TABLE OF CONTENTS

INTRODUCTION ........................................................................................................................................................................ 2
WHAT IS DATA MINING? ............................................................................................................................................................. 2
HOW TO ACHIEVE DATA MINING ............................................................................................................................................ 2
THE ROLE OF DARWIN ............................................................................................................................................................... 3
FEATURES OF DARWIN ............................................................................................................................................................... 4
USER FRIENDLY ............................................................................................................................................................................. 4
SCALABILITY ................................................................................................................................................................................ 6
VISUALIZATION .............................................................................................................................................................................. 8
FUNCTIONALITY ............................................................................................................................................................................ 10
   Data Mining for Telecommunications...................................................................................................................................... 11
   Data Mining for Financial Services ........................................................................................................................................ 11
   Data Mining for Government ................................................................................................................................................... 12
   Data Mining for Database Marketing ................................................................................................................................... 12
   Data Mining for Healthcare ........................................................................................................................................................ 12
SPECIAL FEATURES OF DARWIN .................................................................................................................................................. 13
LATEST FEATURES OF DARWIN .................................................................................................................................................... 14
STRENGTHS & LIMITATIONS OF DARWIN ................................................................................................................................ 16
   STRENGTHS ............................................................................................................................................................................. 16
   LIMITATIONS ........................................................................................................................................................................... 17
SUMMARY .................................................................................................................................................................................. 17
INTRODUCTION

WHAT IS DATA MINING?

First a brief introduction to what is Data Mining.

Data Mining constitutes Knowledge Discovery and Knowledge Deployment.

Knowledge Discovery is the process of finding the hidden patterns that transform data into insight.

Knowledge Deployment is the application of discovered knowledge for business advantage.

The figure below is an illustration of Data Mining:

![Data Mining Diagram]

HOW TO ACHIEVE DATA MINING

Several Data Mining software tools are available to perform data mining. Oracle Corporation’s Darwin is one such Data Mining tool. Oracle Darwin has been in the business of Data Mining for about 10 years. Oracle Darwin is a powerful data
mining software and serves to help companies understand and predict customer behavior to deliver integrated enterprise-wide customer relationship management solutions.

THE ROLE OF DARWIN

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.

Darwin helps you to 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

Darwin's data mining strategy is BUILD, EVOLVE, SELECT, EVALUATE, and APPLY.

Darwin can be used to build predictive models to reveal patterns hidden in data and reveal valuable new information…information that can provide early warning systems for churn and attrition; provide hidden insights about customer behavior, customer demand, and market segments; identify cross-sell and up-sell opportunities; and help combat fraud.

The information thus extracted from data can provide business intelligence that can improve customer relationships and hence generate profits and costs savings that go straight to the bottom line.

Darwin provides various functionality in a user-friendly environment that a business analyst can use, and yet they can always tap into the horsepower of some very powerful UNIX, perhaps SMP, multi-processor environment so they can mine larger amounts of data and extract more information.
FEATURES OF DARWIN

USER FRIENDLY

➤ MS Excel integration

➤ Intuitive GUI

➤ “Wizards” to guide and automate
  • Model wizard
  • Evaluation wizard

Darwin provides an intuitive, easy-to-use user interface. It also provides wizards to simplify and automate the data mining steps. For example, the Key Fields wizard automatically finds the variables that are most influential in addressing a particular question. The Model Seeker wizard automatically builds many data mining models, displays interactive graphs and tables of results, and recommends the best model(s).

Darwin workspace or tree-directory keeps track of data, models and results. Like all windows desktop applications, Darwin supports all point and click, drag and drop functionalities.
Darwin provides a variety of wizards to guide the user through the data mining process. Modeling wizard, along with the evaluation wizard, guides the user through the basic data mining. Modeling wizard allows the user to select the data to be mined, specify the target field, specify the desired model type, and give the model a name. After these, the user can optionally specify model parameters, or start the model building process.

![Model Wizard Interface](image)
The evaluation wizard applies the model created to another set of data to refine the models and display the results. It also allows the selection of the desired output type.

**SCALABILITY**

Darwin uses powerful algorithms and parallelized computing to achieve power and scalability during Data Mining. It uses a powerful, scalable, parallel UNIX server. Darwin runs with the following software:

- Client: Windows
- Servers:
  - Sun Solaris
  - HP-UX
  - Windows NT (Release 4.0)

Darwin data mining software runs in a Windows client and UNIX server architecture. Darwin emphasizes ease of use and yet have the power and depth of functionality to solve large problems.
One of Darwin’s key strengths is its ability to handle large sets of data by using parallel processing environments. When you have parallel processors, say in a Sun SMP environment, you can divide the task up across many processors. Darwin is shown to run in linear scalability. In other words, if there are 64 processors, it can mine the large amount of data 64 times faster than on a single-processor machine. That means it can mine more data and build more accurate models. It can include more variables for analysis as compared to traditional statistical packages (<= 30). With data mining 1,000+ variables can be taken as input variables and mined to identify key relationships and to build predictive models. And because modeling is an iterative process, many different types of models need to be built with different inputs and different assumptions.

**Scalable, Parallel Performance and Model Accuracy**

<table>
<thead>
<tr>
<th>Number of Processors</th>
<th>Darwin Module</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree</td>
<td>1</td>
<td>1.93</td>
<td>3.80</td>
<td>6.30</td>
<td></td>
</tr>
<tr>
<td>Net</td>
<td>1</td>
<td>1.92</td>
<td>3.36</td>
<td>7.37</td>
<td></td>
</tr>
<tr>
<td>Match (train)</td>
<td>1</td>
<td>2.00</td>
<td>4.04</td>
<td>7.97</td>
<td></td>
</tr>
</tbody>
</table>

With Darwin, you can run a large number of models very quickly and very easily to extract more business intelligence more rapidly.
VISUALIZATION

Darwin uses Excel to display the results in a nice, straightforward, user-friendly fashion. It also uses Microsoft Excel for graphing, say, in this case, for comparison between three different models, a Match model, a Net model, and a Tree model. It also supplies Margin and ROI graphs to show the financial patterns for decision making. Darwin’s outputs are geared towards ease of interpretation by business analysts.
FUNCTIONALITY

To solve business questions that SQL programming, statistical packages, and Query/OLAP tools cannot adequately address

Types of data Darwin can access

- Data warehouses
- Relational databases (ODBC)
- Support for SQL queries
- SAS files
- Flat files

Darwin can access data from a variety of network data sources. We can extract information from the database using ODBC drivers. We use DataDirect’s (from MERANT, formerly Micro Focus and INTERSOLV) gateways for database connections and can extract from other databases using their ODBC drivers. We also support SQL queries and we also can import and export SAS files. We can import flat files or ASCII text files. Once you bring the data into Darwin, you often need to prepare the data for data mining, and Darwin supplies a set of transformation functions such as sampling, randomization, generating new computed fields, handling missing values, merging, replacing values, splitting the dataset into train, testing, and evaluation datasets, and so on.

In the model-building phase, Darwin supplies multiple algorithms--C&RT, classification and regression trees, also known as decision trees, neural networks, k-nearest-neighbors technique of memory-based reasoning, and clustering. All of these algorithms are full-featured, tested, accurate, very powerful as they have all been implemented to take advantage of parallel
computing techniques. For speed and the ability to handle large amounts of data—gigabytes and terabytes—they are implemented as parallelized algorithms.

**Data Mining for Telecommunications**

- Which customers are the most likely to churn, or leave for a competitor?
- What are the profiles of most and least profitable customers?
- Which products and services should be offered to which potential customers?
- What is the customer profile that characterizes frequent card users?
- Which patterns of usage indicate fraud?
- Which bundles of products and services will sell best to which current customers?

**Data Mining for Financial Services**

- Which customers am I most likely to lose?
- Who is most likely to transfer an outstanding balance?
- What are the profiles of most and least profitable customers?
- What is the risk associated with this loan applicant?
- What are the profiles of high-risk and low-risk customers?
- Who is most likely to default on a loan?
- What is the debt pattern indicating impending bankruptcy?
- What purchase patterns indicate credit card fraud?
- Which product bundles will sell best to whom?
Data Mining for Government

• Which patterns are associated with tax, welfare, illicit drugs, or other types of fraudulent behavior?

• What profile characteristics of people, equipment, or modes of transportation are indicative of future illegal behavior?

• Which military personnel make the best leaders under battlefield conditions?

• Which patterns help to police stock market activities?

• What are the indicators of investor trading violations?

• Which patterns indicate cases of money laundering?

Data Mining for Database Marketing

• What customer profiles are most promising for this campaign?

• Who is more likely to respond to a given offer?

• What offers are more likely to enlist a positive response?

• How should prospects be segmented for maximum profitability?

• What customer characteristics are associated with the highest lifetime earnings?

• How often should customers be approached, and when?

Data Mining for Healthcare

• What patterns indicate fraudulent claims from patients and/or doctors?

• In which cases do which drugs or treatments work best?

• Which customers are most likely to respond to a promotion for this new drug?
SPECIAL FEATURES OF DARWIN

Darwin also has a unique, special feature called the workflow object. It automatically documents the data mining analysis steps and records them in a graphical function. You can also query those objects to see what the object contains, when it was created, and the information about that object.

Darwin also has a scripting capability that allows you to record, edit, and rerun "macros" of your data mining process, which is the programmatic equivalent of the workflow graphical object. And, with the scripting capabilities of Darwin, you can record data mining scripts of tasks that you want to repeat—say, rerun the January data, the February data and so on. You can also save and/or edit and use these scripts to integrate Darwin into other customer applications so you can develop an entire closed-loop marketing automation solution.
Another unique feature of Darwin is the ability to export the models. Darwin can export C, C++, and Java models which then can be integrated into other applications.

**LATEST FEATURES OF DARWIN**

- Data import wizards for databases and text.

- A missing value treatment wizard that handles the problem of missing data. Darwin supports both field-wise and record-wise treatments.

- A Model Seeker wizard automatically builds many different data mining models, presents the results in interactive graphs and tables, and recommends the best model(s).
• A new computed fields transform allows users to create new derived fields using over 100 statistical, mathematical, comparison, and boolean logic expressions.

• An interactive Tree display feature allows users to view and query decision trees and the rules that Darwin generates.

• The Key Fields wizard automatically sifts through the data using a series of C&RT dives into the data to identify the fields that are most influential in addressing a particular problem. The output of the Key Fields wizard is a useful subset of data fields for input into OLAP.

• Multi-model comparison to easily compare mining models.

• Lift charts for Trees that show the lift by tree segment for more refined control in your market segmentation.

• An editable workflow object to document and communicate your mining activities.

• Clustering wizard – k-means. Clustering, also known as “Unsupervised Learning,” finds groups of records that are similar in some ways.

• ODBC write-back to the database
Interactive cluster reports allow users to view and query the clusters created by Darwin so they can make better decisions.

- Significant enhancement in the Darwin server and ability to run on Windows NT servers.

- Darwin provides native access to the Oracle database, which besides faster access to the data, also facilitates scoring--or placing the prediction results back into the database.

- Darwin provides faster algorithms by taking advantage of “threads” for better parallelism. Additionally, includes support for Naïve Bayes and enhances support for clustering by adding self-organizing maps (SOM).

**STRENGTHS & LIMITATIONS OF DARWIN**

**STRENGTHS**

- Darwin is more powerful in classification and prediction than IBM IM. For example,

  - Darwin’s CART allows user to choose the best among different pruned subtrees, by testing and evaluation.

  - Darwin’s neural network allows user to specify its number of layers, number of nodes at each layer, and learning algorithms. It has also mechanisms to avoid over-training.

  - Darwin’s “build, test, evaluate, and apply” strategy avoids idiosyncratic mining results, and offers a better estimate of the accuracy likely to be obtained when the model is run against new data.

- Darwin has a genetic algorithm. Even it takes more time to train than neural networks, the genetic algorithm is always able to find a global optimization.
LIMITATIONS

- Darwin is dedicated to classification and prediction, and does not support association and time sequence analysis.

- Darwin has little visualization functions.

SUMMARY

Darwin is enterprise-wide, parallel, scalable data mining software that helps you rapidly extract business intelligence hidden within large amounts of customer data.

Darwin’s strength is being able to handle large amounts of data, which is what data mining is all about, not only large amounts of data in terms of volume but also breadth in terms of the number of variables.

Darwin also supports a comprehensive multi-algorithmic approach so as to mine more data for better understanding of customer data behavior.

Darwin is easy to use, with the standard Windows client that provides access to the full horsepower of Darwin on the server.

Darwin can also integrate the results into other existing systems to distribute the business intelligence throughout the enterprise.