionally make multi-million dollar discoveries using predictive analytics, these cases are the exception rather than the rule. Organizations that approach predictive analytics with a "strike-it-rich" mentality will likely become frustrated and give up before reaping any rewards. The reality is that predictive analytics provides incremental improvement to existing business processes, not million-dollar discoveries.

"We achieve success in little percentages," says a technical lead for a predictive analytics team in a major telecommunications firm. She convinced her company several years ago to begin building predictive models to identify customers who might cancel their wireless phone service. "Our models have contributed to lowering our churn rate, giving us a competitive advantage."

The company's churn models expose insights about customer behavior that the business uses to improve marketing or re-engineer business processes. For example, salespeople use model output to make special offers to customers at risk of churning, and the managers to change licensing policies that may be affecting churn rates.

Measuring Value

Our survey reinforces the business value of predictive analytics. Among respondents who have implemented predictive analytics, two-thirds (66%) say it provides "very high" or "high" business value. A quarter (27%) claim it provides moderate value and only 4% admit it provides "low" or "very low" value. (See Figure 4.)[3]

What Is the Business Value of Predictive Analytics to Your Organization?

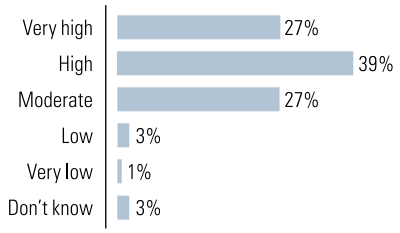


Figure 4. Based on 166 respondents who have implemented predictive analytics.

[3]Our respondents are generally individuals who create predictive models or manage analytic teams, so their perceptions of the business value they provide may be biased or differ from what business executives or managers might say. Nonetheless, I believe the responses generally align with my understanding of the success rates of predictive analytics in most organizations and other research conducted by vendors and research providers like International Data Corp.

How Do You Measure Success?

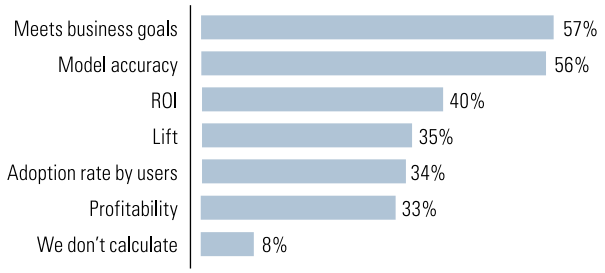


Figure 5. Based on 110 users who have implemented predictive analytics initiatives that offer "very high" or "high" value. Respondents could select multiple choices.

Respondents who selected "very high" or "high" in Figure 4 say they measure the success of their predictive analytics efforts with several criteria, starting with "meets business goals" (mentioned by 57% of respondents.) Other success criteria include "model accuracy" (56%), "ROI" (40%), "lift" (35%), and "adoption rate by business users" (34%.) (See Figure 5.)

Minimizing Churn. Brian Siegel is vice president of marketing analytics at TN Marketing, a firm that produces and distributes books and videos for its clients. He uses "lift" to measure the success of his predictive models. In a marketing campaign, lift measures the difference in customer response rates between customer lists created with and without a predictive model. As a one-man predictive analytics shop at his company, Siegel identifies people from client customer lists and outside lists who are likely to respond to a marketing campaign that his company conducts on behalf of a client.

"We have some cases where we don't need a whole lot of lift to achieve the ROI our president is looking for," says Siegel, who further states that he is successful eight out of ten times in achieving the response rates established by the marketing team. When he's not successful, Siegel says it's usually because the data set is too small or responses too random to offer predictive value.

Siegel is successful eight out of ten times in achieving the desired response rates.

Siegel is quick to translate the lift of his campaigns to business value. "Our response modeling efforts are worth millions," he says. "There have been a number of occasions where we would not have been able to acquire a new client and make the investment required to run a marketing campaign without the lift provided by our predictive models. So, I'm part of the sales process."

ROI. Interestingly, only a quarter of companies (24%) that have implemented predictive analytics have conducted a formal ROI study. This is about average for most BI projects based on past TDWI research. Companies with high-value analytic programs that have calculated ROI invest on average $1.36 million and receive a payback within 11.2 months. (These results are based on responses from only 37 survey respondents.)

The survey also asked respondents how much their group invests annually to support its predictive analytics practice, including hardware, software, staff, and services. The median investment is $600,000 for all respondents that have implemented predictive analytics, but $1 million for respondents with programs delivering "very high" or "high" business value. These results suggest that you get what you pay for. (See Table 1.)

Companies with successful analytics programs invest $1 million annually.

Median Investments in Predictive Analytics

	INVESTMENT
All Companies	$600,000
Companies with "high value programs"	$1 million

Table 1. Companies whose predictive analytics practice delivers "very high" or "high" business value (see Figure 4) invest more money than companies whose programs deliver "moderate" or lower value. Based on 166 and 110 respondents, respectively.

Drilling down more, the survey asked respondents to report their investments in predictive analytics by staff, software, hardware, and external services. Not surprisingly, staff costs consume the lion's share of expenses, followed by software and hardware. Organizations spend only 10% of their total budget on external service providers, either consultants or service bureaus. (See Figure 6.)

Median Breakdown of Expenses on Predictive Analytics

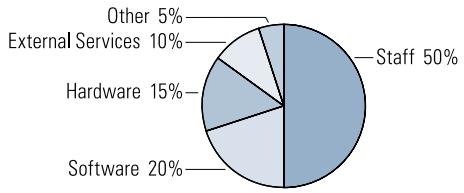


Figure 6. Median numbers are based on 166 respondents whose groups have implemented predictive analytics.

How Do You Deliver Predictive Analytics?

What Now? While some organizations have discovered the power of predictive analytics to reduce costs, increase revenues, and optimize business processes, the vast majority are still looking to get in the game. Today, most IT managers and some business managers understand the value that predictive analytics can bring, but most are perplexed about where to begin.

"We are sitting on a mountain of gold but we're not mining it as effectively as we could."

"We are sitting on a mountain of gold but we're not mining it as effectively as we could," says Michael Masciandaro, director of business intelligence at Rohm & Haas, a global specialty materials manufacturer. "We say we do analytics, but it's really just reporting and OLAP."

Rohm & Haas has hired consultants before to build pricing models that analyze and solve specific problems, but these models lose their usefulness once the consultants leave. Masciandaro says building an internal predictive analytics capability could yield tremendous insights and improve the profitability of key business areas, but he struggles to understand how to make this happen.

"How do you implement advanced analytics so they are not a one-off project done by an outside consultancy?" says Masciandaro. "How do you bring this functionality in house and use it to deliver value every day? And where do you find people who can do this? There are not too many of them out there."

The Process of Predictive Modeling

Most analytic modelers adhere to a methodology to ensure success.

Methodologies. Although most experts agree that predictive analytics requires great skill—and some go so far as to suggest that there is an artistic and highly creative side to creating models—most would never venture forth without a clear methodology to guide their work, whether explicit or implicit. In fact, process is so important in the predictive analytics community that in 1996 several industry players created an industry standard methodology called the Cross Industry Standard Process for Data Mining (CRISP-DM.)[4]

CRISP-DM. Although only 15% of our survey respondents follow CRISP-DM, it embodies a common-sense approach that is mirrored in other methodologies. (See Figure 7.) "Many people, including myself, adhere to CRISP-DM without knowing it," says Tom Breur, principal of XLNT Consulting in the Netherlands. Keith Higdon, vice president and practice leader for business intelligence at Sedgwick Claims Management Services, Inc. (CMS), adds, "CRISP-DM is a good place to start because it's designed to be cross-industry. But then you have to think, 'What makes my world unique?'"

What Methodology Does Your Group Use?

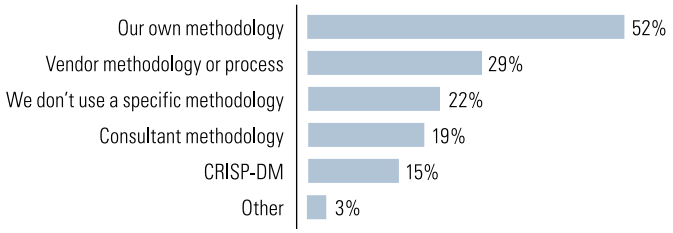


Figure 7. Based on 167 respondents who have implemented predictive analytics.

Regardless of methodology, most processes for creating predictive models incorporate the following steps:

1. **Project Definition:** Define the business objectives and desired outcomes for the project and translate them into predictive analytic objectives and tasks.
2. **Exploration:** Analyze source data to determine the most appropriate data and model building approach, and scope the effort.
3. **Data Preparation:** Select, extract, and transform data upon which to create models.
4. **Model Building:** Create, test, and validate models, and evaluate whether they will meet project metrics and goals.
5. **Deployment:** Apply model results to business decisions or processes. This ranges from sharing insights with business users to embedding models into applications to automate decisions and business processes.
6. **Model Management:** Manage models to improve performance (i.e., accuracy), control access, promote reuse, standardize toolsets, and minimize redundant activities.

Most experts say the data preparation phase of creating predictive models is the most time-consuming part of the process, and our survey data agrees. On average, preparing the data

[4] The impetus for CRISP-DM came from NCR, Daimler Chrysler, and SPSS, who in 1997 formed an industry consortium and obtained funding from the European Commission to establish an industry-, tool-, and application-neutral standard process for data mining. Today, an open special interest group of more than 330 users, vendors, consultants, and researchers supports the CRISP-DM 2.0 initiative. See www.crisp-dm.org.

occupies 25% of total project time. However, model creation, testing, and validation (23%) and data exploration (18%) are not far behind in the amount of project time they consume. This suggests that data preparation is no longer the obstacle it once was. However, if you combine data exploration and data preparation, then data-oriented tasks occupy 43% of the time spent creating analytic models, reinforcing the notion that data preparation consumes the lion's share of an analytic modeler's time. (See Figure 8.)

Project Breakdown: Average Time Spent per Phase

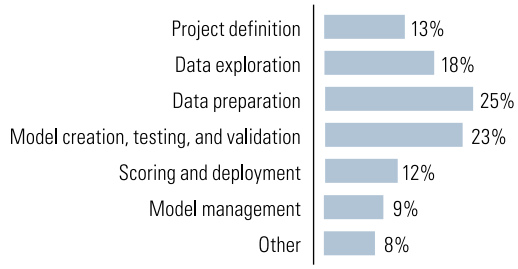


Figure 8. Percentage of time groups spend on each phase in a predictive analytics project. Averages don't equal 100% because respondents wrote a number for each phase. Based on 166 responses.

1. Defining the Project

Although practitioners don't spend much time defining business objectives, most agree that this phase is most critical to success. The purpose of defining project objectives is to discourage analytical fishing excursions where someone says, "Let's run this data through some predictive algorithms to see what we get." These projects are doomed to fail.

Collaboration with the Business. Defining a project requires close interaction between the business and analytic modeler. "I work daily with our marketing people," says a business analyst. To create a predictive model, this analyst meets with all relevant groups in the marketing department who will use or benefit from the model, such as campaign managers and direct mail specialists, to nail down objectives, timeframes, campaign schedules, customer lists, costs, processing schedules, how the model will be used, and expected returns. "There are a lot of logistics to discuss," she says.

2. Exploring the Data

Models are only as good as the data used to create them.

The data exploration phase is straightforward. Modelers need to find good, clean sources of data since models are only as good as the data used to create them. Good sources of data have a sufficient number of records, history, and fields (i.e., variables) so there is a good chance there are patterns and relationships in the data that have significant business value.

On average, groups pull data from 7.8 data sources to create predictive models. ("High value" predictive projects pull from 8.6 data sources on average.) However, a quarter of groups (24%) use just two sources, and 40% use fewer than five sources. Most organizations use a variety of different data types from which to build analytical models, most prominently transactions (86%), demographics (69%), and summarized data (68%). (See Figure 9.)

Fortunately, most of this data is already stored in a data warehouse, minimizing the time required to search for data across multiple systems. According to survey respondents, 68% of the data used to create predictive models is already stored in a data warehouse.

What Types of Data Do You Use to Create Predictive Models?

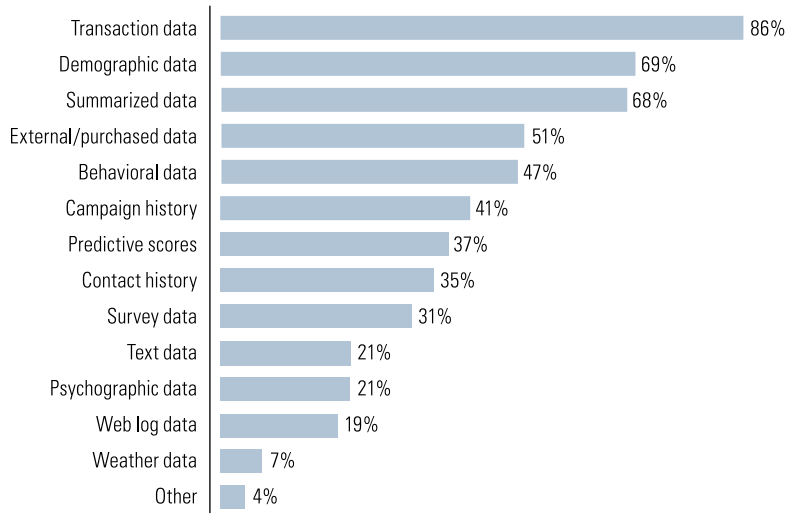


Figure 9. Based on 149 respondents that have implemented predictive analytics.

The tools help Siegel quickly identify the top 20 or so variables out of 200 or more.

TN Marketing's Siegel uses an analytical tool and descriptive statistics to analyze the quality and predictive characteristics of various data sets, which he downloads directly from source systems to his desktop. The tools and techniques help him quickly identify 20 or so variables out of 200 that offer the best chance of delivering good performance (i.e., models that accurately predict values.) However, some data sets don't generate accurate models because (1) there are too many missing values or errors, (2) the data is too random, or (3) the data doesn't accurately reflect what it's supposed to represent.

Tools. Predictive modelers use a variety of tools to explore and analyze source data. Most analytical tools offer some exploratory capabilities. Basic tools enable analysts to compile descriptive statistics of various fields (e.g., min/max values and standard deviation), while others incorporate more powerful data profiling tools that analyze the characteristics of data fields and identify relationships between columns within a single table and across tables. Data profiling tools are common in data quality projects and are offered by most leading data quality and data integration vendors. A small percentage of analysts use advanced visualization tools that let users explore characteristics of source data or analyze model results visually.

3. Preparing the Data

Preparing data means cleaning and flattening it.

Cleaning and Transforming. Once analysts select and examine data, they need to transform it into a different format so it can be read by an analytical tool. Most analysts dread the data preparation phase, but understand how critical it is to their success. "I'm going through the painful process right now of scrubbing data for a project," says Siegel.

Preparing data means first cleaning the data of any errors and then "flattening" it into a single table with dozens, if not hundreds, of columns. During this process, analysts often reconstitute fields, such as changing a salary field from a continuous variable (i.e., a numeric field with unlimited values) to a range field (i.e., a field divided into a fixed number of ranges, such as $0–$20,000, $20,001–$40,000, and so forth), a process known as "binning." From there, analysts usually perform additional transformations to optimize the data for specific types of algorithms. For example, they may create an index from two fields using a simple calculation, or aggregate data in one or more fields, such as changing daily account balances to monthly account balances.

4. Building Predictive Models

Creating analytic models is both art and science.

Creating analytic models is both art and science. The basic process involves running one or more algorithms against a data set with known values for the dependent variable (i.e., what you are trying to predict.) Then, you split the data set in half and use one set to create a training model and the other set to test the training model.

If you want to predict which customers will churn, you point your algorithm to a database of customers who have churned in the past 12 months to "train" the model. Then, run the resulting training model against the other part of the database to see how well it predicts which customers actually churned. Last, you need to validate the model in real life by testing it against live data.

Iterative Process. As you can imagine, the process of training, testing, and validation is iterative. This is where the "art" of analytic modeling comes to the forefront. Most analysts identify and test many combinations of variables to see which have the most impact. Most start the process by using statistical and OLAP tools to identify significant trends in the data as well as previous analytical work done internally or by expert consultants. They also may interview business users close to the subject and rely on their own knowledge of the business to home in on the most important variables to include in the model.

Creating models is very labor intensive and time-consuming.

As a result, most analysts cull the list of variables from a couple hundred in an initial version to a couple dozen in the final model. Along the way, they test a variety of algorithms to see which works best on the training data set. They may find it necessary to add new data types or recombine existing fields in different ways to improve model accuracy. This iterative process makes creating models labor-intensive and time-consuming.

Selecting Variables. Most analysts can create a good analytic model from scratch in about three weeks, depending on the scope of the problem and the availability and quality of data. Most start with a few hundred variables and end up with 20 to 30. This agrees with our survey results showing that a majority of groups (52%) create new models within "weeks" and another third (34%) create new models in one to three months. (See Figure 10.) Once a model is created, it takes about half the groups (49%) a matter of "hours" or "days" to revise an existing model for use in another application and takes another 30% "weeks" to revise a model. In addition, about half (47%) of models have a lifespan shorter than a year, and one-third (16%) exist for less than three months.

How Long Does It Take to Create a New Model from Scratch?

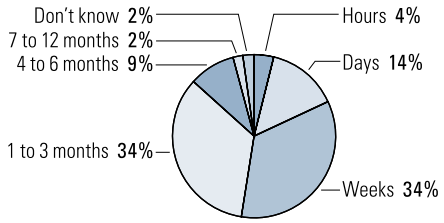


Figure 10. Based on 163 respondents.

How Many Variables Do You Use in Your Models?

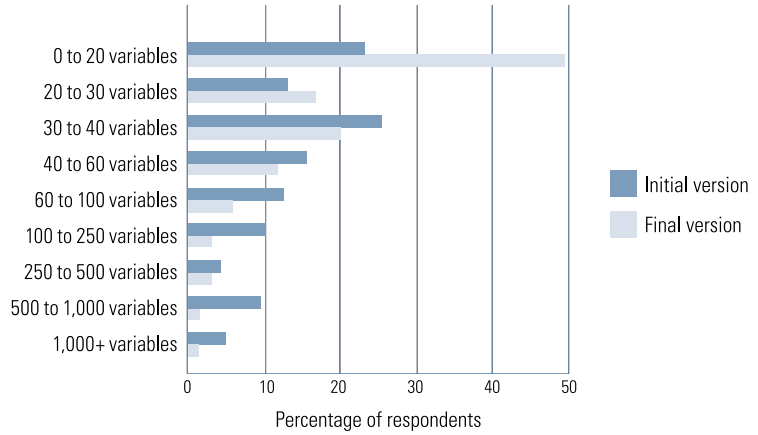


Figure 11. Based on 156 respondents.

5. Deploying Analytical Models

Focus on Business Outcomes. A predictive model can be accurate but have no value. Predictive models can fail if either (1) business users ignore their results or (2) their predictions fail to produce a positive outcome for the business. The classic story about a grocery that discovered a strong correlation between sales of beer and diapers illustrates the latter situation. Simply identifying a relationship between beer and diaper sales doesn't produce a valuable outcome. Business users must know what to do with the results, and their decision may or may not be favorable to the business.

Identifying a data pattern doesn't mean the business will know how to exploit it.

For example, the business could decide to display beer and diapers together at the front of the store to encourage more shoppers to purchase both items. Or they could decide to place the dual beer-and-diaper display at the back of the store to force shoppers to move through more aisles to obtain these items. Or they could place beer and diapers at separate ends of the store to force shoppers to spend the maximum time possible walking through the aisles. Their decision, not the model results, ultimately determines business value.

Clever store managers can also use predictive models to help make this decision. By formulating a simple test with well-matched test and control groups, the manager can accurately anticipate the revenue impact of different beer and diaper placements using predictive modeling techniques.

There are many ways to deploy a predictive model:

A) Share the Model. You can share insights with business users via a paper report, a presentation, or a conversation. For example, Sedgwick CMS creates analytic models to enhance the claims management process and offer recommendations for business process improvements. "Most of our models are part of a consultative approach to minimize our client's cost of risk," says Higdon.

Most organizations transform a predictive model into a SQL statement or programming code.

B) Score the Model. Most organizations transform a predictive model into a SQL statement or programming code and then apply the statement or code to every single record in the company's database pertaining to the subject area of the model. The result is a "score," usually a value between 0 and 1, that gets inserted into the database record as an additional field. A marketing manager, for example, might then select customers for a direct mail campaign who scored above 0.7 in a predictive model that measures customers' propensity to purchase a specific product and respond to mail campaigns. Typically, companies score records on a monthly basis since the scoring process can consume a lot of time and processing power.

A growing number of companies are starting to score models dynamically as new records arrive.

A growing number of companies are starting to score models dynamically as records arrive. Some do this to cut the time and expense of processing large numbers of records in batch, while others find it can be highly profitable. For example, dynamic scoring enables an e-commerce outfit to display cross-sell offers to Web customers who just purchased or viewed a related item. Or a manufacturing company can use dynamic scoring to schedule maintenance for a factory floor machine that is about to break.

Our survey shows that most organizations score models monthly or quarterly. However, 17% score models weekly, 13% score daily, and 19% score dynamically. (See Figure 12.)

How Frequently Do You Score Your Models?

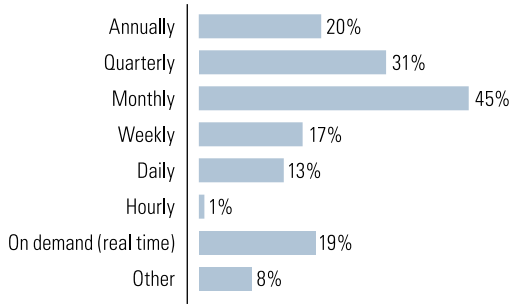


Figure 12. Based on 161 respondents who could select multiple answers.

C) Embed the Model in a BI Report. Predictive models are understood by only a handful of people and dispersed to even fewer. It's not that predictive models and scores exceed the comprehension of the average business user, but most users have never been exposed to these models and don't know what to do with the results. Embedding these results into BI tools and reports is one way to

overcome the hurdle of limited distribution. However, this may require modifying the model or report to make it easier for end users to interpret the results.

Corporate Express, Inc., a business-to-business supplier of office and computer products, created a logistic regression model to predict customer churn using analytical capabilities built into the MicroStrategy 8 BI platform. The company runs the model weekly against its entire database of 60,000 customers and delivers highly accurate results. However, the company had to figure out how to disseminate the results so salespeople could put the results to good use.

"We had to simplify the model to make it usable for the sales reps. These folks speak in terms of average order size, not R-squared values," says Matt Schwartz, director of business analysis at Corporate Express.

"We had to simplify the model to make it usable for the sales reps."

Tripwires. Rather than presenting salespeople the complete churn model for each customer, Corporate Express uses what it calls "tripwires" to highlight one variable in the model that might cause the customer to churn. For example, the report might highlight the model's frequency variable (i.e., time between purchases) to show that a customer who has purchased toner every week for the past year has not made a purchase in the past month. "The tripwires give the sales reps something to discuss with customers," says Schwartz.

Schwartz feared the model might predict churn that the reps already knew about or overwhelm them with too many customers to contact. However, the weekly churn reports only highlight a handful of customers for each salesperson. "We've reduced our attrition and anecdotal evidence from our salespeople is all positive," says Schwartz.

Corporate Express shows it's possible to distribute predictive models to a general population. The challenge is not the technology; it's delivering predictive results in a form that users can understand and apply. Once organizations learn how to display predictive results, BI tools will be a good vehicle for distributing regression models to support forecasting and variance analysis applications, according to Gerry Miller, technology integration services principal at Deloitte Consulting LLP.

Corporate Express shows it's possible to distribute predictive models to a general population.

D) Embed the Model in an Application. Another way to deploy a predictive model is to embed it into an operational application so it drives business actions automatically. To operationalize a predictive model, you need to embed the model results or scores into a set of rules. The rules usually create an "if, then, else" statement around a score. So, a Web recommendation engine might create the rule, "If a customer who just purchased this product exhibits a product affinity score of 0.8 or higher, then display pictures of the following items with the text, 'You may also be interested in purchasing these other items:'"

Some rules are based entirely on model scores, while others use scores as one element in a more complex rule that includes other variables, such as the time of day or month, type of customer, or type of product. For example, Sedgwick CMS uses complex rules composed of model scores and other variables to trigger referrals in its managed care division. It applies the rules on an ongoing basis to alert a claim handler when a claim possesses high-risk characteristics and should be treated differently, says Higdon.

The Holy Grail of operationalized predictive models is to create an automated environment in which models and scores are both dynamically updated and applied as new events occur. Experts refer to this state as a "lights out" decision-making environment or decision automation. This type of online processing is only suitable for well-known processes where the actions are highly scripted.

The Holy Grail of operationalized predictive models is to create an automated environment.

For example, a model might trigger the creation of new purchase orders when inventory falls below a certain threshold. Or, a model might detect fraudulent transactions and recommend actions based on the characteristics of the transaction. In reality, however, a human component will almost always exist in these "automated" decision-making environments. You need people who understand the model and the business and can validate the model results and recommended actions before they are executed. Without human intervention, companies risk making poor decisions based on faulty or spurious model results.

Our survey shows use of all the approaches mentioned above. Two-thirds (65%) use predictive model insights to "guide decisions and plans." A slight majority (52%) use models to "score records," while 41% "import models into BI tools or reports," 36% use scores to "create or augment rules," and 33% "embed rules or models in applications to automate or optimize processes." (See Figure 13.)

What Does Your Group Do with the Models It Creates?

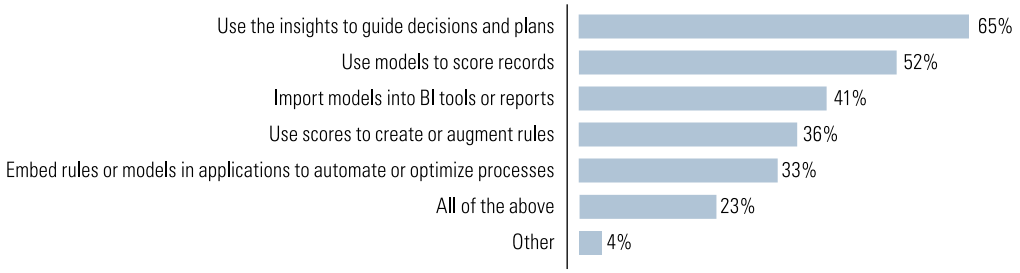


Figure 13. Based on 166 respondents selecting multiple answers.

6. Managing Models

The last step in the predictive analytics process is to manage predictive models. Model management helps improve performance, control access, promote reuse, and minimize overhead. Currently, few organizations are concerned about model management. Most analytical teams are small and projects are handled by individual modelers, so there is little need for check in/check out and version control. "We don't have a sophisticated way of keeping track of our models, although our analytical tools support model management," says one practitioner. She says her four-person team, which generates about 30 models monthly, maintains analytical models in folders on the server.

Demand for model management will increase to comply with new compliance regulations.

However, some expect an increase in the demand for model management to enable compliance and auditability with new standards and regulations. IT managers, in particular, want to impose greater structure on ad hoc analysis activities and multi-vendor analytical environments to minimize risks. Although model management can help teams of analytical modelers work more efficiently, few currently work within a rigorous project environment that adheres to industry standards for designing, creating, and publishing models.

A majority of organizations (61%) still use an ad hoc or project-based approach to developing analytical models, according to our survey. Only 36% have either a program office or Center of Excellence to coordinate predictive modeling tasks. These results expose the relative immaturity of the practice of managing predictive analytics projects. (See Figure 14.)

Which Best Describes Your Group's Approach to Model Management?

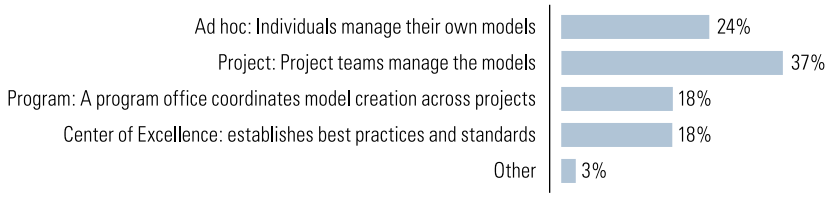


Figure 14. Based on 164 respondents.

Trends in Predictive Analytics

Analytical Immaturity. The BI community has seen a groundswell of interest in predictive analytics in the past two years as more companies seek to derive greater value from their data warehousing investments. However, few organizations have taken the plunge into predictive analytics, and for good reason.

"A lot of companies want to do predictive analytics, but have yet to master basic reporting. Until they get there, investing in predictive analytics isn't a good way to spend their money," says Deloitte Consulting's Miller.

"A lot of companies want to do predictive analytics, but have yet to master basic reporting."

Even organizations that have implemented predictive analytics have yet to harness the technology's potential. Only about one-third of organizations (36%) say they have implemented predictive analytics in a mature fashion that uses well defined processes and measures of success and enables them to continuously evaluate and improve their modeling efforts. (See Figure 15.)

Describe the Maturity of Predictive Analytics in Your Group

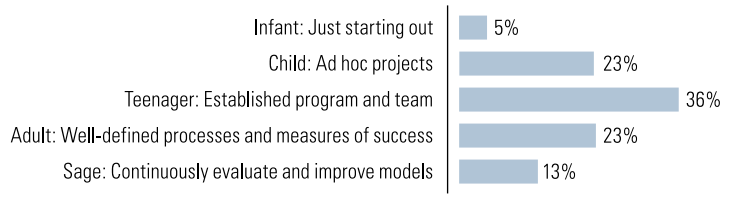


Figure 15. Based on 168 responses from groups that have implemented predictive analytics.

Analytics Bottleneck

One reason for the overall lack of maturity in predictive analytics is that most companies haven't done it for very long, at least among our survey base. Thirty-eight percent of organizations that have implemented predictive analytics have had the technology for two years or less, and 60% have had it in house less than four years. This is not much time to optimize or master a complex discipline!

Barriers to Usage. A host of barriers can prevent organizations from venturing into the domain of predictive analytics or impede their growth. This "analytics bottleneck" arises from:

1. **Complexity.** Developing sophisticated models has traditionally been a slow, iterative, and labor intensive process.
2. **Data.** Most corporate data is full of errors and inconsistencies but most predictive models require clean, scrubbed, expertly formatted data to work.
3. **Processing Expense.** Complex analytical queries and scoring processes can clog networks and bog down database performance, especially when performed on the desktop.
4. **Expertise.** Qualified business analysts who can create sophisticated models are hard to find, expensive to pay, and difficult to retain.
5. **Interoperability.** The process of creating and deploying predictive models traditionally involves accessing or moving data and models among multiple machines, operating platforms, and applications, which requires interoperable software.
6. **Pricing.** The price of most predictive analytic software and the hardware to run it on is beyond the reach of most midsize organizations or departments in large organizations.

Fortunately, these barriers are beginning to fall, thanks to advances in software, computing, and database technology.

Advances in Predictive Analytics Software

Analytical software has taken much of the labor, time, and guesswork out of creating sophisticated analytical models.

Today, you can purchase a single workbench that supports all six steps in the analytic development process.

Integrated Analytic Workbenches. Leading vendors of analytical software have introduced in the past several years robust analytic workbenches that pre-integrate a number of functions and tasks that analytic modelers previously completed by hand or with different tools. Today, modelers can purchase a single analytic development environment that supports all six steps in the analytic development process.

Market leaders SAS Institute and SPSS, Inc., offer the leading analytic workbenches today, followed by a host of second-tier vendors like Fair Isaac, Unica, Oracle, KXEN, Salford Systems, StatSoft, Insightful, Quadstone, Visual Numerics, and ThinkAnalytics. Most of the leading workbenches contain integrated tools that enable developers to create and manage project plans; explore and profile data sets; create, test, and validate models; and deploy and manage the models.

Graphical Modeling. One major advancement offered by these workbenches is their ability to graphically model the flow of information and tasks required to create and score analytic models. In the past, modelers had to hand-code these steps into SQL or a scripting program. "I can't develop models without the types of analytic tools available today since I don't have programming skills," says TN Marketing's Siegel. "Today, I can create one hundred little steps in a graphical workflow, configure each step, and then hit a button to make the program run. The tool builds the programming logic behind the scenes so I don't have to."

Graphical Modeling

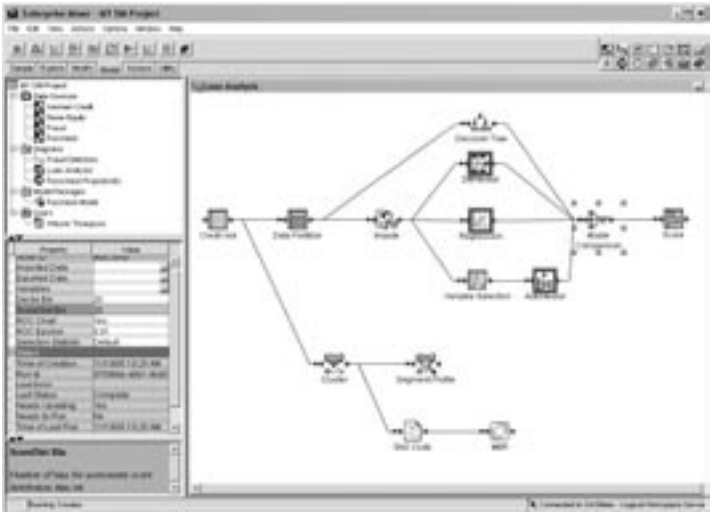


Figure 16. Predictive modelers create models using graphical workflows like this from SAS Institute's Enterprise Miner.

Automated Testing. Analytic workbenches have also improved developer productivity by automatically running multiple models and algorithms against a data set and measuring the impacts to see which provides the best performance. Previously, developers had to spend time testing each type of model and algorithm separately, effectively limiting the options they could test.

Client/Server. Today's analytic workbenches run in a client/server configuration rather than only on a desktop. A client/server architecture consolidates queries onto the server, reducing what analysts must download to their desktops to explore data and create analytic models. This reduces network traffic and redundant queries, which can bog down system performance.

Text Analytics. Predictive text analytics enables organizations to explore the "unstructured" information in text in much the same way that predictive analytics explores tabular or "structured" data. Through text analytics, organizations can uncover hidden patterns, relationships, and trends in text. As a result, companies gain greater insight from articles, reports, surveys, call center notes, e-mail, chat sessions, and other types of text documents. Predictive text analytics also allows organizations to combine structured and unstructured information in the same models or retrieve documents related to specific KPIs.

Most companies deploy analytical data marts that house most of the data that predictive modelers want to analyze.

Analytic Data Marts. Along with the client/server workbench, most organizations implement an analytical data mart to house much of the data that analysts want to analyze. Most organizations refresh these analytical data marts on a monthly basis so modelers can rerun models on new data. Having a dedicated environment for predictive modelers further offloads query processing from a central data warehouse and operational systems, and improves performance across the systems.

"Right now, our analysts each pull data sets from the data warehouse and create models on their desktops," says Dave Donkin, group executive of information management at Absa Bank, a $60 billion financial services company in Johannesburg, South Africa. "But we are moving to a group-based model where we will pull down the data and dimensions the analysts need to a central server

running a rich SAS Institute data environment. This will minimize the time they spend preparing the data and improve the performance of our data warehouse servers, which are running close to 100% capacity."

Database-Embedded Analytics

A second major development that addresses key elements in the "analytical bottleneck" described above is the availability of relational databases that embed nearly all predictive modeling functions. Teradata, a division of NCR, has led the charge by embedding numerous functions into its relational database, most of which are accessible via SQL extensions or PMML. IBM has a rich analytic feature set within DB2, while Sybase, Oracle, and Microsoft are catching up. Some of the functions that Teradata bundles into its database enable business analysts to:

- Profile and describe the data to reveal the quality and suitability of data.
- Transform, reformat, or derive columns.
- Reduce the amount of data required for analytic algorithms by applying correlation, covariance, and other algorithms.
- Restructure tables, create data partitions, and generate samples.
- Visualize model results to enhance understanding.
- Apply analytical algorithms, including linear and logistic regression, factor analysis, decision trees, and clustering algorithms.
- Store and score models directly inside the database.

"Database-enabled analytics is one of the biggest enablers of predictive analytics in the past several years."

It is not just the database vendors who recognize the value of in-database analysis. Some analytic workbenches, such as SPSS Clementine, automatically translate their graphical workflows into optimized SQL that goes to a relational database for execution, increasing scalability of common data mining tasks like scoring. When combined with client/server analytic workbenches, database-embedded analytics allows organizations to distribute analytic processing across client, server, and database tiers to optimize performance, scalability, and usability. Many organizations plan to take advantage of this flexibility to improve performance in the near future.

"Database-enabled analytics is one of the biggest enablers of predictive analytics in the past several years," says consultant Breur. "No matter what tools you use, you still have to spend a lot of time working with the data. As data warehouses provide better support for the types of scrubbing and transformations required to create analytic models, the easier your life becomes."

According to our survey, most organizations plan to significantly increase the analytic processing within a data warehouse database in the next three years, particularly for model building and scoring, which show 88% climbs. The amount of data preparation done in databases will only climb 36% in that time, but it will be done by almost two-thirds of all organizations (60%)—double the rate of companies planning to use the database to create or score analytical models. (See Figure 17.)

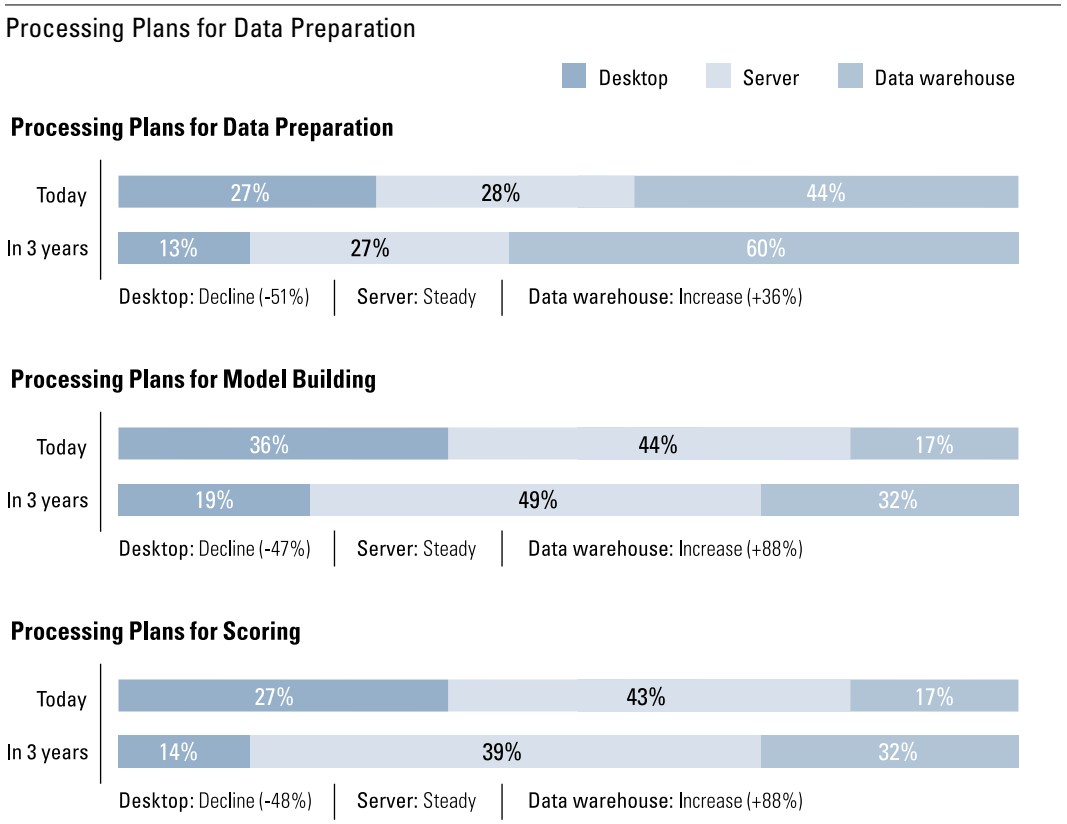


Figure 17. The three graphs are based on 162 responses from companies that have implemented predictive analytics.

The growth in database processing comes at the expense of desktop processing. Most databases can run on high-performance servers that can substantially accelerate performance across large data sets. Surprisingly, the amount of processing performed on an analytical server or workbench will remain constant in the next three years. Nearly half (49%) of all companies will create models on analytic servers, which underscores the power and flexibility of the new analytic workbenches.

Benefits of Relational Database Processing. If processing analytical functions occurs inside a database, users no longer have to extract, move, and load large datasets to a desktop or server machine. "It used to take us several days to score 50 million records, but now it takes half an hour," says one practitioner whose data warehouse runs on a 48-node Teradata machine. Plus, leading relational databases offer greater reliability, scalability, and fault tolerance than desktop machines.

"It used to take us several days to score 50 million records, but now it takes half an hour."

Given the advent of powerful analytic workbenches, it's surprising that about one-third of organizations plan to build analytical models in databases within three years. Most users I interviewed plan to continue creating models on server-based workbenches and then upload them to a database for scoring. However, it's likely that many organizations plan to use relational databases to perform some steps within the model creation process, specifically model testing and verification. This way, they can move the models to the database for testing and verification rather than move the data to the desktop or server.

"We leverage the data warehouse database when possible," says one analytics manager. He says most analysts download a data sample to their desktop and then upload it to the data warehouse once

it's completed. "Ultimately, however, everything will run in the data warehouse," the manager says. Most experts believe it's still best to create and analyze predictive models in analytic workbenches. Analytic servers provide a flexible environment designed for model building and analysis, unlike relational databases that are designed for transaction or query processing and require knowledge of SQL. Consequently, Teradata and IBM have recently de-emphasized the model building capabilities within their database management systems.

BI-Enabled Analytics and Applications

Although BI tools are not an ideal environment for creating sophisticated models with complex algorithms, users can import these types of models into a BI tool environment using PMML. In addition, some BI tools enable users to create regressions and cause/effect models, and display the results in easy-to-read reports. For example, MicroStrategy enables developers with some training to create linear and logistic models within its MicroStrategy 8 toolset and run reports for general business use that leverage the output of these models. (See Figure 18.) MicroStrategy also provides various charts and viewers that let users gauge and compare the accuracy of various models and apply the results.

Analytics Embedded in a Sales Report

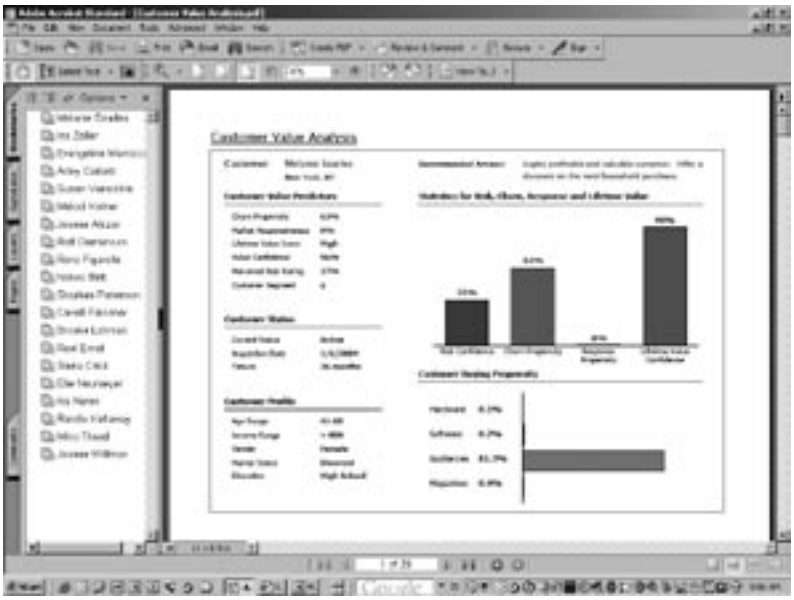


Figure 18. This screenshot shows a MicroStrategy report that leverages various models (churn, risk, response propensity) to analyze customer behavior and recommend actions to salespeople.

Other BI vendors that embed predictive models include Oracle's Siebel Analytics reporting and analysis tool, Business Objects (which uses a separate set analyzer module for creating segmentations), and Advizor Solutions, a visualization vendor that embeds predictive algorithms from KXEN.

Analytic Applications. There is sometimes a fine line between BI tools and analytic applications. In general, BI tools can create reports that support any department or domain. In contrast, analytic applications are "solutions" that generally consist of a set of predefined reports that enable business people to manage a variety of integrated processes within a single domain, such as sales, and sometimes in a single industry, such as sales for consumer packaged goods companies. A sales analytic application, for example, might provide a series of reports for users to track pipeline, sales rep performance, and customer activity.

Many BI and application vendors, including most sponsors of this report, are delivering packaged analytic applications that embed predictive models. OutlookSoft, for example, has embedded root-cause and variance analysis algorithms in its performance management application (e.g., integrated planning, budgeting, consolidation, dashboarding, and scorecarding) and now markets its predictive capabilities as a competitive differentiator. (See Figure 19.)

Analytics Embedded in a Performance Management Dashboard

Figure 19. The top-middle portion of this OutlookSoft screenshot shows the status of KPIs in the finance perspective, including actuals, budget, and predicted values. Highlighting a KPI (e.g., inventory turns) changes the table and gauge below to show the reasons (i.e., dimensional categories) and root causes (i.e., transactions) of the budget variance along with relevant text documents, all of which are generated using predictive algorithms. By moving the slider at left, a user can select new inputs to all predictor variables to create scenarios based on different risk levels.

These analytic applications are easy to use because they don't require users to create analytic models. An expert at the vendor company creates the models and the vendor embeds them into reports or code so results are generated automatically. Users only need to know how to interpret the results.

Embedded models must be generic enough to handle wide variations in customer data.

However, this ease of use comes at a price. The embedded models must be generic enough to handle wide variations in customer data. But generic models are generally less accurate (and may be misleading) than models tuned to individual data sets. To get around this problem, packaged analytic vendors typically narrow the domain of each application so the models work on a limited set of data. Also, they pre-configure the models for each type of industry or domain in which the

application will be used. Third, they require users to tune the models by selecting parameters and configuring other controls to boost model accuracy.

Support Vector Machines. Some vendors are experimenting with support vector machine algorithms, which seek to automatically select an optimal set of variables, transform data into the proper format, and reduce the size of the data to improve model-building performance. A primary goal of support vector machines is to create models that apply to a wide range of data beyond what was used to build the model.

"The algorithms do a nice job of automated variable selection and model building," says Herb Edelstein, president of Two Crows, Inc., a predictive analytics research and consulting firm. "Are they as good as hand-tuned models with appropriately transformed variables? No, but that's not their goal. Their goal is to allow an unsophisticated user to easily come up with a reasonable answer. As a result, these analytical models are good for embedding within packaged analytic applications."

Industry Standards for Interoperability

The bugaboo of predictive analytics is interoperability. A business analyst creates a predictive model using his tool of choice but cannot share it with other analysts who prefer to use other analytic workbenches. Or the business analyst wants to use the model to score customer records in a relational database but the modeling tool generates the model only in a language that is proprietary or is not supported by the organization's database platform (e.g., Cobol, C, Java, SQL).

PMML. In 2002, a loose affiliation of analytics vendors produced version 1.0 of the Predictive Model Markup Language (PMML) to overcome these problems. PMML is an XML schema for describing statistical and analytical models so they can be shared among applications. PMML describes the inputs to predictive models, the transformations to prepare data, and the parameters that define the models. Now produced by the Data Mining Group (www.dmg.org) with its 20 full and associate members (most of whom are analytics vendors), PMML is beginning to be used by a number of vendor and user organizations. Version 3.2 provides rich support of the most common algorithms and transformations involved in building predictive models.

"PMML is a good idea but it's only good if a majority of vendors adhere to the same standard," says Breur. "Right now, PMML has only halfway caught on."

PMML is gaining momentum as an industry standard.

The problem is that PMML currently translates only a limited set of predictive models. In addition, many vendors create proprietary extensions to PMML to support their unique features, diluting the advantages of using a standard. And some don't support PMML across all their product lines. However, the DMG has done a good job of baking these proprietary extensions into the standard with each subsequent version, and PMML is gaining momentum, not losing it, like other standards.

Interoperability Strategy. As a result, some vendors are making PMML a pivotal part of their analytics strategy. MicroStrategy, for example, is a "PMML consumer," meaning it can store PMML models created by other analytic workbenches in its XML metadata repository and apply them to data in its reports. This will allow MicroStrategy customers to apply more complex predictive models to reports than they can build within MicroStrategy 8. Thus, PMML might be the vehicle that makes it possible to distribute the results of predictive analytics to more users. Report designers can embed the algorithms in reports and give users prompts to select data via one or more predefined queries, ensuring that users select appropriate data to run against the algorithm.

Most database vendors can consume PMML models, which makes it easy for developers to upload their models to a database for scoring. Without PMML, developers must code their models in SQL or some other language, then test and debug the code, and get permission from database administrators to upload and run it. Because it's XML-based, PMML skirts many of these time-consuming tasks and is more readily accepted by database administrators.

Bridging the Analytics-IT Gulf. From a user perspective, PMML might help overcome the technical and cultural differences between predictive analytics teams and their IT counterparts. Although both groups are highly technical, they approach problems differently, which can lead to frustration and stalemate.

PMML might help overcome technical and cultural differences between predictive modelers and their IT counterparts.

"We spent 18 months creating, testing, and validating a new way of segmenting customers for our salespeople, but the project is stalled because we can't get the IT department to incorporate it into the production data warehousing environment," says a frustrated director of business analytics at a major pharmaceutical firm. "The amazing thing is that we provided them with the actual [Microsoft Visual Basic] code, which they'll need to translate into SQL and ETL, but they still want to see requirement documents and functional specifications."

To break this logjam, the analytics team could use an analytic workbench that generates PMML code rather than Microsoft Visual Basic. They could then upload the resulting PMML code to the data warehouse database and work with the data warehousing team to schedule it and create the data sets and models needed to drive new sales reports. Meanwhile, the analytics team could export the model results to Microsoft Access or Excel to give BI developers a head start on creating more sophisticated reports. IT folks are generally receptive to industry standards and XML. Nonetheless, the analytics team will need to take its place in the queue for new projects, especially if the models require the data warehousing team to source new data.

The only other alternative is for the analytics team to create its own specialized data mart(s) and generate both models and reports. Of course, this approach brings considerable overhead and redundancy, and most analytics teams do not want to become report developers or report distributors. Despite these drawbacks, consultant Breur says, "I've seen such setups work well at some of my clients who have centralized departments."

Recommendations

Now that we've defined predictive analytics, assessed its business value, and stepped through key trends and processes, it's important to provide specific recommendations to BI managers and business sponsors about how to implement a predictive analytics practice. This section offers five recommendations that synthesize best practices from various organizations that have implemented predictive analytics.

1. Hire business-savvy analysts to create models

Every interviewee for this report said the key to implementing a successful predictive analytics practice is to hire analytic modelers with deep understanding of the business, its core processes, and the data that drives those processes.

"Good analysts need strong knowledge of narrow business processes."

"Good analysts need strong knowledge of narrow business processes," says Higdon of Sedgwick CMS. In claims processing, Higdon says good analysts "understand the business of claims handling; the interplay of variables across claim, claimant, and program characteristics; and what data they can and cannot rely on." Only then, he adds, can analysts create "meaningful models" that result in positive business outcomes.

Three Characteristics. Analytic modelers do not need deep statistical knowledge. "While a PhD in statistics can't hurt, it isn't the key," says Breur. "Many modelers shoot themselves in the foot because they fall in love with their algorithms, not business outcomes." Breur says effective analytic modelers exhibit three overriding characteristics: (1) they understand the business and can translate models into bottom-line outcomes, (2) they have a strong data background, and (3) they are diehard experimentalists.

Breur, who has run analytics departments in large companies and is now an independent consultant, says he has had success hiring social scientists with quantitative backgrounds. He also says the best data miners have traditionally come from the United States, where most of the research and commercial implementations have taken place in the past.

Hire from Within. Higdon says he largely grooms analysts internally after having little luck hiring people from the outside. He looks for people who've worked very closely with clients to create, deliver, and measure the impact of claims processing services and who have some computer experience, such as knowledge of Excel, Access, and an OLAP tool. "It's difficult to find someone with industry knowledge, analytics skills, and the right software experience. Two out of the three is usually the best you can find." On his current team, one member has experience in statistics, reporting, and data warehousing; two have backgrounds in Six Sigma; and four were junior data analysts and report developers who supported programs for large clients or partners.

Getting a new modeler up to speed takes three months or longer if they are right out of college or have minimal experience.

Apprenticeships. Higdon also says he "apprentices" new team members to ease them into the position. After a week of orientation and tool training, he assigns the new hire an experienced analyst as a mentor. New hires are given sample old projects to work on to learn the tool, the processes, and the data without any risk, and then they step up to small live projects. "The process takes a good three months; longer if they are right out of college or have minimal experience," Higdon says.

Advances in analytical modeling tools during the past decade have made the field more accessible to individuals without deep statistical training. Several of the people I talked to were largely self-taught. For example, one predictive modeler started out as a database administrator. After helping guide the implementation of the company's enterprise data warehouse, the person proposed to management to "try a little data mining." The person then purchased an analytics package, took a week of training, hired a consultant to get going, and demonstrated business returns. Today, the company has several analytical modelers that generate 30 models a month, saving the company millions of dollars in marketing and other costs.

2. Nurture a rewarding environment to retain analytic modelers

Since good analytic modelers are difficult to find, it's imperative that managers create a challenging and rewarding environment. Of course, money is requisite. Top-flight analytic modelers often command higher salaries than classic business or data analysts. But good analytic modelers are motivated by things other than money and status, says Breur.

"You don't attract analytic modelers with the same incentives as other people," he says. "They want an opportunity to demonstrate their skills and learn new things so you have to increase their training budgets. I've struggled with human resources departments on this issue."

"You don't attract analytic modelers with the same incentives as other people."

Absa Bank's Donkin oversees a team of 30 analytic modelers. "We've built up a substantial team over the years, but we've battled to get them and it is challenging to retain them."

Offer New Challenges. Donkin says the best way to retain analysts is to provide a stimulating work environment and new opportunities to exercise their skills and demonstrate their talent. Donkin's group provides information services to the entire company, and as a result, enables analysts to work on a variety of business problems, such as anti-money-laundering, lead generation, retention, credit scoring, fraud detection, cross-selling, operational research, and in the future, human resources, procurement, and other departments. "We can provide analysts with a wealth of opportunities and experience, which is one of the advantages of having a central IM group," he says.

Career Paths. It's also important to provide analytical modelers a clear career path to keep them from leaving for greener pastures. This rarely happens since analysts straddle the line between business and technology and are taken for granted by both sides. One strategy is to work with human resources to define two career paths for analytic specialists: (1) A technical track that eventually leads to the company's research and development organization and cutting-edge work in the laboratory that creates next-generation products, and (2) a business track that moves a highly business-driven individual into a management position in the marketing department. The practice of marketing is increasingly adopting quantitative methods and these individuals will be in high demand in the future.

3. Fold predictive analytics into the information management team

The Inside Track. Traditionally, analytics teams are sequestered away in a back room somewhere and report to an ambitious department head (usually sales or marketing) who is seeking to ratchet up sales and get an edge on the competition. Unfortunately, this approach is not optimal, according to most practitioners. Analytic modelers are voracious consumers of data, and must establish strong alliances with the data warehousing team to maintain access to the data.

Analytic modelers are voracious consumers of data and must establish strong alliances with the DW team.

"Since I work in the data warehousing department within IT, I go to my colleagues and say, 'I need this,' and I usually get it. I get to the head of the queue faster than someone from outside, which means I get more [storage] space, quicker access to data, and more privileges. As a result, my projects get pushed faster. Those are the unwritten rules," says an analytical modeler.

She continues, "It is much harder for someone from marketing to get access to data. They have to fill in forms, justify their request, and other things." Since modeling is an iterative process, modelers often find they need additional data in the middle of a project. But analytical modelers who don't have close ties to the data team often must weigh the benefits of requesting new data against the delays in obtaining it. Not surprisingly, "it takes marketing much longer to create predictive models," she says.

Options for Organizing the Analytics Team. Breur lists three ways companies can organize an analytics group: (1) embed analysts within a business department, (2) outsource the work, or (3) create a centralized group that serves the entire company. The last option is the only one that works, Breur says. "You really need a centralized department where specialists can rub elbows."

This is exactly what Absa Bank has done. Its 30-person analytic intelligence team is one of five groups within the information management group that provides data warehousing, business intelligence, knowledge management, analytic services, and geographic information systems to the enterprise. Donkin runs the group and reports to the bank's chief operating officer. "If we reported into finance or marketing, we'd have a bias towards those areas and overlook important initiatives. This way, we're neutral and that is really important."

4. Leverage the data warehouse to prepare and score the data

Once you've hired the people and established the organization, the next important task is to provide a comprehensive solution for managing the data that the analytic modelers may want to use. While it is not necessary to build a data warehouse to support the analytic process, a data warehouse can make the process infinitely easier and faster.

Saving Time. A data warehouse pulls together all relevant information about one or more domains (e.g., customers, products, suppliers) from multiple operational systems and then integrates and standardizes this information so it can be queried and analyzed. With a data warehouse, analysts only have to query one source instead of multiple sources to get the data they need to build models.

A data warehouse can liberate analysts from data drudgery and greatly accelerate the model building.

In addition, a data warehouse loads, cleans, integrates, and formats data, sparing analytical modelers precious weeks and months spent on these data management tasks. As a result, a data warehouse can liberate analysts from data drudgery and greatly accelerate model creation. A data warehouse can also format the data for analytical modelers. For example, analysts typically need to flatten data into a single table with dozens or hundreds of columns; they often need to aggregate or dis-aggregate records depending on the needs of the algorithm or create new derived fields. Or they may need to import external or syndicated data into the data warehouse for analysis. All these steps can be done automatically as part of a data warehousing process.

For example, every month, the data warehousing team at a major U.S. telecommunications company builds specialized tables or data marts within its enterprise data warehouse for the predictive modeling team. "Once you have your data at the beginning of the month, everything runs smoothly. You don't have data quality problems or extra data preparation to do," says the modeler.

Analytic Data Marts. The specialized requirements of analytical modelers lead most companies to spawn a specialized data mart from the data warehouse. "The best way to go is to provide a dedicated data mart," says Breur. In a Teradata-based data warehouse, an analytical data mart consists of a set of tables within the database. In most other environments, however, the data mart runs on a separate database, and sometimes on a different physical server. (See Figure 20.)

Predictive Analytics Architecture

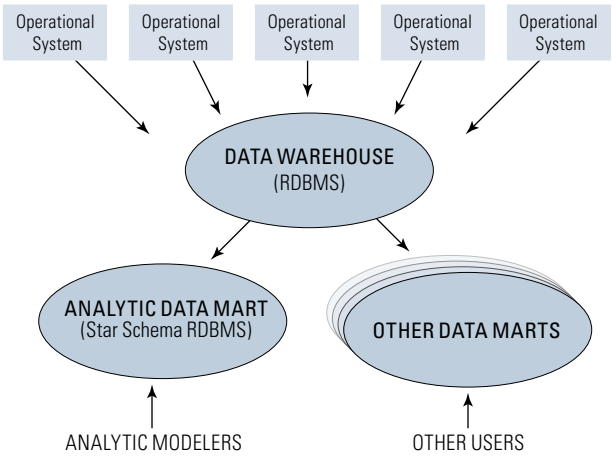


Figure 20. Most organizations create a separate analytic data mart from a central data warehouse to support the specialized queries and data processing required to build analytic models.

One sizable advantage in creating an analytical data mart is performance. Analytical modelers tend to submit complex, expensive queries against large data volumes. This can slow down queries for other users on a shared system, unless the analytical processing is limited to evening hours, which obviously isn't convenient for analysts and is not conducive for the iterative analysis critical to creating good models.

A client/server workbench running against a dedicated analytic data mart is the de facto architecture for supporting analytics projects. It provides a three-layer architecture that allocates different types of processing to the optimal layer, minimizes network traffic, and offloads queries from mission-critical analytical and operational systems, improving query performance for all.

5. Build awareness and confidence in the technology

One of the toughest challenges in implementing analytics is convincing the business that this mathematical wizardry actually works. "Building confidence is a big challenge," says one analytics manager. "It takes awhile before business people become confident enough in the models to apply the results. The ultimate litmus test is when business people are willing to embed models within operational processes and systems. That's when analytics goes mainstream in an organization."

"It takes awhile before business people become confident enough in the models to apply the results."

But getting to a lights-out analytical environment is not easy. Most organizations, even those with large analytic staffs like Absa Bank, still only apply analytics in pockets and have yet to truly operationalize the results. "Our business is increasingly becoming aware of what's possible with analytics, but we still battle there," says Donkin.

How to Spread the Word. Donkin employs several techniques to increase awareness. His group has hired a cartoonist to create an eye-catching visual representation of what it knows about customers and its business significance. "A picture is worth a thousand words," says Donkin. The group also sends business users notifications via Internet banner ads and other vehicles that contain interesting tidbits, such as "Did you know there are 20,000 checking account customers who have home

loans with another bank?" Says Donkin, "We want to stimulate thinking within the business and highlight the data we have."

Internal Sales People. Finally, Donkin's group has a dozen business development managers whose sole job is to interface with the business and explain what is possible from a technical perspective. Each business development manager is assigned to one or more lines of business. They meet regularly with business representatives to inform them how to leverage data and resources to support business strategy, address problems and opportunities, and collect requirements. "They need to know the business and information management—it's a tough job—but some now sit on the executive committee of the business unit they serve, which is a sign of respect and that the business is ready to leverage our resources fully," says Donkin.

Conclusion

Predictive analytics reinstates "gut feel" into corporate decision making.

Applying these five recommendations should enable any organization to implement predictive analytics with a good measure of success. While many people seem intimidated by predictive analytics because of its use of advanced mathematics and statistics, the technology and tools available today make it feasible for most organizations to reap value from predictive analytics.

In many respects, predictive analytics is something we all do intuitively. Many of us have "gut feelings" about people or situations, and often these gut feelings turn out to be uncannily accurate. Malcolm Gladwell's bestselling book *Blink* (Little Brown & Co., 2005) provides many examples of how our subconscious mind collects events, analyzes patterns, and predicts the future, often reflexively in ways that defy our conscious knowledge.

For example, Vic Braden, one of the world's top tennis coaches, has an uncanny ability to predict when a professional tennis player is about to double fault. Braden doesn't know how he does this, but he wants to find out. He is currently videotaping tennis players and analyzing the film in super slow motion to see if he can identify the traits or characteristics of a double fault that his mind sees and processes subconsciously.

It is not an exaggeration to say that Braden is an analytical modeler; he simply uses a different set of tools than do modelers who work on corporate data sets. In this context, predictive analytics is nothing more than "slowing down the tape" and dissecting events one at a time to find the key characteristics that have the most predictive power. Ironically, while predictive analytics leverages highly cerebral disciplines of statistics and mathematics, it enables our organizations to respond more intuitively and instinctively to customers and business events. In an odd way, predictive analytics reinstates "gut feel" in corporate decision making on an enterprise scale.

SAS
100 SAS Campus Drive
Cary, NC 27513-2414
919.677.8000

www.sas.com

SAS gives you the power to integrate data across an enterprise—THE POWER TO KNOW®—and to quickly transform that data into a competitive advantage. As the leader in business intelligence software and services, SAS reduces uncertainty, predicts with precision, and optimizes performance. Our comprehensive suite of analytics includes solutions for statistics, data and text mining, model management, forecasting, econometrics, quality improvement, and operations research.

SAS celebrated its 30th anniversary with an unbroken track record of revenue growth and profitability, and worldwide revenue of $1.68 billion. Customers at 40,000 sites use SAS® software to improve performance through insight into vast amounts of data, resulting in faster, more accurate business decisions; more profitable relationships with customers and suppliers; compliance with governmental regulations; research breakthroughs; and better products. Only SAS offers leading data integration, intelligence storage, analytics, and traditional business intelligence applications within a comprehensive enterprise intelligence platform.

As this TDWI report indicates, analytics will play an increasingly important role, so it is critical that organizations reduce the likelihood of erroneous model output or incorrect interpretation of model results. SAS provides the tools and solutions to execute every step of the predictive modeling process described in the TDWI report:

1. **Project definition**
2. **Exploration**
3. **Data preparation**
4. **Model building**
5. **Deployment**
6. **Model management**

Specific predictive analytics products include:

- **SAS Forecast Server** can automatically generate large quantities of statistically based forecasts without the need for human intervention—unless so desired. It automatically chooses the best forecasting model, optimizes the model parameters, and produces the forecasts. An easy-to-use graphical user interface lets users manually build or adjust forecasting models, or update forecasts.
- **SAS Enterprise Miner**™ is a powerful, complete data mining solution with unparalleled model development alternatives and extensive integration opportunities. It is designed for data miners, marketing analysts, database marketers, risk analysts, fraud investigators, engineers, and scientists who need to solve critical business or research issues.
- **SAS Text Miner** provides a rich suite of tools for discovering and extracting knowledge from text documents. It transforms unstructured data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between documents, clustering into categories, and incorporating text with structured data to enrich predictive modeling endeavors.
- **SAS Model Manager**, with its secure model repository and rich metadata structure and project templates, streamlines the tedious and often error-prone steps of creating, managing, and deploying analytical models. As predictive models are deployed, performance metrics verify compliance with regulatory requirements such as the Sarbanes-Oxley Act and the Basel II accord.

SAS is pleased to sponsor this TDWI report on predictive analytics. We invite you to visit us at www.sas.com to learn more about our products, our educational offerings, and the award-winning technical support services that come with our software.

TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

1201 Monster Road SW
Suite 250
Renton, WA 98057

T 425.277.9126
F 425.687.2842
E info@tdwi.org

www.tdwi.org