Oracle9*i* Data Mining

An Oracle White Paper
June 2001

ORACLE®

Oracle9*i* Data Mining

BENEFITS AND USES OF DATA MINING

In today's competitive marketplace, it is critical that companies manage their most valuable assets — their customers and the information they know about their customers. That's where data mining can help. Data mining can sift through massive amounts of data and find hidden information — valuable information that can help you better understand your customers and anticipate their behavior. Oracle9*i* Data Mining™ software helps you build applications to uncover this hidden information about your customers. Armed with this information, you can build a close relationship with and understand your customers, which helps you to:

- Better retain customers and avoid churn
- Profile customers and understand behavior
- Maintain and improve profit margins
- Reduce customer acquisition costs
- Target profitable customers with the right offer

Oracle9*i* Data Mining allows you to build applications that go beyond traditional reporting tools — that tell what has already happened in the past —by analyzing the past and using it as a predictor of the future. By analyzing the profiles of your best customers, Oracle9*i* Data Mining allows you to build applications that identify other customers who may not represent your most valuable customers today but fit the profiles of those who have become your best customers. Knowing the "strategic value" of customers — who should be profitable in the future and who is not likely to be profitable in the future, or predicting which customers are likely to churn or likely to respond to a marketing offer — and integrating this information into your operations are key to managing your business proactively.

WHAT IS DATA MINING?

Now let's describe what data mining is and how it both differs from and complements other business intelligence products — query and reporting, OLAP, and statistical tools. Let's also look at some common definitions of business intelligence tools.

"Simply put, data mining is used to discover [hidden] patterns and relationships in your data in order to help you make better business decisions."

-- Herb Edelstein, Two Crows Corporation

Query and reporting tools help you to get information out of your database or data warehouse. These tools are good at answering questions such as "Who purchased a mutual fund in the last 3 years?"

OLAP tools go beyond this and allow users to interactively "drill down" into the data for summaries, comparisons, analysis, and forecasts. OLAP is good at drill-downs into the details to find, for example, "What is the average income of mutual fund buyers by year by region?"

Statistical tools are used to draw conclusions from representative samples taken from larger amounts of data. Statistical tools are useful for finding patterns and correlations in "small to medium" amounts of data but fall short when the amount of data begins to overwhelm the tool. Typically, when you deal with greater than, say, 25 input variables and tens of thousands of records, traditional statistical regression techniques struggle. Because statistical tools can't analyze all the data, they force data analysts to use representative samples of the data and to eliminate input variables from the analysis. By throwing out variables and using samples of the data, you are throwing away "information."

Query and reporting, OLAP, and statistical tools are good at allowing the user to drill down and understand what has happened in the past. With both tools, if you know what you are looking for, have a good analytical tool, are a good analyst, and have a lot of time, you can eventually find the answer.

Data mining doesn't have these limitations. Data mining goes deep into the data. Data mining uses machine-learning algorithms to automatically sift through each record and variable to uncover patterns and information that may have been hidden. Data mining goes beyond traditional business intelligence tools and analyzes the details of past (e.g. whether an individual purchases item "A"). Data mining builds models and uses the models as predictors of the details in the future (e.g. the likelihood that a customer will purchase "A" in the future). Data mining is good at providing detailed insights and making individual predictions, such as "Who is likely to buy a mutual fund in the next six months and why?"

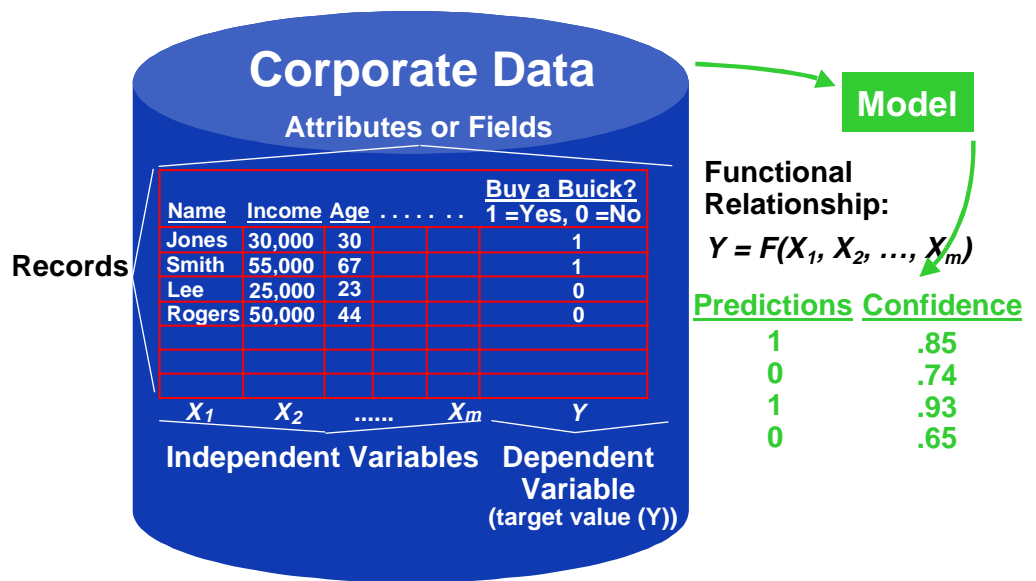| **Query and Reporting**	**OLAP**	**Data Mining**
Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
"Information"	"Analysis"	"Insight and Prediction"
Who purchased mutual funds in the last 3 years?	**What is the average income** of mutual fund buyers by region by year?	**Who will buy** a mutual fund in the next 6 months and why?

DATA MINING CONCEPTS

Data mining can be separated into "supervised learning" and "unsupervised learning" techniques. Supervised learning requires the data analyst to identify a target field or dependent variable. The supervised-learning technique then sifts through data trying to find patterns and relationships between the independent variables and the dependent variable. (Oracle9*i* Data Mining provides the Naïve Bayes data mining algorithm for supervised-learning types of problems.)

In "unsupervised learning," the user does not indicate the objective to the data mining algorithm. Associations and clustering algorithms make no assumptions about the target field. Instead, they allow the data mining algorithm to find associations and clusters in the data independent of any
a priori defined business objective. (Oracle9*i* Data Mining provides the Assoociation Rules data mining algorithm for unsupervised-learning problems.)

Using the Past to Predict the Future

In the supervised learning example below, a dependent variable is defined to describe which customers purchased a new car; a "1" for "yes" and a "0" for "no." The data mining algorithm sifts through the data finding hidden patterns and intuits a "model" that best describes the functional relationships. Typically the data is separated into three equal parts — for training, testing, and evaluation. The initial model is built using the "training" sample of the data. Then the model is applied to the "testing" sample of the data to refine the model. Finally, the model is applied to the evaluation sample of data that is used to make predictions. Because we already know the outcome — who purchased a car and who didn't — we can "evaluate" the model's accuracy. Models with acceptable prediction capability often have high economic value.

Corporate Data
Attributes or Fields

Name	Income	Age	. . . . . .	Buy a Buick? 1 =Yes, 0 =No
Jones	30,000	30		1
Smith	55,000	67		1
Lee	25,000	23		0
Rogers	50,000	44		0

Records

X_1 X_2 ...... X_m Y

Independent Variables **Dependent Variable** (target value (Y))

Model

Functional Relationship:

$$Y = F(X_1, X_2, …, X_m)$$

Predictions	Confidence
1	.85
0	.74
1	.93
0	.65

Unsupervised learning (association and clustering) models usually are applied only to the historical data to reveal clusters and associations, regardless of any specific business application or purpose. Sometimes, association and clustering models can be used to make predictions, as in Oracle9*i*AS Personalization, which makes "recommendations" for both registered customers and anonymous Web visitors. The recommendations are other "items" (e.g. products within an e-tailer's catalog of products) that are *associated* with other items (demographics, purchases, "clicks," and ratings data) for that customer or visitor.

ORACLE9*i* DATA MINING: AN OVERVIEW

Oracle9*i* Data Mining allows companies to build advanced business intelligence applications that mine corporate databases, discover new insights, and integrate that information into business applications. Oracle9*i* Data Mining is a priced option to the Oracle9*i* Database Enterprise Edition (EE) that embeds data mining functionality for making classifications, predictions, and associations. All model-building and scoring functions are accessible through a Java-based API.

Oracle9*i* Data Mining helps companies build business intelligence applications that find meaningful patterns and associations in corporate data — patterns that help you better understand and predict customer behavior. With it, companies can forge strategies to:

- Prevent customer attrition
- Cross-sell to existing customers
- Acquire new customers
- Detect fraud
- Identify most profitable customers
- Profile customers with greater accuracy

With Oracle9*i* Data Mining, companies can tap information hidden in their corporate databases to reveal new insights about their customers and their businesses. At every stage of the customer life cycle, Oracle9*i* Data Mining delivers value — that goes straight to your bottom line.

Oracle9*i* Data Mining can also find patterns hidden in scientific, government, manufacturing, medical, and other types of data. Applications of data mining in these areas include:

- Predicting the quality of a manufactured part
- Finding associations between patients, drugs and outcomes
- Identifying possible network intrusions

Insights discovered by Oracle9*i* Data Mining can be revealing, significant, and valuable.

DATA MINING EMBEDDED IN ORACLE9*i* DATABASE

Oracle9*i* Data Mining simplifies the process of extracting business intelligence from large amounts of data. It eliminates off-loading vast quantities of data to external special-purpose analytic servers for data mining and scoring. With Oracle9*i* Data Mining, all the data mining functionality is embedded in Oracle9*i* Database, so the data, data preparation, model building, and model scoring activities remain in the database.

Because Oracle9*i* Data Mining performs all phases of data mining within the database, each data mining phase results in significant improvements in productivity, automation, and integration. Significant productivity enhancements are achieved by eliminating the extraction of data from the database to special-purpose data mining tools and the importing of the data mining results back into the database. These improvements are notable in data preparation, which often can constitute as much as 80% of the data mining process. With Oracle9*i* Data Mining, all the data preparation can be performed using standard SQL manipulation and data mining utilities within Oracle9*i* Data Mining.

"Scoring" to make predictions is greatly simplified now, as the data and the model are in the same location — the database. There is no need to move the data to the model nor export the model as code (typically C, C++, or Java) and then move the model back to the data for scoring in the database. With Oracle9*i* Data Mining, everything occurs within the database.

Benefits of "in-database mining" become apparent in the first steps of a data mining project. Data preparation includes the analysis and assembly of relevant data, definition of a target field, and possibly creation of new summary or calculated fields. Since mining will occur in the database, any tools or applications can contribute to this process. It is simple to use lists and measures from existing business applications as the starting point for data mining.

Oracle9*i* Data Mining accepts both transactional and non-transactional (summarized, single-record) tables. Internally, Oracle9*i* Data Mining uses transactional tables and automatically handles any conversion necessary.

Non-transactional, summarized, single-record data format

caseId	Age	Income	Gender
100	20	1000	M
200	30	2000	F

Transactional data format

caseId	Attribute	Value
100	Age	20
100	Income	1000
100	Gender	M
200	Age	30
200	Income	2000
200	Gender	F

Oracle9*i* Data Mining supports binning, the process of collecting the values in a field into meaningful groups. A utility allows programmers to bin the data prior to building models for more problem-specific purposes. For example, for a driving safety problem, the age field might be binned into the ranges 0-16, 17-21, and greater than 21. For a life-insurance problem, the same age field might be binned into the ranges 0 to 35, 35-55, and greater than 55.

ORACLE9*i* DATA MINING ALGORITHMS

Oracle9*i* Data Mining provides programmatic access to two data mining algorithms embedded in Oracle9*i* Database through a Java-based API. Data mining algorithms are machine-learning techniques for analyzing data for specific categories of problems. Different algorithms are good at different types of analysis. Oracle9*i* Data Mining provides two algorithms: 1) Naïve Bayes for Classifications and Predictions and 2) Association Rules for finding patterns of co-occurring events. Together, they cover a broad range of business problems.

Classification and Prediction

Naïve Bayes is a supervised-learning technique for classification and prediction that builds models that predict the probability of specific outcomes. Naïve Bayes uses historical data to find patterns and relationships and then makes predictions. Oracle9*i* Data Mining enables application developers to build data mining models to classify and predict for a variety of purposes, such as:

- Identifying which customers are likely to purchase a certain product
- Identifying customers who are likely to churn
- Predicting the likelihood that a part will be defective

Customer ID	Response	Probability
4567332	YES	0.92
5463433	YES	0.92
1123444	YES	0.85
9565333	YES	0.85
2346557	YES	0.72
2257673	YES	0.55

Oracle9*i* Data Mining's Naïve Bayes algorithm can predict binary or multi-class outcomes. In binary problems, each record either will or will not exhibit the modeled behavior. For example, a model could be built to predict whether a customer will churn or remain loyal. Naïve Bayes can also make predictions for multi-class problems where there are several possible outcomes. For example, a model could be built to predict which class of service will be preferred by each prospect.
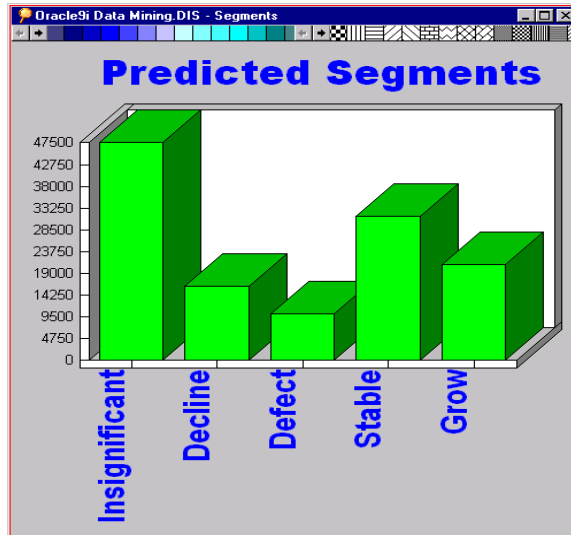
Binary model example

Q: Is this customer likely to become a high-profit customer?

A: Yes, with 85% probability

Multi-class model example

Q: Which one of five customer segments is this customer most likely to fit into — Grow, Stable, Defect, Decline, or Insignificant?

A: Stable, with 55% probability

Oracle9*i* Data Mining can identify customers who are likely to churn so companies can examine these customers' common traits and develop detailed profiles using other business intelligence tools, such as Oracle9*i*AS Discoverer as shown above.

Binary and multi-class predictions together cover a huge range of important business problems, including campaign response, sales offers, fraud detection, profitability prediction, customer profiling, credit rating, churn anticipation, inventory requirements, failure anticipation, and many others.

Oracle9*i* Data Mining also provides utilities for evaluating models in terms of model accuracy and "lift" — or the incremental advantage of the predictive model over the naïve guess.
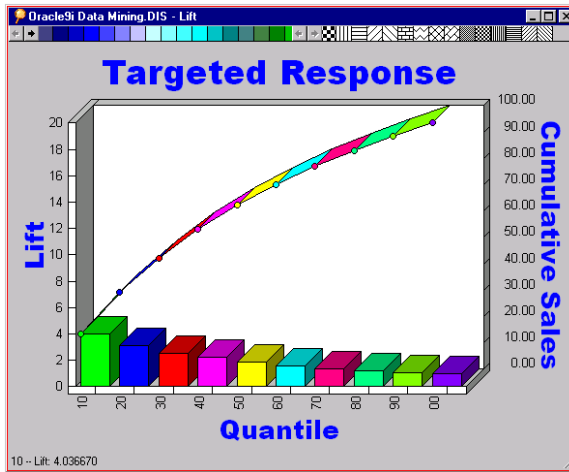
Oracle9*i* Data Mining's models can be evaluated to measure their "lift," or incremental improvement over the naïve guess, as shown here using Oracle9*i*AS Discoverer.

Finding Associations

Association rules detect "associated" or co-occurring events hidden in databases. Association analysis, or unsupervised learning, is often used to find popular bundles (e.g. market basket analysis) of products that are related for customers, such as "milk" and "cereal" being associated with "bananas." Oracle9*i* Data Mining's Associations Rules can be used to identify co-occurring items or events in a variety of business problems, such as:

- Which manufactured parts and equipment settings are associated with failure events?
- Which patient and drug attributes are associated with which outcomes?
- Which items or products is this person most likely to buy or like?

Oracle9*i* Data Mining's association rules are stored in Oracle9*i* Database and may be examined using other query and reporting tools, such as Oracle9*i*AS Discoverer as shown here.

The Association Rules algorithm generates a set of antecedent and consequent pairs in the form of A implies B with a probability of *n*%. This modeling technique allows users to discover associations between items or events. Oracle9*i* Data Mining includes utilities for examining and reporting the association rules. The associations or "rules" thus discovered are useful in designing special promotions, product bundles, and store displays.

Model Scoring

Once an Oracle9*i* Data Mining Naïve Bayes model has been built, new data records can be scored. Scoring is the process of predicting outcomes. In Oracle9*i* Data Mining, the data mining algorithm is taken to the data, rather than taking the data to the algorithm as in most traditional ad hoc data mining tools. This is a more efficient architectural strategy when dealing with large amounts of data.

Oracle9*i* Data Mining can score in batch mode or on demand. Batch scoring takes a table of records as input and stores the predictions in another table. On-demand scoring takes a single record and returns a prediction that can be fed directly into the "calling" application.

ENHANCING APPLICATIONS WITH DATA MINING

Oracle9*i* Data Mining can automate the extraction and integration of new insight and predictions into a variety of business applications, including call centers, Web sites, campaign management systems, ATMs, enterprise resource management (ERM), and other operational and business planning applications. Rather than having special departments of advanced data analysts who work on ad hoc data mining projects, the true value of data mining is realized when the new insights and predictions are integrated and "operationalized" into existing business applications.

Automation of data mining tasks is facilitated by Oracle9*i* Data Mining's Java-based API. Application programmers can control all aspects of data mining — they can expose complex settings for advanced users or completely automate the process for business users. Programmatic control extends from data preparation and model building to on-demand scoring of single records and batch scoring of large data sets. Batch scores may be stored in relational tables for access by other business applications (e.g. call centers or marketing campaign systems) or called "'on-demand" in interactive applications where new information is collected that must be factored into the predictive model.
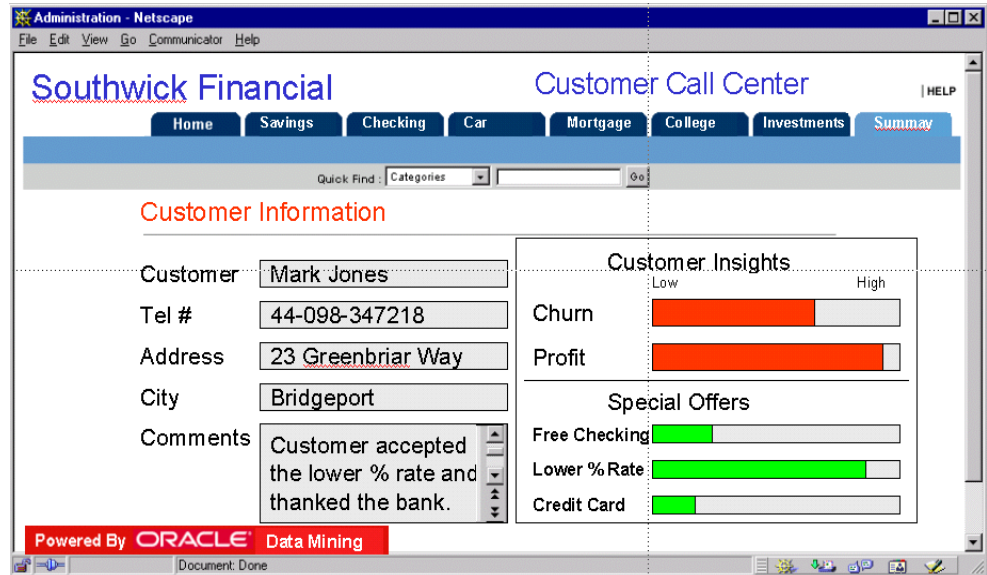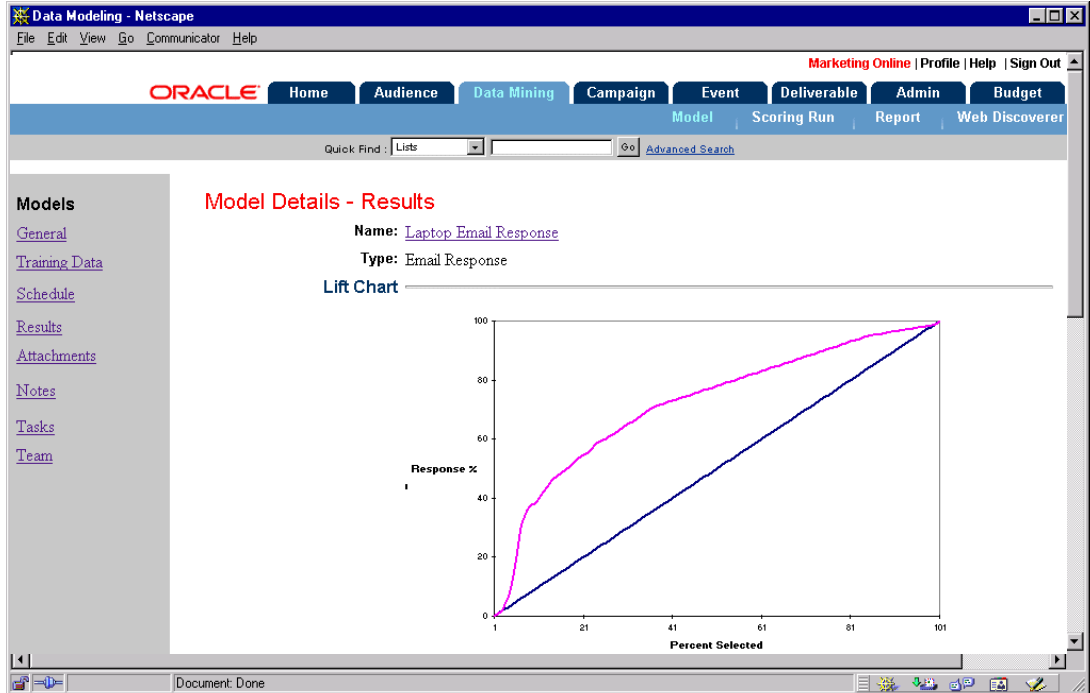
Oracle9*i* Data Mining's Java-based API enables application developers to enhance business applications with insights and predictions, such as this example call center that highlights a customer's likelihood to churn and to be a profitable customer. The probability that the customer will accept the special offers is also displayed for the customer service representative to better service the customer.

Capabilities for integrating data mining with other analytical tools and applications is greatly expanded with. Because all results are created and stored in an open relational database, users have access to data mining results using a wide variety of business intelligence tools.

ORACLE CRM 11i APPLICATION ADDS DATA MINING

The next major release of Oracle CRM 11i application adds integrated data mining to automate model-building and model-scoring activities for a set of predefined business problems. By constraining the problem and having a known schema, data mining can make certain modeling assumptions to automate the model-building and model-scoring process. CRM 11i's Marketing Online application allows business analysts and marketing professionals to take advantage of the power of data mining without the need to learn complex analytical techniques or have to perform extensive data preparation. CRM 11i now provides easy and accurate answers to four business questions:

- Which customers are likely to respond to an email offer?
- Which customers are likely to respond to a direct mail offer?
- Which customers are likely to respond to a telesales offer?
- Which customers are likely to remain loyal?

Model Details - Results

Name: Laptop Email Response

Type: Email Response

Lift Chart

(Response % vs Percent Selected)

Models
- General
- Training Data
- Schedule
- Results
- Attachments
- Notes
- Tasks
- Team

CRM users can design targeted campaigns by selecting "Audience" and answering a few simple questions. All of the data preparation, data mining, and data "scoring" occur automatically for the users.

Score Details

Model Name: SF Email Response

Model Type: Email Response
Created Date: 8/27/99
Update Date: 11/30/99
Last Run Date: 11/30/99

Score Status: Scored
Performance: Fastest
Owner: Mary Shuemyer

Decile Results

Responses 25,000   Scored List: Techno Savvy Customers

Decile	Number of Records	Cumulative Records	Response Score	Predicted Responders	Cumulative Predictive Responers
1	145,000	145,000	4.15%	6,018	6,018
2	145,000	290,000	3.76%	5,452	11,470
3	145,000	435,000	3.54%	5,133	16,603
4	145,000	580,000	3.12%	4,524	21,127
5	145,000	725,000	2.65%	3,843	24,969
6	145,000	870,000	2.48%	3,596	28,565
7	145,000	1,015,000	2.26%	3,277	31,842
8	145,000	1,160,000	2.01%	2,915	34,757
9	145,000	1,305,000	1.85%	2,683	37,439
10	145,000	1,450,000	1.50%	2,175	39,614

Market managers can simply select the top deciles of customers identified to be most likely to respond or behave in a particular way.

SUMMARY

Oracle9*i* Data Mining provides a powerful, scalable infrastructure for building applications that automate the extraction of business intelligence and its integration into other applications. Oracle9*i* Data Mining's Naïve Bayes and Association Rules data mining algorithms can solve a wide variety of business problems. Data mining, embedded in the Oracle9*i* Database and accessible through a Java-based API, provides a powerful infrastructure for building advanced enterprise business intelligence applications.

By automating, integrating, and "operationalizing" the extraction and distribution of new business intelligence insights, companies can leverage their investment in data, operate more effectively, and obtain greater competitive advantage.

ORACLE