User's Guide

WizWhy®

Version 2014

WizSoft®

1. Welcome!

Many data sets contain valuable information that is not readily obvious. Examples of such information might be:

- Patterns of high-risk companies within financial data;
- Types of customers in a direct mailing list who are most likely to make a purchase;
- The relationships between patients' personal data and their medical diagnosis.

The search for these valuable, yet hidden, patterns and relationships within the data is known as *data mining*.

Users may be interested in data mining applications for several reasons:

- Some users are interested in data mining for issuing a *summary* of the data: When the data is too numerous to be reviewed record by record, there is a need for a summary, and revealing the main patterns in the data provides a useful summary.
- Other users expect data mining to reveal *interesting* phenomena in the data. These users wish to ignore trivial cases and concentrate rather on unexpected phenomena.
- Still other users are interested in issuing *predictions* for new cases. They wish to reveal the patterns in the data in order to use them for issuing predictions for new cases. For example, revealing the patterns of risky customers in financial data, enables one to predict to what extent a new customer is risky.
- Finally some users may also be interested in using data mining for *auditing* purposes. The records that deviate from the discovered patterns in the data might be cases of data entry error or fraud.

How does *WizWhy* work?

Prior to using *WizWhy*, one should have a data set that he or she wishes to analyze. *WizWhy* will determine how the values of one field are affected by the values of other fields.

For example, suppose that you maintain a customer data set where each record contains a range of fields relating to a customer, such as: Customer Name, Address, City, State, Field of Business, Salesperson, Amount Purchased, % Growth Since Last Year. One of these fields, say, % Growth Since Last Year, should be defined as the d*ependent variable*, while the other fields are the i*ndependent variables*.

In this example, the aim might be analyzing customer retention, that is, revealing the patterns of those customers where the % Growth Since Last Year is negative. The % Growth Since Last Year might then be analyzed as either Boolean or continuous. In a Boolean analysis the aim might be revealing the patterns of the customers where the Growth Since Last Year is less than, say, 0%. A continuous analysis is more detailed: it calculates the specific % growth of each customer as a function of the other fields.

On analyzing the data, *WizWhy* performs the following operations:

1. It first reads the data. The user selects the dependent variable (in the above-mentioned example - % Growth Since Last Year) and can fine-tune the analysis by defining parameters such as the minimum probability of the rules, the minimum number of cases in each rule, and the cost of a miss vs. the cost of a false alarm. *WizWhy* follows these "instructions" when issuing the rules.
2. Within a short time, *WizWhy* lists the rules that relate between the dependent variable and the other fields. The rules are formulated as "if-then" and "if-and-only-if" sentences. On the basis of the

discovered rules *WizWhy* also points out the main patterns and the unexpected phenomena and cases in the data.

3. *WizWhy* can now make predictions for new cases; for instance, given the data of a new customer, *WizWhy* can calculate the expected % Growth. These predictions can be either Boolean (for instance, whether or not the % Growth is above 0%) or continuous (for example, the % Growth is between 20% - 30%).

What is an if-then rule?

WizWhy starts analyzing the data by revealing all the if-then rules that relate between the Dependent Variable and the other fields. An example of an *if-then* rule is:

> *If* City *is* New-York
> *and* Amount Purchased *is* 200 … 300 (average = 250)
> *and* Salesperson *is* Dave
> *Then*
> Growth Since Last Year *is* less than 0%
>
> *Rule's probability*: 0.70
> *The rule exists in* 370 *records*
> *Significance Level: Error probability* < 0.001

This rule says that for 70% of the customers, residing in the City New York, and purchasing between 200 and 300, and the Salesperson is Dave, the Growth Since Last Year is negative. There are 370 such customers.

The term "probability" designates what other data mining tools call "Confidence Level". Obviously, this probability should be significantly higher than the overall frequency of the value under analysis (i.e., the frequency of customers, where the Growth Since Last Year is less than 0%, is much lower than the rule probability, 70%).

"Error probability" indicates the degree to which the rule can be relied upon as a basis for predictions. Assuming that the data under analysis is a representative sample of an infinite population, the error probability quantifies the chances that the rule does not hold in the entire population and exists accidentally in the file under analysis.

Numeric fields, such as Amount Purchased and % Growth, are automatically segmented into intervals, and these intervals are the values in the if-then rules. For example, in the if-then rule above, the second condition refers to the case where the value in the Amount Purchased field is the interval between 200 and 300. *WizWhy* employs a unique algorithm for the optimal segmentation of numeric (continuous) fields.

Revealing all the if-then rules is known as the "association rules" method. One of the main challenges of such a method is to validate each possible relationship, in a *reasonable time-span*. For instance, the data might contain:

- 10,000 records
- 20 fields in each record
- An average of 10 possible values for each field

Using conventional means to check every possible relation in such data would require thousands of years. *WizWhy* employs a sophisticated algorithm that reveals all the rules in an astonishingly short time. In the previous example, it would take *WizWhy* just a few minutes to discover *all* the if-then rules under investigation.

What is an if-and-only-if rule?

On the basis of the if-then rules *WizWhy* proceeds to search for if-and-only-if rules. An example of an if-and-only-if rule is:

> The following conditions explain when the Growth Since Last Year is less than 0%:
>
> If at least one of these conditions holds, the probability that the Growth Since Last Year is less than 0% is 0.9
>
> If none of these conditions holds, the probability that the Growth Since Last Year is *not* less than 0% is 0.95
>
> The conditions are:
>
> 1. The Amount Purchased is between 0 … 199 (average = 100)

2. The Salesperson is Dan
   and the City is Boston

In other words, the Growth Since Last Year is negative, if and only if, the Amount Purchased is between 0 … 199 *or* the Salesperson is Dan and the City is Boston.

If-then rules represent sufficient conditions (the "if" condition is a sufficient condition for the result). If-and-only-if rules go one step further: they represent necessary and sufficient conditions. In the previous example, the two conditions, (1) and (2), are necessary and sufficient conditions for the growth being negative. If at least one of them holds there is a high probability that the Growth is negative, and if the two of them do not hold, there is high probability that the Growth is *not* negative.

Obviously such a relation cannot be accidental, and therefore might be relied upon when issuing predictions. Indeed, when *WizWhy* reveals if-and-only-if rules it takes them into account when issuing predictions for new cases.

Revealing the if-and-only-if rules also helps in pointing out the main patterns in data. One common "complaint" against if-then rules is that they are too numerous. Indeed, in many data sets *WizWhy* may discover thousands of if-then rules. All of them are valid and may be used for issuing predictions, but cannot be manually reviewed in any practical way. Revealing if-and-only-if rules solves this problem. Each value of the dependant variable is explained by one or two if-and-only-if-rules. These if-and-only-if rules are optimal in the sense that they cover the maximum number of both positive and negative cases.

How does *WizWhy* summarize the data?

As already mentioned, some users are interested in data mining in order to issue a *data summarization.* They are interested in a report that presents the main patterns in the data.

WizWhy meets this target by listing the relations between all the values in each field and the dependent variable. Consider the Amount Purchased field in the above-mentioned example. *WizWhy* segments the field into

intervals (as mentioned, *WizWhy* employs a unique algorithm that segments numeric fields in an optimal way), and displays the relation between each interval and the value under analysis (Growth Since Last Year is less than 0%).

For example, when analyzing the Amount Purchased field, *WizWhy* may reveal the following relations:

IF The Amount Purchased is between:	*THEN* The probability, that the Growth Since Last Year is less than 0%, is:
0 - 199	50%
200 - 300	40%
301 - 480	30%
481 - 791	15%

Each line is a one-condition if-then statement. Some or even all of these if-then statements may not be rules, since they may not meet the requirement that the rule probability be significantly higher or lower than the primary frequency. Still, all of them represent basic trends in the data. *WizWhy* applies such an analysis to each of the fields.

WizWhy also calculates the *explanatory power* of each field. The explanatory power designates to what extent the field explains the dependent variable. When sorting the fields by this parameter one can see the fields listed according to their "importance" in explaining which customers are likely to leave for a competitor.

WizWhy illustrates graphically these one-condition relations: each value is represented as a bar, where the height denotes the probability of the customer leaving for a competitor, and the width signifies the number of cases (i.e., customers) having this value.

This analysis of the basic rules and trends results in a data summarization. These basic rules and trends summarize the data, in the sense that they explain most of the other rules. More accurately they explain all the rules having more than one condition with the exception of the *unexpected* rules. The idea of unexpected rules will be further discussed in the next section.

How does *WizWhy* reveal interesting phenomena?

Revealing interesting phenomena relies on the assumption that *unexpected* phenomena are interesting. For example, an event that is inconsistent with (namely unexpected by) an accepted theory is an interesting event. Now, each rule can be viewed as an event, and the one-condition rules and trends, discussed in the previous section, can be viewed as the "basic theory" that describes the data. Therefore, by calculating how unlikely each rule is relative to the basic rules and trends, the *unexpected rules* can be revealed. These unexpected rules signify the interesting phenomena in the data.

The unexpected rule has at least two conditions. The basic rules have fewer conditions (in many cases they have one condition only), and the basic trends, by definition, have one condition only, as well. Each of the conditions of the unexpected rule appears in the basic rules and trends. The unexpected rule is an unlikely event relative to the basic rules and trends.

The level of unlikelihood is computed in the following manner: consider a data set of 1000 records, where each record refers to one patient, and contains the information regarding whether the patient shows either symptom A or B and the diagnosis (whether or not the patient suffers from the disease D). Suppose also that 30% of the patients have the disease D, and the following three rules were discovered:

1. If a patient shows symptom A, the probability that he or she suffers from the disease D is 60%.
2. If the patient shows symptom B, the probability that he or she suffers from the disease D is 60%.
3. If the patient shows both, symptom A and symptom B, the probability of the disease D is 10%.

In this example, rules no. 1 and 2 are the basic rules, and rule #3 is the unexpected rule.

WizWhy calculates what *should have been* the probability of having the disease D among the patients showing both symptoms, A and B, on the basis of rules 1 and 2, contrary to the *actual* probability in rule 3. This is the *expected* probability. To calculate the expected probability *WizWhy*

measures the dependency between symptom A and B. The difference between the expected probability and the actual probability signifies how unlikely rule 3 is in regard to rules 1 and 2.

WizWhy measures this unlikelihood in an additional way. *WizWhy* calculates the *conditional probability* of the event described in rule 3, given the conditions described in rules 1 and 2. In the example under discussion, the conditional probability is almost 0, and therefore, the level of unlikelihood, which is 1 minus the conditional probability, is almost 1. Note that the probability of the Unexpected Rule may be much higher or much lower than the expected probability. Any significant deviation is unexpected.

How does *WizWhy* issue predictions?

WizWhy makes use of the rules discovered in the data set in order to issue predictions for new cases. When a new record is entered, *WizWhy* applies the rules on the values of this record and calculates the expected value of the dependent variable. For example, one can run *WizWhy* on financial data, where the dependent variable is a field signifying whether the company has gone bankrupt. *WizWhy* will reveal the rules that relate between the company data and the probability of going bankrupt. When the data of a new company is entered, *WizWhy* will then apply the relevant rules, and calculate the probability of this company going bankrupt. When issuing the predictions *WizWhy* can list the rules that entail each prediction. These rules serve as the explanations for the predictions.

The predictions can be either *Boolean* (for instance, whether the company will go bankrupt or not) or *multi-value* (for example, given the patient's symptoms, what the disease is) or *continuous* (for instance, given the financial data, what the expected rate of growth is).

Still, when issuing the predictions on the basis of rules one faces the following two problems:

(i) How can rules representing noise or overfitting be avoided?

For any rule discovered in the data, one may ask, what is the *reason* or the *explanation* for the existence of this rule. The two possible extreme answers are (i) the rule exists in the entire population, where the data under analysis is a representative sample, or (ii) the rule is the result of a chance, i.e., it is a case of a *noise* or *overfitting*. Obviously, when issuing predictions for new cases, rules existing in the entire population should be taken into account while rules representing noise should be ignored.

WizWhy measures the probability that a rule is the result of a chance by considering two parameters:

- The *error probability* of the rule: The lower the error probability, the lower the probability that the rule is the result of a chance, and therefore the higher the probability that the rule exists in the entire population.
- The *level of unlikelihood* of the rule (in case the rule is unexpected): The higher the level of unlikelihood, the lower the probability that the rule is an accidental phenomenon, and therefore the more the rule can be relied on when issuing predictions.

Note however that above-mentioned problem does not relate to the if-and-only-if rules. Obviously these rules cannot be accidental and therefore predictions that are based on if-and-only-if rules do not suffer from the problem of overfitting.

Avoiding overfitting is one of the main features that differentiate between *WizWhy* and other prediction applications based on neural nets, decision trees or genetic algorithms. As a result, the accuracy level of *WizWhy* predictions is usually much higher in comparison with other approaches.

(ii) How can cases of rule inconsistency be solved?

On using if-then rules to issue predictions one may face cases where some of the rules predict one way while others predict the other (for example, some of the rules predict that a certain company tends to go bankrupt, while other rules predict that it does not). *WizWhy* solves this problem by using the *error probability* and the *level of unlikelihood*. Since these two parameters signify the extent to which the rule can be relied on, *WizWhy*

weighs the rules according to these parameters and calculates the prediction in cases of inconsistency.

Once again, note that the problem does not apply to the case where the predictions are based on the if-and-only-if rules. The if-and-only- if rules are always consistent.

How does *WizWhy* point out unexpected cases?

On top of issuing predictions for new cases, the rules can be used to point out *unexpected cases* in the data. *WizWhy* issues predictions regarding the records of the data set under analysis and points out cases where the dependent variable's value deviates from the value anticipated according to the rules. Such a deviation may be the result of noise, but it can also indicate a data entry error, a fraud or another type of error. *WizWhy* lists those cases, displaying the expected value together with the relevant rules. The user can review the cases in order to audit the data.

Who can benefit from *WizWhy*?

Data mining in general, and *WizWhy* in particular, have many practical applications. You can use *WizWhy* in most cases where simple or complex data analysis and predictions are required. For instance:

- *WizWhy* can assist professionals in the fields of medicine and social sciences to enhance their diagnostic and research efforts.
- Banks and financial institutions can use *WizWhy* to indicate risky customers.
- Corporations implementing direct marketing will find that *WizWhy* is the ideal tool for increasing the success rate of direct mailing.

Scientific research: *WizWhy* can be applied for inferring rules in a wide range of scientific fields, including medicine, economics, psychology, sociology and geology. For example:

- In medical research, *WizWhy* can reveal laws relating between symptoms and diseases.
- In geological research, *WizWhy* can reveal rules that show the relationships between soil data and mineral location.
- Researchers in social sciences can use *WizWhy* to discern patterns and characteristics that designate students' aptitude for academic success.

Generally, in research entailing large quantities of data, *WizWhy* can save significant time and effort by assuming the scientist's burden of revealing rules. Data mining in general, and *WizWhy* in particular, may just revolutionize traditional methods of conducting research in scientific fields.

Banks, credit and insurance companies: These organizations can use *WizWhy* to discern financially unstable customers and to predict the degree of financial risk of a new customer. To do so, *WizWhy* reads the customer data file and reveals the patterns of the financially risky customers. Rules describing these patterns are saved to file. Later, company personnel can analyze a new customer by entering the customer's data and having *WizWhy* calculates the extent of that customer's financial risk.

Market research: Market researchers can use *WizWhy* to improve the success rate of direct marketing campaigns. Prior to mailing a large quantity of marketing literature, a small batch can be sent to a representative sample of the population, and *WizWhy* can then discover the patterns of both those customers more likely and less inclined to purchase, relative to the average purchasing rate. For example, if the purchasing rate is 2%, *WizWhy* can be instructed to discover those customers whose purchasing rate is 4% and above, or 1% and below. The result can then be applied to the master file, in order to define the optimal direct mailing list.

The contents of this guide

This manual is organized in a manner that familiarizes you with the *WizWhy* processes.

Chapter 2 explains how to install your *WizWhy* system to operate either as a stand-alone application or a network version.

Chapter 3 helps you launch *WizWhy* and familiarize yourself with the *WizWhy* desktop and sequence of operations.

Chapter 4 is a tutorial that shows you how to perform the basic *WizWhy* operations, step by step.

Chapter 5 tells you how to open data sets and prepare them to be analyzed by *WizWhy*.

Chapter 6 explains how to set the parameters that *WizWhy* will use to issue the rules.

Chapter 7 describes how to review the reports.

Chapter 8 shows how to generate predictions.

Chapter 9 contains a number of frequently asked questions about using *WizWhy*. This chapter will probably assist you in clarifying questions you may have and enable you in making the most out of your application.

At the end of the manual:

The **Glossary** provides definitions and explanations of terms used throughout this guide and in the *WizWhy* application.

The **Quick Reference** appendix describes the main *WizWhy* functions organized according to the *WizWhy* menu options. It should prove useful after you are “up and running,” as a fast reminder of how to perform certain operations.

The **Appendix** describes the mathematics behind *WizWhy*.

2. Installing *WizWhy*

WizWhy can be installed to operate as a stand-alone system or as a network application. A CD containing both the *WizWhy* application and the software for installing ODBC and OLE DB drivers is included in the installation kit.

Before you begin installation, be sure to close all open applications.

Software Installation

Insert the CD into the appropriate drive of your computer. Usually the installation will start automatically. If not,

- From the Start menu, select Run to display the Run dialog box.
- In the Open text box, select d:\setup or e:\setup, as appropriate, and click OK.

The Master Setup Dialog window will be displayed, presenting the following installation options:

- Click on the Install WizWhy Analyzer button to install the basic *WizWhy* program that analyzes the data and issues predictions.
- Click on the Install WizWhy Predictor button to install the independent predictor program. (See chapter 8 for explanations about this option).
- The Install Microsoft Internet Explorer ver. 5 button is active if the Microsoft Internet Explore ver. 5 (or later) is not already installed. This program is needed for the *WizWhy* online Help.

After installing one program you can repeat the above steps to install another.

Read the displayed Welcome screen and click the Next command button when you are ready to continue. The Registration window will be displayed.

Enter your Name and Company, click Next and then confirm your selections by clicking Yes. The Setup Install Type dialog box will be displayed.

Select the type of *WizWhy* installation that you require. First-time users should usually select the Typical option.

In the Destination Directory block, check the displayed path for installing *WizWhy*. If it is not appropriate, use the Browse button to indicate the desired path. Clicks Next to continue.

Follow the displayed instructions to complete the installation process.

3. Getting Started

Before you begin using *WizWhy*, be sure that you have a data set containing the data you wish to analyze.

The *WizWhy* Menu

The *WizWhy* menu contains the following options:

File - for managing the *WizWhy* files and for opening data tables.

Edit - for performing common editing functions.

View - for displaying or hiding the *WizWhy* bars.

Issue - for activating the issuing rules and issuing predictions commands.

Settings - for selecting fonts and properties.

Windows - for selecting the active window.

The Action Bar

You can use these four command buttons to instruct *WizWhy* what to issue.

Click the Issue Trends button to instruct *WizWhy* to issue the trend report (without issuing the rules).

Click the Issue Rules button to instruct *WizWhy* to reveal the rules and issue the reports (including the trend report).

Click the Predict to File button to update predictions in another file.

Click the Predict on-line to have *WizWhy* make a prediction for a new case that is entered manually.

The Report bar

You can use these seven command buttons to move between *WizWhy*'s main window and the report windows.

Click on the left button to move to the main window.

Click on the button with the R to move to the Rule report

Click on the button with the T to move to the Trend report.

Click on the button with the U to move to the Unexpected rules report.

Click on the button with the C to move to the If-and-only-if rules report (the if-and-only-if rules).

Click on the button with the S to move to the unexpected Cases report.

Click on the button with the B to move to the Prediction report.

The Window bar

You can use this Window bar to move between *WizWhy*'s main window and the report windows like the Report bar.

Click on the + and then select the window.

Sequence of operations

There are four main steps in operating *WizWhy*:

First, you open the data set to be analyzed. You have several options for this step. *WizWhy* can --

- Read a text file using the *WizWhy* ASCII reader.
- Directly read a file with a .dbf extension.
- Directly read MS Access table(s)
- Directly read MS SQL table(s)
- Directly read an Oracle data set.
- Read a data set (having one or more tables) through ODBC.
- Read a data set (having one or more tables) through OLE DB.

Once the data set is open, you select the dependent variable and determine the parameters to be used in the analysis.

You can now select one of the following options:

Click on the Issue Trends button if you are interested in viewing the Trend report only.

Click on the Issue Rules button if you are interested in viewing all the reports (including the trend report) and issuing predictions.
Once the rules have been issued you can generate predictions.

Click on Predict to File button to update the predictions in another file. If you click on this command before the rules have been issued, *WizWhy* will issue all the reports (as if you clicked on the Issue Rules buttons) *and* update the predictions.

Click on the Predict on-line button if you are interested in issuing predictions on the screen for new cases entered manually.

Chapters 5 through 8 describe these operations in detail.

4. Tutorial

This chapter teaches you how to perform the basic *WizWhy* operations. In this tutorial, you will:

- Open a data set
- Define the analysis parameters
- Review the reports
- Issue predictions

Before you begin, be sure that your system has been properly installed and that Stock investor 1000 companies 1.mdb and Stock investor 1000 companies 2.mdb reside in your Wizfiles folder. The first table contains financial data on 1000 companies in the US stock market. You will analyze the annual sales of these companies. The second table contains data on another 1000 companies in the US stock market. You will issue predictions to the companies in this second table on the basis of the rules discovered in the first table.

Access the *WizWhy* application by double clicking on the *WizWhy* icon on the desktop. Alternatively, open the Start menu and from the Programs list, select WizSoft Applications and then select WizWhy. The *WizWhy* desktop will be displayed.

Opening the data set

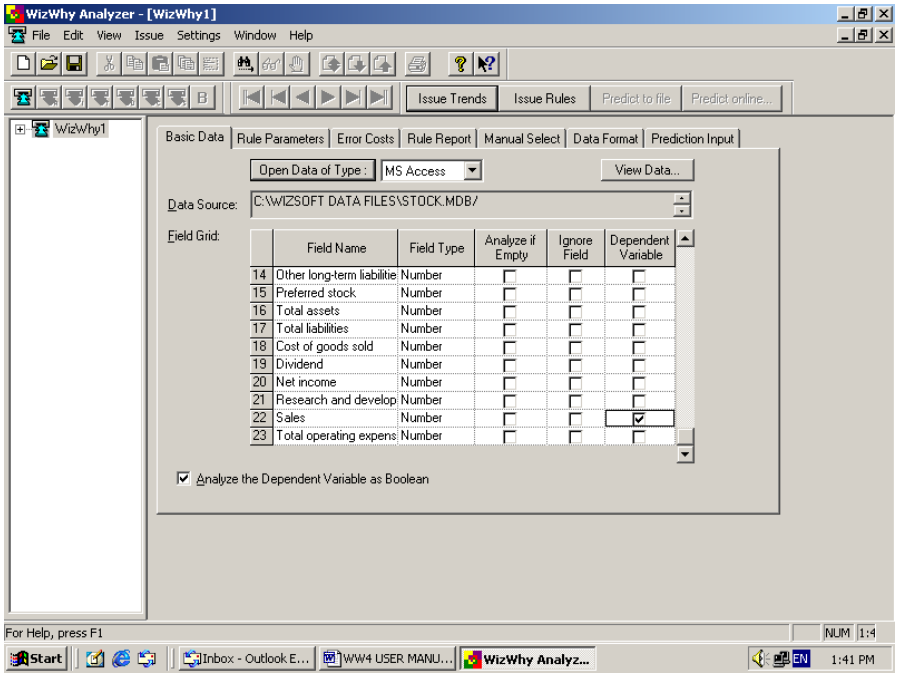
As mentioned, you will open two Microsoft Access data sets, both containing financial data on 1000 companies in the US stock market. You will reveal the patterns in one data set, and issue predictions in the other. The dependent variable will be the annual sales.

Click on the Open Data of Type button. The Open dialog box will be displayed.

Click on the Stock data source name and click OK. The MS Access Data Source dialog box will be displayed.

Click on the Stock data source name in the left pane, and select the Stock investor 1000 companies 1 table in the right pane. Click OK. *WizWhy* will open the table and display its fields in the *WizWhy* work area.

The name of the table that you selected is displayed in the Data Source text box. Under that is the field grid, which includes the names of all the fields listed in the Field Name column.



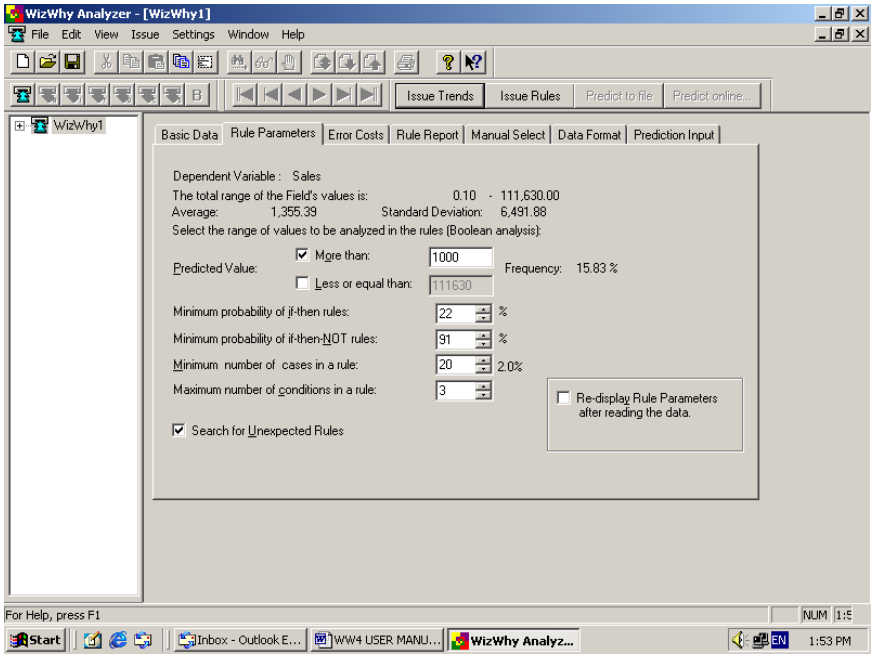
Determining the analysis parameters

Once the data set to be analyzed is open, the analysis parameters can be determined. At this stage you have to select one of the fields as the dependent variable.

In the Dependent Variable column (on the right side) select the check box in the Sales line (the 22nd line). This field will be the dependent variable.

WizWhy can perform both Boolean and multi-value analysis. In a Boolean analysis the dependent variable is analyzed as if it has two values (True, False) while a multi-value analysis analyzes *all* the values of the dependent variable. The Boolean analysis is the default.

Click the Rule Parameters tab. Since the dependent variable (the annual sales) is continuous, while you perform a Boolean analysis, you have to define the range of values that will be analyzed as a Boolean value. Enter 1000 in the More than box, that is, the annual sales are more than $1000 millions. *WizWhy* will reveal the rules explaining when the annual sales are more than 1000 or not.



All the other analysis parameters will be defined by their default values. (You can read about these parameters in chapter 6).

Click on the Issue Rules button

WizWhy will begin its analysis and display a progress indicator. Upon completion, it will display the number of rules discovered. Click OK to instruct *WizWhy* to display the reports.

Reviewing the reports

You can now review the reports (read chapter 7 for more details).

Click on the + sign in the Window bar in the left pane. *WizWhy* will list 6 reports.

The Summary Report summarizes the rules' *explanatory power*.

The If-Then Rules Report lists the discovered *if-then rules*.

The Trend Report presents graphically and textually the one-condition trends in the data. These trends *summarize* the data.

The Unexpected Rule Report displays the rules that are unexpected relative to more basic rules and trends. These unexpected rules describe *interesting* phenomena in the data.

The If-And-Only-If Rules Report lists the if-and-only-if rules (i.e., *necessary and sufficient conditions*).

The Unexpected Cases Report displays the records where the dependent variable's value *deviates* from the expected value according to the discovered rules.

Issuing predictions

Once the rules have been issued you can issue predictions. If you wish to issue predictions on the screen for data that you enter manually, click the Predict On-Line button. The Predictor dialog box will be displayed.

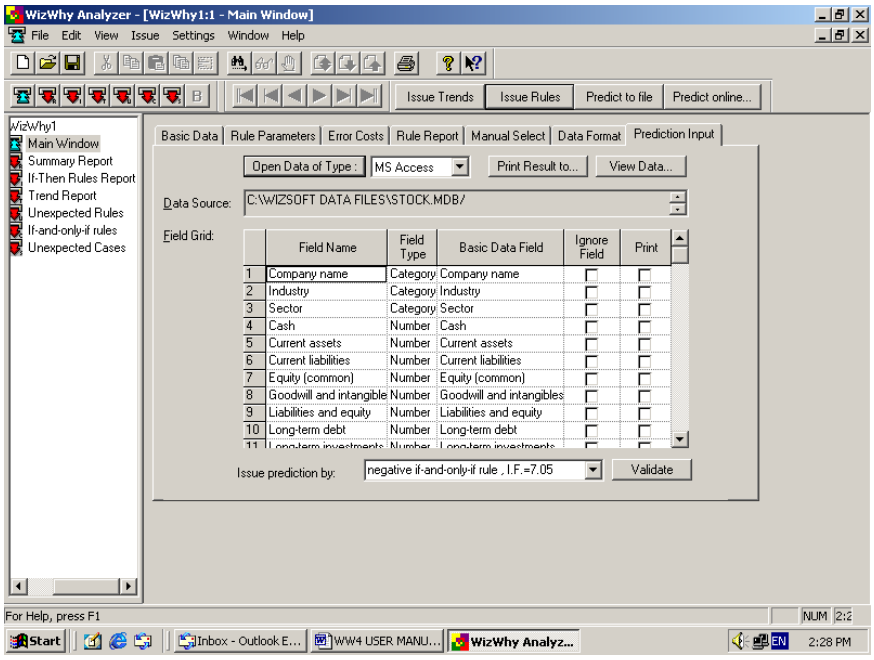
To issue a prediction for a new record (another company), you need to provide some values of this company. *WizWhy* will apply the discovered rules on these values and issue a prediction as to the expected sales of this company. Enter values in some or all the fields, and then click on the Issue Report button. The Prediction report will be displayed.

You will now issue predictions to another file. The data for the prediction will be the data of the 1000 companies in the second table. Since the sales of these companies are already known, you will be able to check the accuracy of the *WizWhy* predictions. We will refer to this table as the Test file (while the first table is the Train file).

Click on the Main window in the Window bar in the left pane. The Main window where you entered the analysis parameters will be displayed.

Click the Prediction Input tab. The Prediction Input window will be displayed.

Following the instructions presented above in regard to opening the first table, select the Stock investor 1000 companies 2 table.



After clicking OK in the MS Access Data Source dialog box, *WizWhy* will display another dialog box. Here you enter the name of an ASCII file in which the predictions will be updated. *WizWhy* will read the data in the Test file, apply the rules to issue a prediction for each record, and update the predictions in the file that you determine in the current dialog box. Enter a name (for example, Stock predictions), and click the Save button. *WizWhy* will now displays the fields of the Test file. (These fields are identical to the Train file's fields).

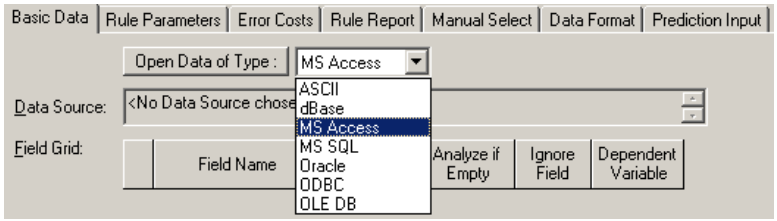
Click the Validate command button in the bottom of the screen. *WizWhy* will read the Test file, apply the rules on the data of the records in this table, and compare the actual values with the predicted ones. A summary of this validation will be displayed on the screen.

If you wish to update the predictions in the Stock Predictions ASCII file, click the Predict to File button. The file will contain the predicted sales (more than 1000 or not) and the conclusive probability. You can open this file by an application such as Microsoft Access or Excel.

5. Opening the Data Set

WizWhy can read seven types of data

- Files produced through dBase, FoxPro, Clipper etc. (*.dbf files) are read directly.
- Microsoft Access (*.mdb) tables are read directly.
- Microsoft SQL Server tables are read directly.
- Oracle data tables are read directly.
- ODBC (Open DataBase Connectivity) compliant databases (produced by Access, SQL, Oracle, Sybase, Informix, DB/2 and other programs) can be read through the ODBC interface.
- OLE DB compliant databases (produced by Access, SQL, Oracle, Sybase, Informix, DB/2 and other programs) can be read through the OLE DB interface.
- ASCII-type text files can be read through the *WizWhy* ASCII reader.



WizWhy can read Microsoft Access, Microsoft SQL and Oracle data both directly, and through its ODBC or OLE DB utility. The direct reading is simpler.

Direct reading of *.dbf files

To open a *.dbf file -

- In the Open Data of Type list box select dBase, or click on the Open Data of Type button if dBase has been already selected. The Open dialog box will be displayed.
- In the Files of type text box at the bottom, be sure that the option dBase, FoxPro, Clipper Files (*.dbf) is displayed.
- Using the Look in list at the top, locate the directory containing the database file. All *.dbf files in that directory will be displayed.
- Select the file to be analyzed from the list.
- Click Open. The data set you selected will be displayed in the *WizWhy* work area.

Direct reading of a single Microsoft Access (*.mdb) table

WizWhy can read one or several tables. In the following lines, the instructions how to open a single table are presented. The instructions how to join several tables can be found in pages 28 - 30.

- To open a single Access table, in the Open Data of Type list box select Access, or click on the Open Data of Type button if Access has been already selected. The MS Access Open dialog box will be displayed.
- Select ACCESS Files (*.mdb) from the Files of type drop-down list. Click on the file name and then click Open. The Access Data Sources dialog box will be displayed.
- Click on the data source name. The tables of this data source will be displayed in the right pane.
- Select the table to be analyzed in the right pane and click OK.

Direct reading of a single Microsoft SQL table

WizWhy can read one or several tables. The instructions how to open a single table are presented below. The instructions how to join several tables can be found in pages 28 - 30.

- To open a single Microsoft SQL table, in the Open Data of Type list box select MS SQL, or click on the Open Data of Type button if MS SQL has been already selected. The Microsoft SQL Login box will be displayed.
- Select or type the server service name in the Server box.
- Enter the User Name and Password for this service, or select the Trusted Connection check box if you are a member of the SQL Server users (your Windows Login Username & Password will be used).
- Click the Connect button. The MS SQL Data Sources dialog box will be displayed.
- Locate your database tables. The database tables will be displayed in the Contents list in the right pane of the SQL Data Sources dialog box.
- Select the data set table(s) to be analyzed in the right pane and click OK.

Direct Reading of a single Oracle table

WizWhy can read one or several tables. The instructions how to open a single table are presented below. The instructions how to join several tables can be found in pages 28 - 30.

- To open a single Oracle table, in the Open Data of Type list box select Oracle, or click on the Open Data of Type button if Oracle has been already selected. The Oracle Login box will be displayed.
- Select or type the server service name in the Server list box.
- Enter the User Name and Password for this service.

- Click the Connect button. The Oracle Data Sources dialog box will be displayed.
- Locate your database tables. The database tables will be displayed in the Contents list in the right pane of the Oracle Data Sources dialog box.
- Select the data set table(s) to be analyzed in the right pane and click OK.

Combining several tables into one *WizWhy* data set

You can combine two or more tables by defining the relationships between them. The Database Relationships window is displayed automatically when multiple tables are selected. Two means of definition are available:

- Defining the table relationships directly (text view)
- Defining the links graphically

Defining relationships directly

- Select the Text View tab.
- In the displayed dialog box, select the Primary Table and Primary Field from the first two lists. The primary table and field(s) are the basis from which the relationships are defined. These connections can be either from the primary table/field to a single related table/field (one-to-one) or from the primary table/field to a number of other related tables/fields (one-to-many).
- From the subsequent lists, select the Related Table and Related Field and click Add Join. The related fields will be displayed in the list at the center of the dialog box.
- Repeat the previous steps to define all the relationships. You can use the command buttons at the right as follows:

To *change* a defined relationship, select the defined relationship in the list, revise the relationship by redefining the Primary Table/Field and the Related Table/Field and then click Change.

To delete a relationship, select it from the list and click Delete.

- With Access files, you may have *WizWhy* define relationships automatically, by clicking Autodetect. The system will build the relationships according to identical field names in the tables *and* field types. If a relationship between two tables of non-similar names and types has been defined previously, Autodetect will not delete the relationship and will not create any new relationship between those tables.
- When reading tables directly (*not* through ODBC), *WizWhy* will identify relationships defined previously in the database itself.
- To revise your selection of tables, click the Tables button. The Tables dialog box will be displayed.
- The Selected Tables list at the right displays the tables that you selected in the ODBC Data Sources dialog box. The Non- Selected Tables are those that were not selected. Use the Add and Delete buttons to revise your list of Selected Tables and click OK.
- Click OK to continue

Defining relationships graphically

- Select the Graphical View tab. The graphical Database Relationships dialog box will be displayed.
- The fields of each selected table will be displayed. To create a link between the fields, select a field in the first table and drag the cursor to the related field in the next table.
- To create additional relationships, repeat the previous step.
- You can use the command buttons at the right as follows:

  With Access files, you may have *WizWhy* define relationships automatically, by clicking Autodetect. The system will build the relationships according to identical field names in the tables *and*

field types. If a relationship between two tables of non-similar names and types has been defined previously, Autodetect will not delete the relationship, but it will not create any new relationship between those tables.

When reading tables directly (*not* through ODBC), *WizWhy* will identify relationships defined previously in the database itself.

To revise your selection of tables, click the Tables button. The Tables dialog box will be displayed.

The Selected Tables list at the right displays the tables that you selected in the ODBC Data Sources dialog box. The Non- Selected Tables are those that were not selected. Use the Add and Delete buttons to revise your list of Selected Tables and click OK.

To enlarge the display to full size, click Full Screen size.

To view the tables in the order, in which they were chosen, click Arrange.

To remove certain relationships, select them and click Delete.

To remove all defined relationships, click Delete All.

- When you have defined all relationships, click OK.
- To redefine the database relationships at any subsequent point, select File - Database Relationships to display the Database Relationships window.

Reading through ODBC

ODBC (Open Database Connectivity) is a software standard that enables one program to read a data set created by another program. Each database maintains its own specific driver. *WizWhy* comes with a set of ODBC drivers that are included in Microsoft's ODBC Software Development Kit.

To work with ODBC data sources, you must first be sure that you have installed the ODBC drivers for the database(s) that you wish to open. Once the drivers are installed, you may define ODBC data sources.

WizWhy operates like all standard Windows programs that incorporate ODBC-reading drivers. If *WizWhy* fails to open a table through ODBC, first check if another application – such as Microsoft Access – can open it using the same ODBC driver. If that application cannot open the table, you should consult with your dealer concerning your ODBC installation. The Appendix of this manual, "Working with ODBC" may also be of assistance.

- In the Open Data of Type list box select ODBC, or click on the Open Data of Type button if ODBC has been already selected. The Data Sources dialog box will be displayed.
- This dialog box displays all previously defined data sources. A data source is the name given to a connection between a database driver and the database file or directory. Following the Windows 95/NT standard, this window is divided into two panes:
- The left pane displays the tree of data sources and their tables and views (queries). To display the contents of the current level in the right pane only, click once on the source name or icon. To display the contents of the current level in both the right *and* left panes, double-click on the source. To view the contents of the data source in the *left* frame only, click on the sign at the left of the icon to display its sub-levels.
- The right pane displays the contents of the selected data source.
- Up to three levels can be displayed for a data source: the data source, its tables and the fields of the tables (for display only). Each level is displayed by clicking the sign or double-clicking on the name of the upper level.
- If the data source that you require is displayed, skip to the next step. If you have not yet defined your data source for a file or if you wish to reconfigure an existing data source, click the Data Sources button. A Data Sources dialog box will be displayed.
- Use this dialog box to add a new data source or reconfigure an existing one.

Adding a new data source

Click the Add button in the Data Sources dialog box. The Add Data Source dialog box will be displayed for you to select the data sources you will use with *WizWhy*.

From the Installed ODBC Drivers list, select the driver you need and click OK. A driver-specific Setup dialog box will be displayed for you to define the data source.

In the Data Source Name text box, enter a name for the data source, which will be used in *WizWhy*. You may add a description of the source in the Description text box. Always provide a unique name – such as Payroll or Accounts Payable – for each source, and provide the specifications required to connect the driver to the database.

After configuring your data source, click OK. The Data Sources dialog box will be displayed again.

Select the required database and click Close to return to the ODBC Data Sources dialog box.

Modifying an existing data source

From the Data Sources (Driver) list, select the data source to be reconfigured and click Setup. The relevant Setup dialog box will be displayed, as in the following example:

Make your modifications in the dialog box and click OK. The Data Sources dialog box will be redisplayed.

Click Close to return to the ODBC Data Source dialog box.

From the Data Sources list on the left, click *once* on the name of the data source that you require. The list of tables within that database will be displayed in the right pane.

Do *not* double-click on the file name to open it, because this will display the tables only in the left pane of the ODBC Data Sources dialog box. They must be displayed in both panes in order to select them.

From the Contents list on the right, click once on the database table(s) to be analyzed.

Do *not* double-click on a table name, because this will display the names of all the fields in the table and will deactivate the OK button for selecting the table.

You may select consecutive tables by highlighting the first name, holding down the Shift key and selecting the last name. Select non-consecutive databases by highlighting the first name, holding down the Ctrl key and selecting the other file names.

Depending on the type of data source, a number of different dialog boxes may be displayed, requesting you to define certain parameters such as database file or password.

Click OK. If you selected a single table, it will be displayed in the *WizWhy* work area. If you selected multiple tables, a Database Relationships window will be displayed. Follow the instructions outlined in the previous section – "Combining Several Tables into One *WizWhy* Database" – to define the relationships between the tables.

Reading through OLE DB

OLE DB is a set of interfaces that expose data from a variety of sources by using the Component Object Model (COM).

WizWhy comes with a set of OLE DB data providers that are included with the Microsoft Data Access Components (MDAC 2.5).

The Microsoft OLE DB data providers allows access to data, such as a Microsoft Access,, Microsoft SQL Server ,Oracle, DB2 or access to data through ODBC. (OLE DB data provider for DB2 is not included with MDAC 2.5 Kit).

- In the Open Data of Type list box select OLE DB or click Open Data of Type button if OLE DB has been selected. The Data Link Properties dialog box will be displayed. The OLE DB Provider List lists all the OLE DB providers detected on your hard disk.

- Select an OLE DB provider and click on the Next button. The data link properties may vary depending on your OLE DB provider. (To open the Microsoft Data Link Help click on the Help button). Use the Connection tab of the Data Link Properties dialog box to specify how to connect to your data and click OK.
- The OLE DB Data Sources dialog box will be displayed. The left pane displays the tree of data sources and their tables and views (queries). The first list in the tree includes all tables and views of the database. The other lists include tables and views of several schemas. A schema is a collection of database objects (tables, views) that are owned or have been created by a particular user. To display the contents of the current level in the right pane only, click once on the schema name or icon. The right pane displays the contents of the selected schema. The database object's name contains the table or view name and the schema name. From the Contents list on the right pane, click once on the database object name(s) to be analyzed and click OK.
- If you selected a single database object, it will be displayed in the *WizWhy* work array. If you selected a multiple database objects, a Database Relationships window will be displayed. Follow the instructions outlined in the previous section - "Combining several tables into one *WizWhy* database" - to define relationships between the database objects.

Reading ASCII files

WizWhy includes a unique ASCII file converter that you can use to read a standard text file (with a *.txt extension) created by a word processor, spreadsheet or any other program.

Opening the ASCII File

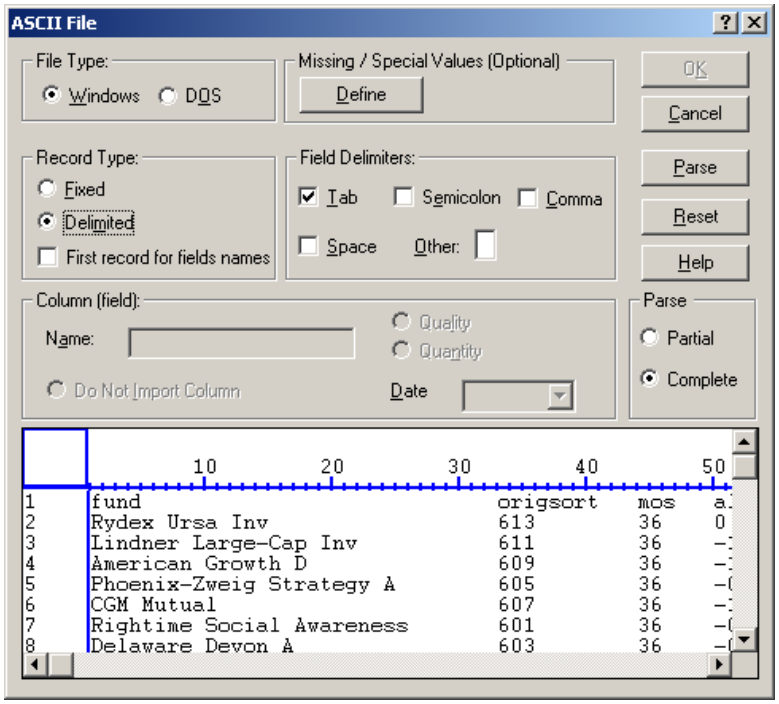
In the Open Data of Type list box select ASCII, or click on the Open Data of Type button if ASCII has been selected. The Open dialog box will be displayed.

In the Files of type text box at the bottom, select the ASCII Files (*.txt) option.

Using the Look in list at the top, locate the directory containing the ASCII (*.txt) file. All *.txt files in that directory will be displayed.

Select the file to be analyzed from the list.

Click Open. The ASCII file that you selected will be displayed in the *WizWhy* work area, as in the following example:



This dialog box is divided into two main sections:

The top section is used to define parameters for reading the data.

The bottom section is the ASCII file displayed as a table. Any change in parameters in the top section of the dialog box is instantly reflected in the bottom section.

Each line in the ASCII table represents a single record; the number at the left in the first column is therefore the serial number of each record. The numbers in the blue horizontal ruler are the character serial numbers. You may use the horizontal and vertical scroll bars to browse through the table and view all records and characters.

Partial vs. complete parsing

WizWhy re-reads the file all over again after each step in order to update the parsing and the field types. This may take considerable time when the file is large. You can shorten the time by instructing *WizWhy* to refer to the first 1000 records in the file only. The selection between partial and complete parsing is determined in the Parse frame at the right bottom corner of the top section. Select Partial to instruct *WizWhy* to read the first 1000 records only, or select Complete to indicate that the whole file should be read after each step.

Converting the file into a readable table

The procedure for converting the ASCII file into a table that *WizWhy* can read consists of three main steps:

- Filling in the top section.
- Indicating lines for exclusion.
- Performing field parsing.
- Defining and modifying the field features.

Filling in the top section

- In the File Type block at the top of the dialog box, indicate whether the file was created in DOS or in Windows. This will enable *WizWhy* to interpret special characters that are represented differently in each environment.
- In the Record Type block, indicate if the records in the ASCII file are:

Fixed: All records in the file are of the same length. This is the default.

Delimited: The records in the file are not of a fixed length. If the data in the table appears to be displayed incorrectly, select this option.

- If you selected Delimited, in the Field Delimiters block at the right, indicate the character by which *WizWhy* should delimit the fields: Tab, Semicolon, Comma, Space or Other. Once you select the delimiter character, *WizWhy* will update the data table, using the selected character to separate the fields in each record.
- If the first line in the file contains the field name, select the First record for field name line.
- In a number of instances, missing values are converted in ASCII files to symbols other than a blank or zero (0), such as an asterisk (*). If you leave these symbols, *WizWhy* might read files that were originally numeric as alphanumeric instead. To avoid such a situation, click the Define command button in the Missing/Special Values block at the top of the dialog box. The String/Value Replacement dialog box will be displayed.
- In the String in original ASCII file text box, enter the character that represents a missing value in the ASCII file, and in the Replace with text box, enter a *numeric* value that *does not exist* elsewhere in the data and is quite distant in value from the other numeric values. For example, if an asterisk (*) represents a missing value in the ASCII file, and if all the numeric values are positive, you might enter the value -1000 in the Replace with text box. After you click OK, the replacement value that you entered will be displayed immediately in the table at the bottom.
- You can use this option to replace other characters as well.

Indicating lines for exclusion

Certain lines in the table – such as subtotals – may contain information that is irrelevant to the data analysis or that interferes with the parsing. These lines must be deleted at this stage.

To indicate lines for exclusion:

- Browse through the table to view all the lines.
- Click on the serial number (in the first column) of each line that you want to exclude from the data set. The text of that line will change color. To cancel your selection, click a second time on the line number.

Performing field parsing

The next step is to define the fields, or perform *parsing*. To assist you, *WizWhy* automatically identifies each field in the records and suggests the parsing, in order to create a column that represents a field. The procedure is as follows:

- Click the Parse command button at the top right. *WizWhy* will read the records and "reformat" (parse) the table to create and identify fields.

  The black vertical lines separate each field. Above the columns, *WizWhy* suggests one of three field types:

  QLT - Qualitative (categorical, alphanumeric characters)

  QNT - Quantitative (numeric data)

  D-M-Y - Date type (date formatted using numbers)

- Check the *WizWhy* parsing by scrolling horizontally through the records to view each field. To modify the parsing, you need to add or delete the vertical field separators.
- To *add a field separator,* click the pointer (↑) at the column position of the new separator. Click the body of the table and not in the header area. The new field heading (type) will be displayed as a

question mark (?), which means that this field type is yet unknown to *WizWhy*.

- To *indicate the field type* of the new column, you can either re-parse or indicate the type in the Column (field) block of the dialog box, as explained in the next step.
- To *delete a field separator,* simply bring the pointer on the separator that you wish to delete, click on the right mouse button, and select delete separator.
- To *move a field separator,* bring the pointer on the separator that you wish to move, and then either press on the left mouse button, and drag the separator while the button is pressed, or click on the right mouse button, and select one of the move commands.
- After creating new fields, you should perform parsing again. To do so, click the Parse command button to update the table. Note that after re-parsing, most question marks are replaced by the field types QLT, QNT or D-M-Y. Fields that remain undefined must be set "manually" in the Column (field) block of the dialog box.

Defining and modifying the field features

To modify the definition of a field, select the column by placing the pointer on the column header and highlighting it.

You may use the Column (field) block to perform three functions on the selected field:

- *Change or define the name of the column (field):* The field name will be printed in the table displayed in the Main dialog box of the *WizWhy* work area and later in the *WizWhy* reports. The default name for the field is a number, as in field1, field2 and so on. It is recommended to use meaningful names for each field, as this will facilitate reading the reports later on. Also, you must assign names to each new field, since any added field is unnamed. To name the selected field, type a header for it in the Name text box.
- *Change or define the field type:* You may:

Change a quantitative field (QNT) into qualitative (QLT)

Change a date field (D-M-Y) into qualitative

Change the date field format

To make your selection, click either the QNT or QLT radios buttons, or pull down the Date list and select D-M-Y format.

- *Exclude the selected field from the data:* Select the Do Not Import Column radio button.

If you decide to keep the field in the report, simply unclick the column.

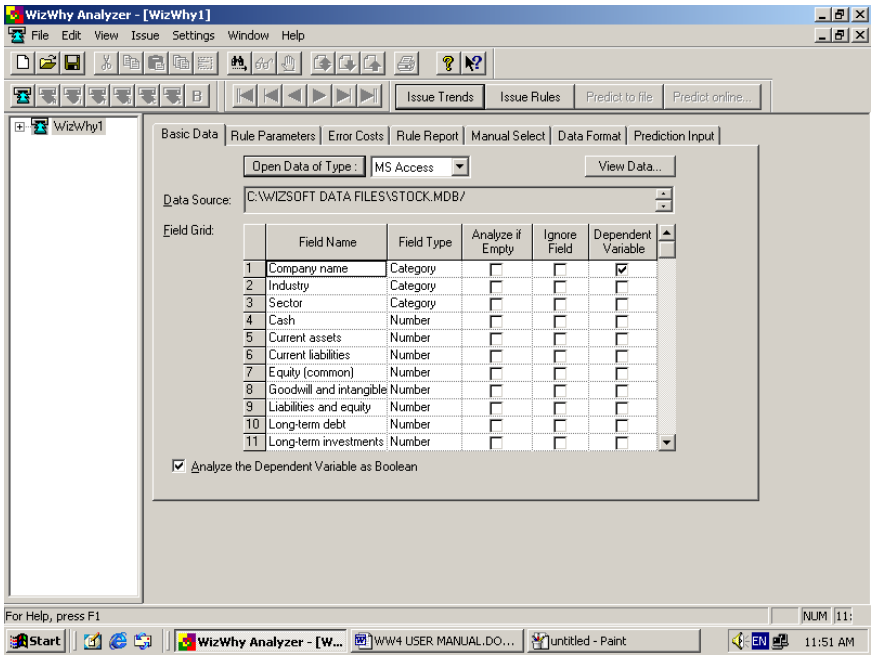
Be sure that you have indicated the field type of each column, so that no question marks remain along the top. In addition, be sure that each field has been assigned a name.

To cancel all formatting that you have performed, click the Reset button at the right. The only modifications that will remain are parsing lines.

Once you are satisfied with the parsing and field definitions, click OK. The ASCII file will be displayed as a data set table in the *WizWhy* work area.

6. Defining and Issuing the Reports

When you open a data set, it is displayed in the *WizWhy* work area, as follows:



Six tabbed dialog boxes are displayed to assist you in defining the parameters that will be used in analyzing the data and issuing the reports and the predictions.

- In the Basic Data dialog box, you select the *dependent variable* and define the parameters of the data set fields.

- In the Rule Parameters dialog box, you set the parameters of the rules to be revealed (for example, the minimum number of cases in a rule).
- In the Error Costs dialog box, you enter the cost of a miss versus a false alarm.
- In the Rule Report dialog box, you determine how many rules will be displayed, and whether positive or negative examples will be included in the Rule Report.
- In the Manual Select box, you can turn on the Manual Select stage, where you have the option to filter out if-then rules' conditions, and to take part in the selection of the if-and-only-if rules.
- The Data Format dialog box contains settings for the report format, printing, visual design and font selection.
- In the Prediction Input dialog box, you can select a data set that will be an input for issuing predictions. This tab should be filled in only when you wish to issue predictions in another file.

The sequence for issuing the reports and the predictions is as follows:

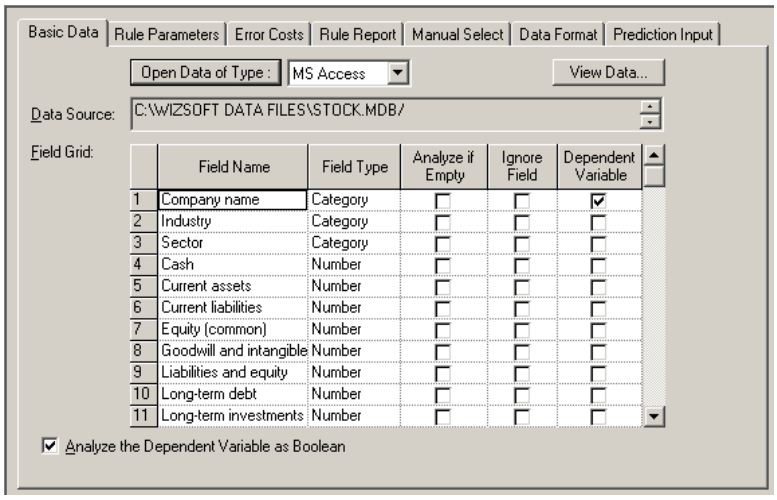
- Fill in the above mentioned tabbed dialog boxes.
- Click on one of the following buttons to instruct *WizWhy* what kind of reports will be issued.

Click on	**To**
Issue Rules	Issue *all* the reports
Issue Trends	Issue the Trend report only
Predict to File	To issue *all* the reports and update the predictions in another file.

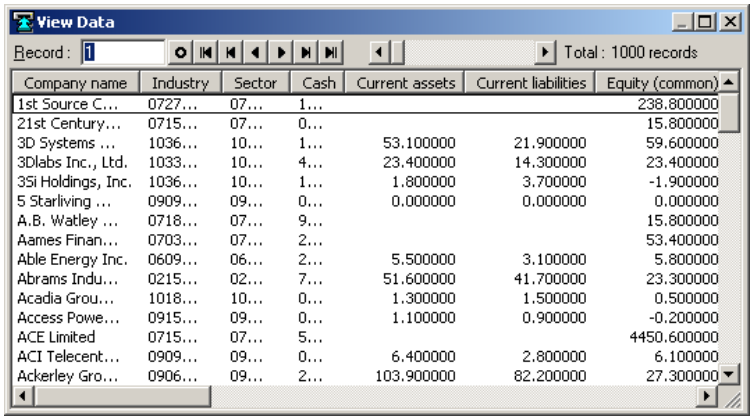
- Once the rules have been issued, you can go on and issue predictions on line and in another file (see chapter 8).

Basic Data

The Basic Data dialog box lists the fields of the data set. Each line refers to a single field.



You can click on the View Data button, if you wish to view the contents of the records. *WizWhy* displays the contents of the first 100 records. Click on the arrows on the top of the window to view additional records.



In the Dependent Variable column select the dependent variable, i.e., the field to be explained and predicted. You should select only one field. The other fields will be analyzed as condition fields (the independent variables).

The Field Name is the editable name to be printed in the reports. By default, *WizWhy* displays the name as it appears in the data. To edit the field name, click on it and type in a new name.

The Field Type column defines whether the field type is Category, Number or Date

- Category indicates alphanumeric data; for example, Item Number, Address, Reference Number, Occupation, or Yes/No fields.
- Number indicates numeric data to which mathematical computations can be applied; for example, Total $, Weight, Length or Percent.
- Date is the date data in one of the standard formats (D-M-Y, M-D-Y, etc.).

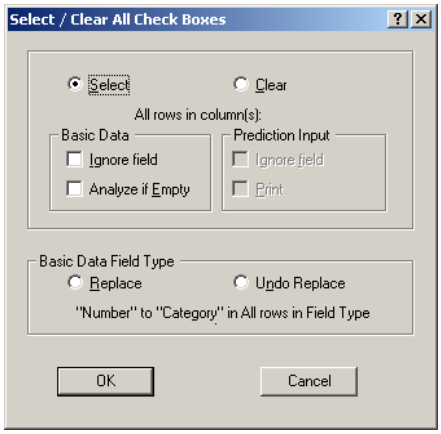
When *WizWhy* reads the data set, it assigns one of these types to each field. When a field contains letters, *WizWhy* determines that the field is categorical. When a field contains only digits, *WizWhy* assumes that the field is numeric. If a field containing only digits is in fact categorical, such as a phone number, change the field type.

If you are not sure whether a certain field should be analyzed as numeric or categorical, you can click on the View Data button, to see the contents of the records in the data.

Select the Analyze if Empty column to indicate that cases of missing values are informative, namely, when you wish *WizWhy* to look for rules such as: If Field A is empty, then the dependent variable is 1. Both numeric and categorical fields are considered empty if they contain no data. Numeric fields that contain 0 (zero) are not considered empty.

To exclude a field from the analysis, click in the Ignore Field column.

You can use the Update check boxes option from the Edit menu to update *all* the rows by one command: select or clear all the rows of the Analyze if Empty or Ignore Field columns, or change the field type in all the rows.



For example, to select all the rows in the Analyze if Empty column, select the Select and the Analyze if Empty check boxes. To change the Number field type into Category in all the rows, select Replace in the bottom pane. Select Clear to clear all the rows or Undo Replace to revert to the original field type.

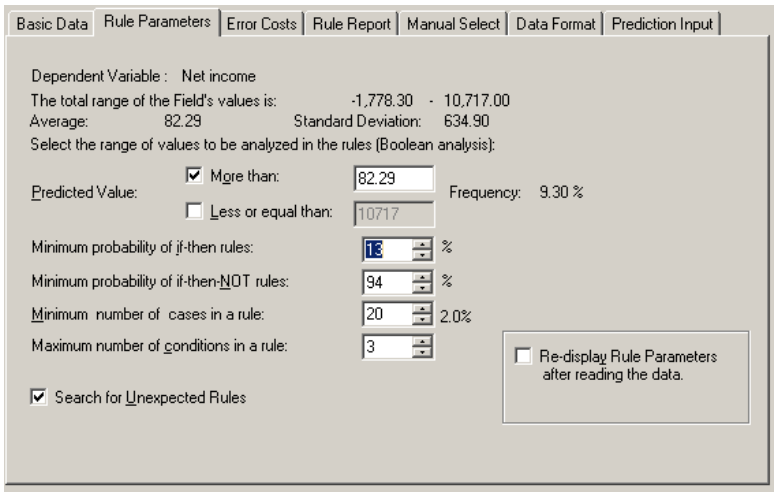
Select the Analyze the Dependent Variable as Boolean check box to analyze the dependent variable as if it has two values only. Clear this check box to analyze the dependent variable as multi-value (when it is a categorical field) or continuous (when it is numeric). When the analysis is Boolean, you select one value (when the field is categorical) or one range of values (when the field is numeric). *WizWhy* will reveal the patterns predicting when the selected value (or range of values) holds or does not

hold. When the analysis is not Boolean, *WizWhy* will reveal such patterns for each value (if the dependent variable is categorical) or for each range of values (if the dependent variable is numeric).

Rule parameters

In the Rule Parameters dialog box you can fine-tune the following parameters of the if-then rules to be discovered. You can select -

- The value (or range of values) to be analyzed (when the dependent variable is analyzed as Boolean).
- The minimum probability of the if-then rules (when the dependent variable is analyzed as Boolean).
- The minimum number of cases in a rule.
- The maximum number of condition in a rule.



Determining the value to be analyzed

If you selected to analyze the dependent variable as Boolean, you now have to select the value to be analyzed as Boolean. This is done in the Predicted Value line. (If the dependent variable is not analyzed as Boolean, *WizWhy* analyzes all the values of the dependent variable, and therefore this line is inactive).

When the dependent variable is categorical (alphanumeric), *WizWhy* lists the values of the field, and you have to select one of them. *WizWhy* will use this value for the Boolean analysis. It will reveal the rules describing the conditions when this value holds or does not hold. The values are limited to those having a frequency higher than the minimum cases in a rule (see page 49).

When the dependent variable is a numeric field, you have to select a range of values as the Boolean value to be analyzed. For example if you analyze a company data set, where the dependent variable is the annual sales, you can select the range of "more than 1,000,000" in order to reveal the patterns of the successful companies.

WizWhy displays the total range of the dependent variable's values, along with the Average and the Standard Deviation. From the displayed information, you should define the Range of values to be analyzed by clicking the More than check box and typing in the lowest value of the range and/or clicking the Less than check box and typing in the highest value of the range.

Once you select the value (or range) to be analyzed, *WizWhy* prints its % frequency. This % frequency will be relevant later, when you determine the minimum probability of the if-then rules.

If the dependent variable is a categorical field, *WizWhy* reads the data when you select the value to be analyzed, to calculate its % frequency. If the dependent variable is a numeric field, *WizWhy* reads the data twice, first to calculate the range of values, average and standard deviation, and later, when you select the range of values to be analyzed, to calculate the % frequency. When *WizWhy* reads the data, a progress bar is displayed. If the file contains many records, reading might take a rather long time. You can stop the reading by clicking on the Cancel button. *WizWhy* then calculates on the basis of the records read till the Cancel button had been

clicked. You can instruct *WizWhy* to read the data once again and re-calculate by clicking on the Refresh statistical data button.

Determining the rules' minimum probability

WizWhy reveals both if-then and if-then-not rules. If-then rules have a structure such as the following:

> *If* City *is* New-York
> *and* Amount Purchased *is* 200 … 300 (average = 250)
> *and* Salesperson *is* Dave
> *Then*
> Growth Since Last Year *is* less than 1%
>
> *Rule's probability*: 0.70
> *The rule exists in* 370 *records*
> *Significance Level: Error probability* < 0.001

If-then-*not* rules have the same structure except for the result line, which is negative. For example:

> Growth Since Last Year *is not* less than 1%

"Rule's Probability indicates the percentage of cases in which both the "if" *and* the "then" sections hold, within the total number of cases in which the "if" sections holds. For example, if the rule's conditions appear in 20 records and the "then" section occurs in only 19 of those records, then the rule's probability is 95%. Some data mining literature uses the term "Confidence Level" rather than "Rule's Probability." However, the two terms are synonymous.

You can determine the threshold of the rule's probability in the Minimum probability of if-then rules, and the Minimum probability of if-then-NOT rules lines.

There is no limit to the lowest acceptable minimum probability. However, when entering these values, you should refer to the basic frequency of the value you are analyzing. For example, if the basic percentage frequency

of a value to be analyzed is 10%, it is recommended to look for if-then rules in which minimum probability is 14% (40% more than the basic frequency) and if-then-*not* rules in which the minimum probability is 94% (40% less than the basic frequency). These are *WizWhy* default values. Note that if-then rules in which the rule probability is 10% (or if-then-*not* rules in which the rule probability is 90%) are *un*informative.

> **Tip**: The lower the minimum probability, the more rules *WizWhy* will find.

After selecting the predicted value, *WizWhy* displays its basic frequency. In order to save time, *WizWhy* estimates the basic frequency by reading the first 1000 records. In the lower right corner you can instruct *WizWhy* to Re-display Rule Parameters after reading the data. That is, during the issuing rules process, after reading the whole data set, display a window containing once more, the primary probability of the predicted value and the minimum probability of the if-then and if-then-not rules. At this stage the calculation of the primary probability of the predicted value is accurate, and therefore you can check and correct the minimum probability thresholds if needed.

Note that the minimum probability lines are active only when the dependent variable is analyzed as Boolean. When the dependent variable is analyzed as multi-value the minimum probability (of both if-then and if-then-not rules) should be determined for each of the values. As determining it manually might be quite tedious, *WizWhy* selects the default minimum probabilities for each value (40% more and 40% less than the basic frequency).

Minimum number of cases in a rule

On the Minimum number of cases in a rule line, define the minimum number of cases required to establish a rule. For example, if you select 50, *WizWhy* will establish a rule on condition that it holds for at least 50 records in the data set. This means that at least 50 records satisfy both the rule's conditions and conclusion. The lowest number of cases is 4 (since a rule holding for less than 4 cases might be accidental). Some data mining

literature uses the term "Support Level" rather than "Number of cases in a rule." However, the two terms are synonymous.

Tip: The lower the minimum number of cases in a rule, the more rules *WizWhy* will find.

Maximum number of conditions

On the Maximum number of conditions in a rule line, you can define the maximum number of the "if" conditions in a rule. For example, if you select 3, *WizWhy* will reveal all the rules having one, two or three conditions, and then stop the search. Since each condition refers to a different field, a rule having 3 conditions has following structure:

If field A is . . . and field B is . . . and field C is . . . then…

Tip: The higher the maximum number of conditions, the more rules *WizWhy* will find.

Defining the minimum number of cases in a rule in conjunction with the minimum probability and the maximum number of conditions is a powerful tool for increasing or decreasing the calculation time and the number of rules that *WizWhy* finds in the data.

Searching for unexpected rules

An unexpected rule is an if-then rule having at least two conditions, that is an unlikely event relative to more basic rules having a fewer number of conditions (see pages 81 - 86 for more details). The search for these rules might take a rather long time.

Clear the Search for Unexpected Rules check box, to instruct *WizWhy* to skip the stage of searching these rules.

Note that this option is relevant only when the dependent variable is analyzed as Boolean. If it is not analyzed as Boolean, *WizWhy* does not search for unexpected rules.

Tip: If you have selected the check box, you can still instruct *WizWhy* to skip this stage, when *WizWhy* searches for the rules (see page 60).

Error costs

In the Error Costs tab you can enter the cost of a miss versus a false alarm.

Both misses and false alarms refer to prediction errors. For example, when diagnosing a patient, if one predicts that the patient does not suffer from a certain disease, and it turns out that he or she does, the diagnosis (or prediction) is considered as a *miss* (the diagnosis missed the disease). On the other hand, if one predicts that the patient suffers from the disease although in fact he or she does not, the diagnosis is an example of a *false alarm*. Some data mining literature uses the terms "false positive" rather than "false alarms" and "false negative" rather than "misses". However, these terms are synonymous.

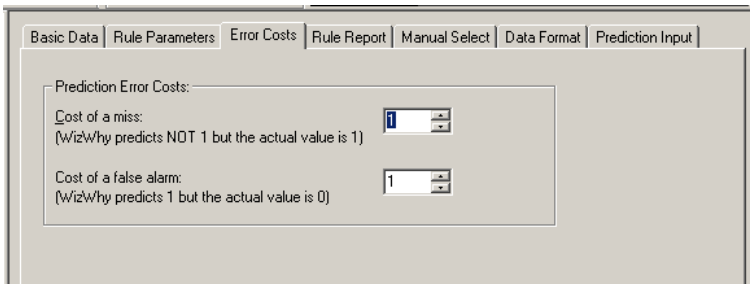
#	**The prediction**	**The actual value**	**Error type**
1	Sick	Sick	--
2	Sick	Not sick	False alarm
3	Not sick	Sick	Miss
4	Not sick	Not sick	--

In many cases the cost of a miss is different from the cost of a false alarm. For example, missing to diagnose a disease might be much worse than a false alarm.

In some cases, such as the above example, the error costs are determined on rather subjective grounds. There are, however, cases where the error costs can be determined accurately. Consider for example direct marketing. Suppose the cost of an envelope (together with the other related costs) is $1, while the profit on selling the product is $100. In such a case the cost of a miss is 100, while the cost of a false alarm is 1.

When issuing predictions *WizWhy* can trade off between misses and false alarms. The cost of errors entered in the Error Costs dialog box tells *WizWhy* what is the trade off.

When the dependent variable is analyzed as Boolean, *WizWhy* displays the following dialog box where you can enter the cost of a miss and the cost of a false alarm.

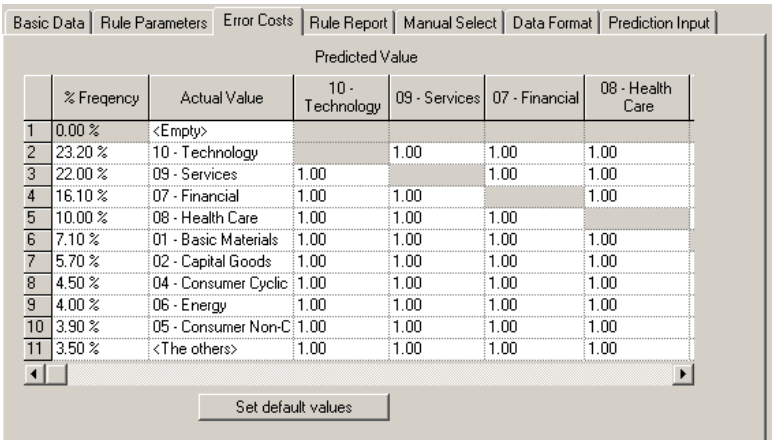


When the dependent variable is categorical and analyzed as multi-value, there are several possible misses and several possible false alarms. *WizWhy* displays the following dialog box, which lists all the values under analysis. You can enter the cost of each possible miss and each possible false alarm. The columns refer to the predicted values, and the rows refer to the actual values. The error costs should be entered in the cells. For example, the cell in the first column and the second row contains the cost of the error of predicting the value of this column, while the actual value is the one related to the row.

By default, all the error costs are 1, that is, the cost of every false alarm and error is equal. If you click on the Set default values button, *WizWhy* enters 1 in all the cells.

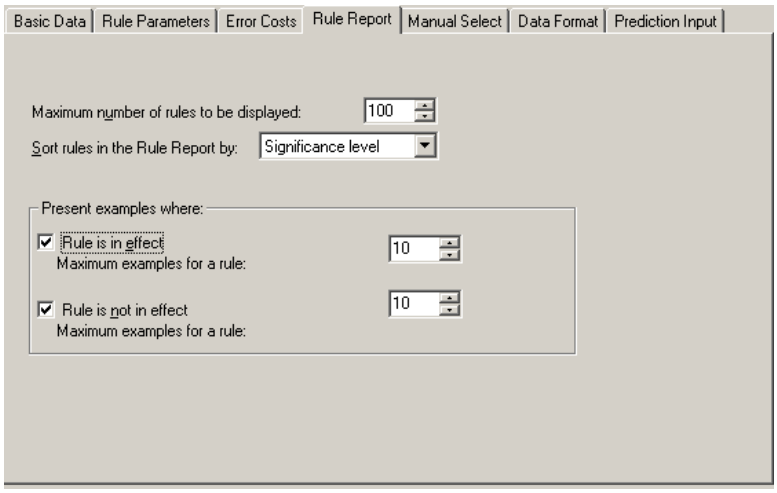
> **Tip**: When all the error costs are equal, the number of errors in the data under analysis is minimized.

When the dependent variable is continuous and analyzed as non-Boolean, the error costs are automatically set to the default value.



Rule report

The If-Then Rule Report contains the if-then rules. In the Rule Report tab you determine how this report will be displayed.



Maximum number of rules

On the Maximum number of rules to be displayed line, you determine the maximum number of rules that will be displayed on the screen in the If-Then Rule Report. For example, if you enter 100, *WizWhy* will display in the report just the first 100 rules (according to the sort determined in the next line). Because *WizWhy* might reveal several thousands rules, and as reviewing all of them is impractical, it makes sense to limit the number of the displayed rules.

No matter what the number you enter in this line is, *WizWhy* always saves all the rules. The number entered in this line refers only to the rules that will be displayed. All the rules are taken into account when *WizWhy* issues predictions and reveals unexpected phenomena, and all the rules can be printed on the printer or exported as an ASCII or DAO (Microsoft Access) table (see page 76 for more details).

Sorting the rules

On the Sort rules in the Rule Report by line, select the criterion for sorting the rules in the If-Then Rule Report. The options are:

- Significance Level, which is the probability that the rule exists accidentally in the data under analysis. (See page 73 for more details).
- % Probability, which is the percentage of cases in which both the "if" *and* the "then" sections hold, within the total number of cases in which the "if" sections holds. (See page 73 for more details).
- No. of Cases in a Rule, which is the number of records where the rule holds. (See page 73 for more details).

Displaying examples

Use the Present Examples block to determine which, if any, examples should be provided in the If-Then Rule Report. *WizWhy* can print two lists of examples after each rule:

- Those in which the rule holds (positive examples): Rule is in effect

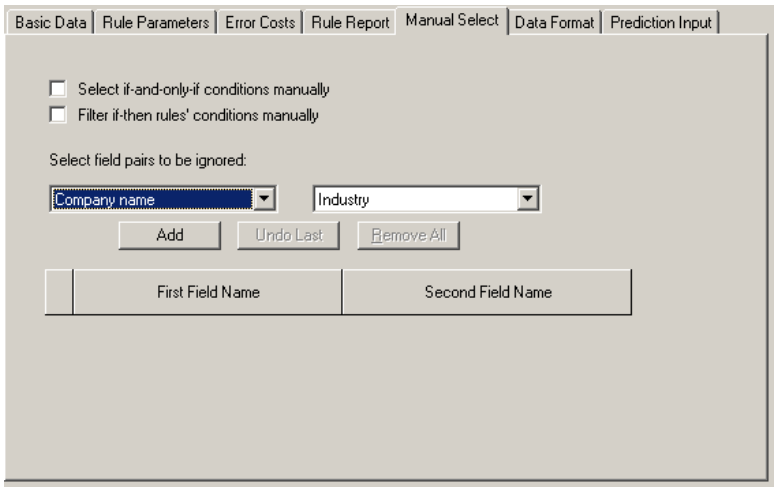
- Those in which the rule does not hold (negative examples, when the rule probability is less than 100%): Rule is not in effect

The examples in both lists are printed per record, according to the records' serial numbers. When the report is displayed in the *WizWhy* viewer, you can display the contents of each record by double clicking on the record number.

If you select either option, in the Maximum examples for a rule text box, enter the maximum length of each list of examples. The default is 10. This means that a maximum of 10 records' numbers of examples will be printed. If the number of actual examples is greater than 10, *WizWhy* will select the examples to be printed at random.

Manual Select

WizWhy reveals the rules automatically. However, in the Manual Select tab you can instruct *WizWhy* to let you manually control the rule revealing process.



Select the Select if-and-only-if conditions manually line, to control the issuing of the if-and-only-if rules. When *WizWhy* will reach the stage of revealing the if-and-only-if rule, it will open a window where you will be able to control the selection of the conditions composing the if-and-only-if rules.

Select the Filter if-then rules' conditions manually line, to control the issuing of the if-then rules. When *WizWhy* will complete the stage of building the two conditions table, it will open a window where you will be able to select field pairs. Any rule containing one of these field pairs will be filtered out. The idea behind this process is to give you an option to define non-interesting rules and to filter them out.

You can select these pairs in the current window. However, since the number of possible pairs might be very large, it makes sense to postpone it to the stage of revealing the two conditions rules. At this stage *WizWhy* will display only those pairs holding for at least N records, where N is the minimum number of cases in a rule. All other pairs cannot be included in the rules, and therefore there is no need to review them.

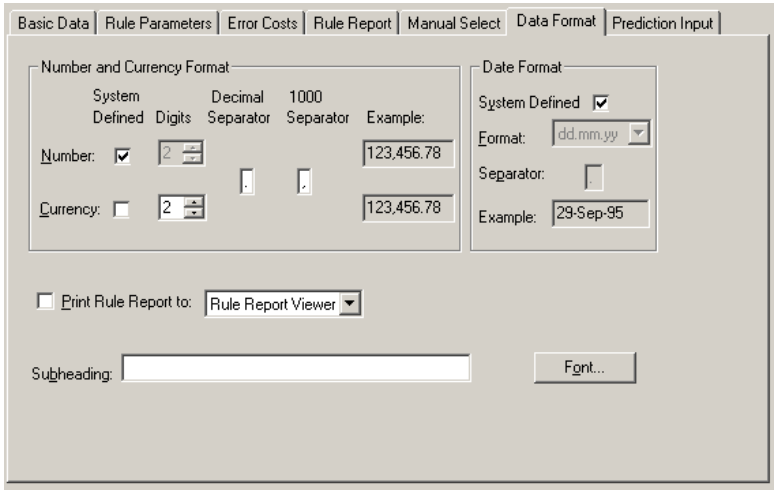
As mentioned, you can select field pairs in the current screen as well. The selection is done in the two list boxes at the center of the screen. Each box contains the data set fields. To select a field pair -

- Select the first field of the pair from the left list box.
- *WizWhy* will display the next fields in the right list box. Select the second field from this list box.
- Click on the Add button. The field pair will be displayed at the bottom part of the screen. Any rule containing this field pair will be filtered out.
- Repeat the above steps to add other pairs.

In case you selected a wrong pair, click on the Undo Last button to delete the last selected pair, or click on the Remove All button to restart.

Data format

The Data Format dialog box contains settings for the report format, printing, visual design and font selection, which determine the overall appearance in which the data is presented.



Number

In the Number and Currency Format block, you can change the formats for currency and numbers. System Defined (default) are the defaults set in the Regional Settings option of the Control Panel.

To change the format:

Clear the relevant check box (Number and/or Currency).

In the Digits text box, indicate the number of digits to be displayed after the decimal separator.

In the Decimal Separator box, type the character that will be used as the decimal point.

In the 1000 Separator box, type the character that should be used as the thousands separator.

The Example box will illustrate the result of your selections.

Date

In the Date Format block at the right, you can change the format for dates.

Clear the System Defined check box.

In the Format box, select the desired date format from the drop-down list box.

In the Separator box, type in the character that should be used to separate the day, month and year.

The Example box will illustrate the result of your selections.

Print rule report to

By default *WizWhy* displays the If-Then Rule Report on the screen. To change it:

- Select the Print Rule Report to check box.
- In the Print Rule Report to list box, select one of the following options to indicate where the If-Then Rule Report should be printed:

  Rule Report Viewer (on screen)
  Rule Report Prints (printer)
  ASCII Text File
  RTF File (Rich Text Format file)
  MS Access File (Microsoft's Access table)

Note that when the If-Then Rule Report is displayed on the screen you can still instruct *WizWhy* to send it to one of these options (see pages 76 - 77).

Heading

In the Subheading text box, you may type a title that will appear at the beginning of the reports.

Font

To change the font type, style and/or size, click the Font command button to display the standard Font dialog box.

Select the Font, Font style and Size and click OK. The reports will display the data in the font and styles selected.

Prediction input

In the Prediction Input tab, you can select a file that will be an input for issuing predictions. This dialog box should be filled in only when you wish to issue predictions in a different file. You can select the file either before or after issuing the rules. For more details see chapter 8.

Issuing the rules

Once all the parameters are set, you can generate the rules and issue the predictions. To do so click on one of the following buttons in the action bar.

- Click the Issue Rules button to generate all the *WizWhy* reports. (See the next chapter for explanations about the *WizWhy* reports).
- Click the Predict to File button to generate all the reports and update predictions in the file you selected in the Prediction Input dialog box (if you did not selected a file in this dialog box, the Predict to File button is inactive).

- Click the Issue Trends button if you are interested in issuing the Trend Report only. (See pages 78 - 81 for more details about this report).

WizWhy will display a progress bar where you can see the process stage and the % progress.

The process of issuing rules is composed of various stages:

The Reading phase, where *WizWhy* reads the data and builds tables needed for calculations.

The Calculation phase, where *WizWhy* builds the basic tables.

The Building 2-conditions tables phase, where *WizWhy* builds the tables needed for revealing rules having 2 conditions (in the "if" part of the if-then rules).

The Revealing 2-conditions rules phase, where *WizWhy* searches for the rules having 2 conditions (in the "if" part of the if-then rules).

The Building and Revealing stages continue for 3, 4 and more conditions. The maximum number of conditions is determined in the Rule Parameters dialog box (see page 50).

The Deleting Redundant Rules phase where *WizWhy* deletes rules having N conditions that can be explained by simpler rules having N-1 conditions.

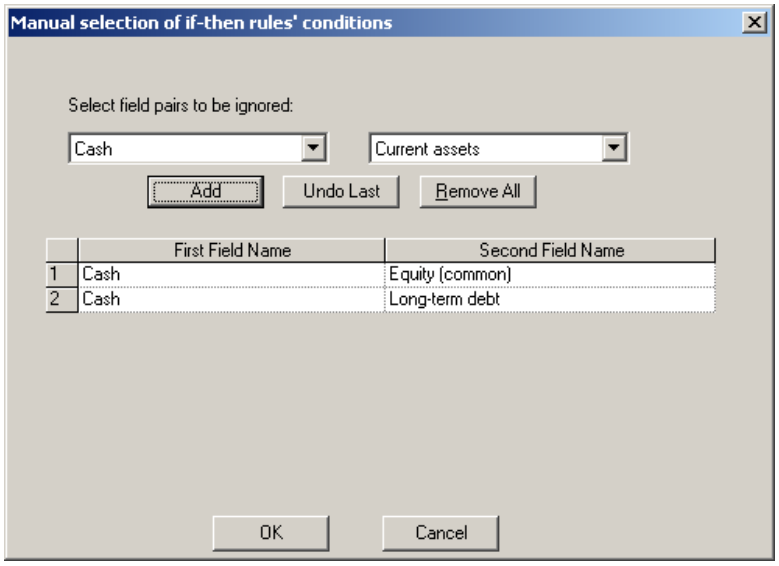
The Update Rule Parameters and Validating phases, where *WizWhy* reviews the rules prediction power.

In some stages you can click on the Move Forward button to instruct *WizWhy* to move on to the next stage (without completing the current stage).

You can stop the process by clicking the Cancel button.

Manual filtering of if-then rules' conditions

If you selected the option of filtering the if-then rules (in the Manual Select tab, see pages 55 - 56), then after building the 2 conditions table, *WizWhy* opens a window such as the following:



In this window you can select field pairs. Any rule containing one of these field pairs will be filtered out. The idea behind this process is to give you an option to define non-interesting rules and to filter them out.

Note that at this stage *WizWhy* displays only those pairs holding for at least N records, where N is the minimum number of cases in a rule. All other pairs cannot be included in the rules, and therefore there is no need to review them.

Each of the two list boxes contains the data set fields. To select a field pair -

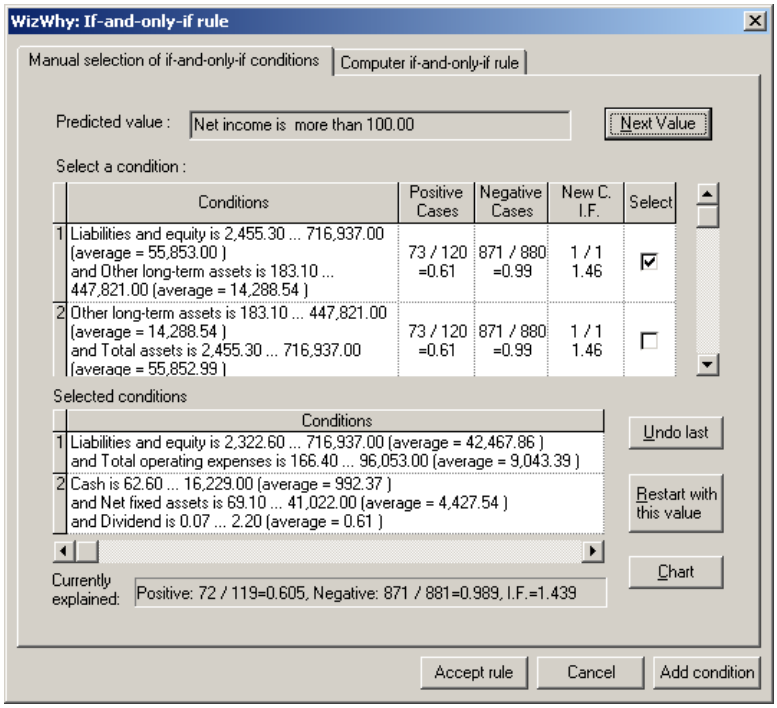
- Select the first field of the pair from the left list box.
- *WizWhy* will display the next fields in the right list box. Select the second field from this list box.

- Click on the Add button. The field pair will be displayed at the bottom part of the screen. Any rule containing this field pair will be filtered out.
- Repeat the above steps to add other pairs.
- Click on the OK button to continue.

In case you selected a wrong pair, click on the Undo Last button to delete the last selected pair, or click on the Remove All button to restart.

Selecting the if-and-only-if conditions

If you selected the option of manually issuing the if-and-only-if rules (in the Manual Select tab, see pages 55 - 56), then during the Rule Validation stage, *WizWhy* opens a window for selecting the conditions of the if-and-only-if rules.



Each screen refers to one if-and-only-if rule. The subject of the if-and-only-if rule is displayed in the Predicted Value line.

The if-and-only-if rule recommended by *WizWhy* is displayed in the Computer if-and-only-if rule tab. You can select the *WizWhy* recommended if-and-only-if rule or replace it with another if-and-only-if rule that you issued in the Manual selection of if-and-only-if conditions tab.

What is an if-and-only-if rule?

When the dependent variable is analyzed as Boolean, *WizWhy* searches two kinds of if-and-only-if rules. The first if-and-only-if rule lists the conditions that meet the following requirement: If at least one of the conditions holds, there is a high probability that the dependent variable's value is *r*, and if none of them holds, there is a high probability that the dependent variable's value is *not-r*. The other if-and-only-if rule lists the conditions that meet the opposite requirement: If at least one of them holds the dependent variable's value is *not-r*, and if none of them holds, the dependent variable's value is *r*.

When the dependent variable is *not* analyzed as Boolean, *WizWhy* applies the above-mentioned search for each of the dependent variable's values. In other words, for each value of the dependent variable *WizWhy* searches two sets of necessary and sufficient conditions. For example, if one of the values is *s*, *WizWhy* searches one set that refers to *s*, and another set that refers to *not-s*.

For more explanations about if-and-only-if rules, see pages 87 - 92.

At the bottom of the *Computer* if-and-only-if tab *WizWhy* displays some parameters in regard to the if-and-only-if rule. These parameters can help you when considering to what extent the if-and-only-if rule explains the cases.

If the value under analysis is, for example, *r*, *WizWhy* lists the following lines:

In the first item, *WizWhy* displays the probability that *r* holds, if at least one of the conditions holds, that is, if the first condition holds, *or* the second condition holds and so on. *WizWhy* also prints the total number of the cases entailing this probability, that is the total number of cases where at least one of the conditions holds, and the total number of cases where at least one of the conditions holds, and R holds as well.

In the second item, *WizWhy* displays the probability that *not-r* holds if all the conditions do not hold, that if the first condition does *not* hold, *and* the second condition does *not* hold and so on. *WizWhy* also prints the relevant number of cases.

In the Total number of cases lines, *WizWhy* prints the sum of the numbers presented above: the sum of the positive and negative cases explained by the conditions, and the total number of the positive and negative cases. By dividing these two numbers *WizWhy* calculates the success rate. This number denotes to what extent the conditions correctly explain the predicted value.

The next line lists the primary positive and negative probabilities of the predicted value (note that the sum of the two probabilities is 1).

Finally, the Improvement Factor line denotes to what extent the if-and-only-if rule explains the dependent variable values relative to an explanation based on the primary probability only.

Issuing the if-and-only-if rule

In the Manual selection of if-and-only-if conditions tab *WizWhy* displays list of conditions that are candidates to be included in the if-and-only-if rule. You have to select one of these conditions. *WizWhy* then will display a second list of conditions that are candidates as additional conditions for the if-and-only-if rule, and so on, till you either determine that the rule is complete, or there are no conditions that can improve the rule.

To select a condition, check the Select column, and click on the Add Condition button.

To confirm the if-and-only-if rule, click on the Accept Rule button. The rule you issued will replace the *WizWhy* recommended rule. *WizWhy* will continue by displaying the next predicted value, or move on to completing the issuing reports process.

If you prefer to leave the *WizWhy* recommended if-and-only-if rule, click on the Next Value or the Cancel buttons. *WizWhy* will ignore the rule issued manually. Once again, *WizWhy* will continue by displaying the next predicted value, or move on to completing the issuing reports process.

In many cases *WizWhy* will not display all possible predicted values. During the revealing of the if-then rules, *WizWhy* checks which of these rule conditions can be used as necessary and sufficient conditions. The list of predicted values is limited to those having at least one candidate for being a necessary and sufficient condition.

When selecting the conditions you can move one step backward by clicking on the Undo last button, or restart selecting the conditions (of the current rule) by clicking on the Restart with this value button.

Note: Contrary to the selection of conditions, once you select a *rule,* you cannot undo the selection.

In order to help you in selecting between the candidate conditions, the following information is presented for each condition:

The Conditions column displays the contents of the condition.

The Positive Cases column refers to the records where the condition holds. 3 numbers are displayed, for example: 180/200 = 0.9. These numbers denote that out of the 200 records, where the condition holds, in 180 records the predicted value holds as well. That is, if the condition holds, there is a 0.9 probability that the predicted value holds.

The Negative Cases column refers to the records where the condition does *not* hold. Once again, 3 numbers are displayed, for example: 9,800/10,000 = 0.98. These numbers denote that out of the 10,000 records, where the condition does *not* hold, in 9,800 records the predicted

value does *not* hold as well. That is, if the condition does *not* hold, there is a 0.98 probability that the predicted value does *not* hold.

The third column, New C. / I.F. denotes the number of *new* cases (New C) explained by the condition, and the improvement factor (I.F.). For example, if the numbers in the first line (new cases) are 30/50 it means that out of the 50 cases not explained by the previous conditions 30 new cases are explained by the condition under consideration. The third number (the improvement factor) denotes the expected improvement factor of the if-and-only-if rule composed of the previously selected conditions together with the condition under consideration.

At bottom of the window at the Currently explained box, *WizWhy* summarizes how well the selected conditions (together) explain the data.

When considering a new condition, you can click on the Chart button, to view a graphical presentation of the relation between this condition and the previously selected ones. See pages 90 - 91 for more explanations about this chart.

7. Reviewing the Reports

The *WizWhy* viewer displays six reports:

- The Summary Report contains an analysis of the rules' *explanatory power*.
- The If-Then Rule Report lists the discovered *if-then rules*.
- The Trend Report presents graphically and textually the one-condition trends in the data. These trends *summarize* the data.
- The Unexpected Rule Report displays the rules that are unexpected relative to more basic rules and trends. These unexpected rules describe *interesting* phenomena in the data.
- The If-and-only-if Rules Report lists the *necessary and sufficient conditions* (i.e., the if-and-only-if rules).
- The Unexpected Cases Report displays the records where the dependent variable's value *deviates* from the expected value according to the discovered rules.

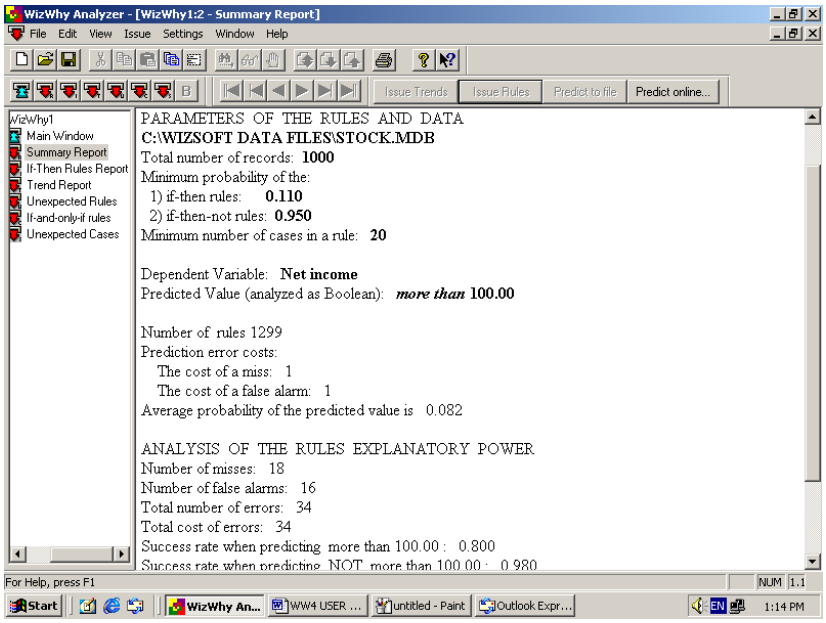
When the dependent variable is *not* analyzed as Boolean, the following reports are *not* issued:

- The Trend Report.
- The Unexpected Rule Report.

All the reports can be printed on the printers and exported as ASCII files, RTF (Rich Text Format) files or Microsoft Access (mdb) tables. To print or export a report, open the report and click on the printer icon.

The Summary Report

The Summary Report displays an analysis of the rules explanatory power. It contains two sub-sections:



The Parameters of the Rules and Data sub-section, is where *WizWhy* lists the parameters of the analyzed data together with the parameters that you entered. These are the parameters according to which the rules were issued.

The Analysis of the Rules Explanatory Power sub-section, is where *WizWhy* analyzes to what extent the discovered rules explain the data. For each record in the basic data, *WizWhy* reads the field values and applies the relevant rules to deduce the outcome entailed by the rules. In other words, *WizWhy* issues predictions in regard to the records of the data set that was analyzed for revealing the rules. At this stage *WizWhy* also reveals which set of rules explain the dependent variable most accurately -

the if-then rules or the if-and-only-if rules. The final predictions are based on the most accurate set. *WizWhy* then checks for each record whether the dependent variable's actual value matches the predicted value. All the records where the predicted value fits the actual value are explained by the discovered rules, while all the records where the predictions fail are not explained by the rules. This section summarizes the results of these internal predictions.

Analysis of the Rule Explanatory Power

The Analysis of the Rule Explanatory Power section contains the following items:

Decision Point: This line is printed only if (1) the dependent variable was analyzed as Boolean, and (2) The final prediction was based on the if-then rules (and not the if-and-only-if rules). When *WizWhy* issues a prediction on the basis of if-then rules, it calculates the probability that the value of the dependent variable in the record under analysis is 1 (assuming that 1 is the predicted value). This probability is called the *Conclusive Probability*. Now, when the analysis is Boolean, there should be a decision point, that distinguishes between the cases where the dependent variable is 1 and the cases where it is not 1. The number printed in this line signifies this decision point: When the conclusive probability is above the decision point *WizWhy* predicts that the dependent variable is 1, and when the conclusive probability is below the decision point, *WizWhy* predicts that it is not 1. When calculating the decision point *WizWhy* takes into account the cost of errors (the cost of a miss versus the cost of a false alarm).

Number of Misses: Cases where *WizWhy* predicts R while the actual value is *not*-R.

Number of False Alarms: Cases where *WizWhy* predicts *not*-R, while the actual value is R.

Total Number of Errors: The sum of the above two numbers.

Total Cost of Errors: The number of misses multiplied by the cost of a miss plus the number of false alarms multiplied by the cost of a false

alarm. When the dependent variable is continuous and the analysis is not Boolean, the user does not enter error costs. In this case the values of the dependent variable are automatically segmented into intervals, and the error costs are calculated according to the absolute difference between the actual range and the predicted one. For example, if the actual value of the dependent variable is 30, and this value falls in the second interval, while the value predicted by *WizWhy* is 70, and this value falls in the fifth intervals, the cost of this miss is 3.

Success Rate when predicting 1: The % success of the positive predictions. This number is calculated from the number of misses and the total number of positive predictions. This line is printed only if the analysis was Boolean.

Success Rate when predicting NOT 1: The % success of the negative predictions. Like the previous line, this line is printed only if the analysis was Boolean.

The proportion between the success rate of the positive and negative predictions is the result of the proportion between the price of a miss versus the price of a false alarm.

Number of Records with No Relevant Rules: The number of records where no rule applies. The field values in these records meet none of the conditions in the rules. When *WizWhy* issues predictions in regard to these records, it determines the prediction on the basis of the % frequency of the predicted value and the error costs. The number of records with no relevant rules can usually be reduced by reducing the minimum number of cases in a rule, and the minimum probability of if-then and if-then-not rules.

Average Error Cost (per record): The total cost of errors divided by the number of records in the data. When issuing predictions *WizWhy*'s object is to minimize this number.

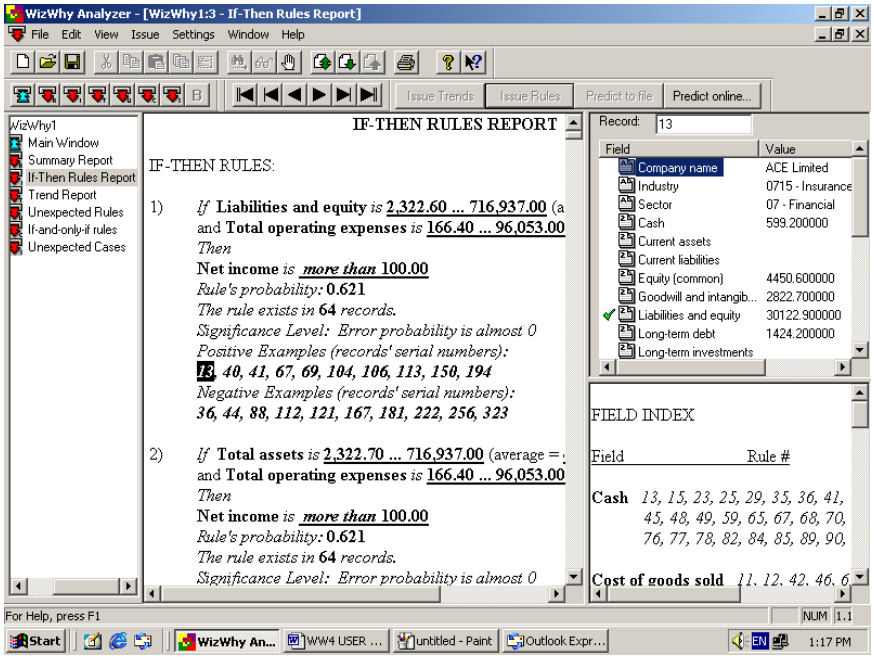
Expected Average Error Cost (per record): The average error cost resulting from issuing predictions on the basis of the % frequency of the predicted value, the cost of a miss, and the cost of a false alarms only. In other words, this is the expected average error cost when no rule is known. For example, if the frequency of the predicted value is 15%, and a cost of a miss is 2, while the cost of false alarm is 1, the best way for

issuing predictions when no rule is known, is to predict that the predicted value holds in *none* of the records. Such a prediction will result in misses in 15% of the records, no false alarms, and since the cost of a miss is 2, the average cost per record will be 0.3. This is the expected average error cost.

Improvement Factor: The expected error cost divided by the average error cost. This number signifies the contribution of the rules beyond what is known without the rules.

The If-Then Rule Report

The If-Then Rule Report displays the discovered rules. The screen is divided into three panes.



The left pane, the Rule List, lists the rules.

The top right pane displays the Record Details Grid, which shows the contents of each record, per field.

The bottom right pane, the Field Index, lists the fields that appear in the rules found.

> **Tip**: You can move the vertical bar between the Rule List and the two right panes – and the horizontal bar between the Record Details Grid and the Field Index – to display the various lists more clearly.

The rule list

The left pane of the Rule report window contains the discovered if-then rules.

All the rules relate between the dependent variable and the other fields by means of if-then statements, such as:

> *If* Field of business *is* High Tech
> *and* Number of Employees *is* 200…230 (Average = 215)
> *and* Annual Sales *is* 38,000…42,000 (Average = 40,000)
> *then*
> Company Value *is* More than 500,000
>
> *Rule's Probability:* 0.90
> *The Rule exists in* 370 *records*
> *Significance Level: Error Probability* <0.01
>
> *Positive Examples (record's serial number):*
> 4, 11, 17, 25, 38, 45, 87, 100, 106, 116
> *Negative Examples (record's serial number):*
> 14, 15, 27, 63, 67, 70, 96, 124, 129,139

The rules are either if-then rules (such as the example above) or if-then-*not* rules. If the example above had been an example of an if-then-*not* rule, the conclusion would have been: Company Value is *not* more than 500,000.

Each condition in the rule refers to a single field. The conditions are linked by an "and."

When the condition refers to a numeric field (such as the second and third conditions in the example above), the value is an interval. The intervals in the number fields are automatically determined by *WizWhy*.

Rule probability

The probability of a rule indicates the number of cases in which both the "if" *and* the "then" sections hold, within the total number of cases in which the "if" sections holds. The rule probability is equal to or higher than the Minimum probability of the if-then rules and the Minimum probability of the if-then-*not* rules, as determined in the Rule Parameters dialog box (see pages 48 - 49).

The term "Rule Probability" is synonymous with "Rule Confidence Level" that is used in many data mining textbooks and papers.

Rule exists in . . .

This line indicates the number of records in which the if-then rule exists; that is, the number of records in which both the "if" *and* the "then" clause of the rule applies. This number is equal to or higher than the number entered in the Minimum number of cases in a rule line in the Rule Type dialog box (see page 49).

The term "Number of cases in a rule" is synonymous with "Rule Support Level" that is used in many data mining textbooks and papers.

Error probability and significance level

The error probability of a rule indicates the percent chance that the rule under discussion exists accidentally in the data. The significance level of the rule equals 1 *minus* the error probability. The lower the error probability, the higher the degree the rule can be "trusted" when issuing a prediction.

The concept of the error probability is similar to the concept of the α probability (confidence level) in classical statistical tests.

The Rule Report displays rules whose error probability is less than 0.5; when the error probability equals 0.5, the probability that the rule exists in the data accidentally equals the probability that it does *not* exist accidentally; therefore such a rule is irrelevant for issuing explanations and predictions.

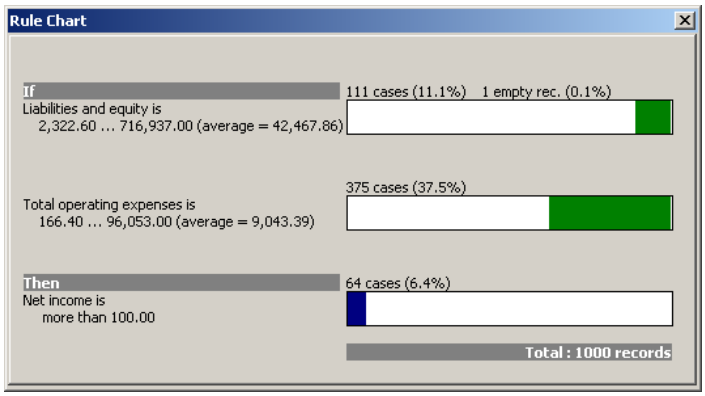
Positive examples / negative examples

If you selected one or both of the Present Examples options in the Rule Parameters dialog box (see pages 54 - 55), *WizWhy* provides lists of examples after each rule. An "example" is actually the serial number of a record in which the rule either holds or does not hold. To view the content of the record, double-click on the record serial number. The fields and their values of that record will be displayed in the Record Details Grid at the top right pane.

You can use arrow toolbar to review each of the records in the data set.

Visualizing a rule

WizWhy can present any rule by a chart. To visualize a rule, select the rule (it suffices to select one character in a rule), click on the left mouse button, and select Rule Chart. *WizWhy* will open a window such as the following.



The left pane contains the rule in text mode. The right pane is a visualization of the rule. Each bar refers to either one condition or the conclusion. The green section in a bar signifies the value in the condition (or conclusion), the red section represents the missing values, and white section(s) are the other values in this field. The length of each section denotes the number of records, and therefore, the total length of each bar denotes the total number of records.

When the condition refers to a numeric field, the value is placed along the bar at its relative location in the continuation of the field's values. When the condition refers to a text field, the value is always placed at the beginning of the bar.

Number of displayed rules

The number of rules displayed in the Rule List is limited by what you determined in the Rule Type screen at the Maximum number of rules to be displayed on the screen line. For example, if you entered 100, *WizWhy* displays in the Rule Report just the first 100 rules (according to the sort determined in the Sort rules line). To display additional rules –

- Select the Number of displayed rules button on the toolbar  .
- Change the number in the Maximum number of rules to be displayed on the screen line, and click OK. *WizWhy* will refresh the Rule List in accordance with this number.
- You can now select the Rule Page Down button  , to display additional rules, and the Rule Page Up button  to go back.

Record details grid

The Record details grid displays the contents of each record of the data set, according to its serial number.

You can use the arrow keys on the navigation bar (described in Chapter 3) to skip back and forth through the records.

If your rule list contains examples of records after each rule, you can double-click on the serial number of one of the records and the contents of that record will be displayed in the Record details grid, as follows.

A red X checkmark at the far left indicates the dependent variable, and green checkmarks indicate the condition fields. In the icon immediately to the left of the field name, an A indicates a categorical (alphanumeric) field, a 2 indicates a numeric field and a D indicates a date field.

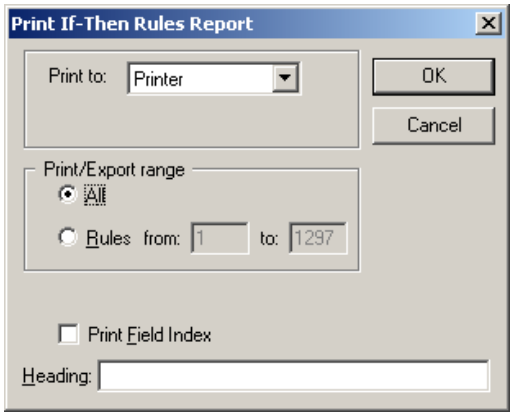
The field index

The Field Index at the bottom right pane lists by field name, the serial number of the rules in which that field appears.

The rule numbers correspond to the serial numbers of the rules displayed in the Rules List at the left pane. To display a rule in the Rule List, double-click on the rule number in the Field Index.

Printing and exporting the rules

To print the rules on the printer or export the rules to another file select the print option  . *WizWhy* will display the Print Rules dialog box.



In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file; or MS Access to export the rules to a database.

In the Print / Export Range box select either All to print (or export) all the rules, or enter a range of rules serial number in the Rules from… to… to print (or export) some of the rules.

Select Field Index to print the field index (see page 76).

You can enter any text in the Heading box, which will be printed as the title of the report.

Exporting rules to an SQL statement

WizWhy can translate the if-then rules generated in the Rule Report into an SQL (Standard Query Language) statement so that you can apply the rules to another similar data set for issuing queries.

When you create an SQL statement, only the "if" part of the rules is translated into an SQL query. The SQL statement can be written to an ASCII file or to the Clipboard, for use by database programs such as Microsoft Access.

To create an SQL query file –

- Select Edit - Select Rules for SQL, and from the cascade menu, define the rules to be exported to SQL using the range of options provided:

  Select All Rules

  Deselect All Rules

  Select All If-Then Rules

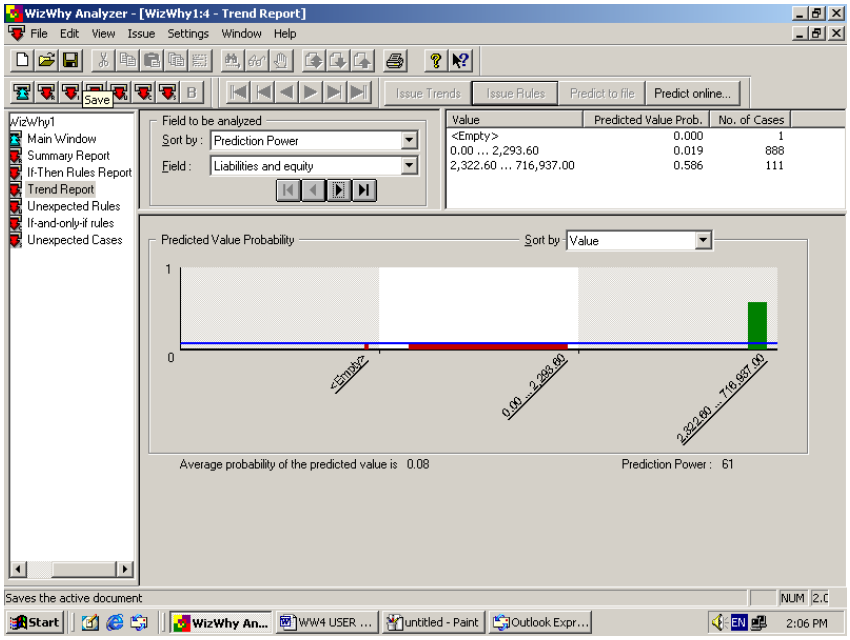
  Select All If-Then-Not Rules

  Inverse Selection of All Rules (selects all rules not selected in the previous step and deselects all rules that *were* selected)

- Select Issue - SQL Statement. The SQL Statement dialog box will be displayed.
- To print the query to an ASCII file, select the Write to File radio button; otherwise, you can select Copy to Clipboard and paste the query into another application.
- Click the Issue command button to write (or copy) the file. Then select Close to exit.

The Trend report

The Trend report presents the one-condition relations in the data, and as such it summarizes the data.



The Trend Report screen is divided into three panes.

- In the top left pane you select the field that you wish to review.
- The top right pane displays the values of the field under analysis and their relations with the dependent variable in characters.
- The bottom pane displays the same information graphically.

Tip: You can move the vertical bar between the panes to display the various data more clearly.

Field to be analyzed

In the top left pane you can select the field you wish to analyze. All the data in the screen refers to the selected field.

The field is selected in the Field line. You can move from the present field to the next or to the previous field by clicking on the arrows.

The fields are sorted by the criteria selected in the Sort by line. The options are:

- Field name: The fields will be sorted alphabetically by their names.
- Field number: The fields will be sorted by the original sort in the data. This is how they are presented in the basic data dialog box.
- Prediction power: The fields will be sorted by an index, calculated by *WizWhy*. This measures to what extent the field explains the dependent variable. If one has to issue predictions by using one field only, this index designates which field is the best predictor, the second best, etc. When calculating this index, *WizWhy* takes into account the expected number of errors and the cost of errors (misses and false alarms).

Predicted value probability as a function of field values

In the top right and bottom panes *WizWhy* displays the relations between the values of the analyzed field and the dependent variable. Assuming that the predicted value is True, these relations can be read as one-condition statements:

If the value in the analyzed field is … then the probability that the dependent variable's value is True is ….

There are …. records where this relation holds.

The right top pane displays these relations in text, while the bottom pane displays them graphically.

In the graphical representation the vertical axis refers to the probability that the dependent variable's value is True, and the horizontal axis presents the values, and the number of cases in each value. The chart is divided into squares. Each square refers to a single value of the analyzed field. Inside each square you see a bar. The height of the bar signifies the degree of probability that the dependent variable's value is True, while the width denotes the number of cases having this value. (The total width of the square signifies the total number of cases in the data). The blue line represents the primary frequency of the predicted value. When the bar is above this line, it means that the field value under analysis is a positive predictor of True, and if it is below the blue line, then the field value is a negative predictor. The higher the difference (up or down) between the height of the bar and the blue line, the better predictor the field value is.

When you click on the left mouse button, while the cursor is in one of the squares, *WizWhy* marks the matching line in the right top pane.

When the field is numeric the values are intervals. These intervals in the numeric fields are automatically determined by *WizWhy*.

When the field is categorical (alphanumeric) having many values, *WizWhy* displays up to 20 values. These are the first values according to the selected sorting.

Note that some of the relations presented in the Trend report might not meet the Minimum number of cases in a rule or the Minimum probability requirement, and as such are not qualified to be rules. Such relations are called *trends* (contrary to *rules*).

You can sort the values by selecting the sort method in the Sort by box. The options are:

- Difference of Probability: The field values will be sorted by the difference of their height (namely the probability that the dependent

variable's value is True) from the blue line (the primary frequency of the predicted value). In this way you see the best predictors first.

- Number of cases: The field values will be sorted by the number of cases in each value.
- None: The field values will be displayed according to the original sort.

Printing and exporting the Trends

To print the Trends on the printer or export the trends to another file select the print option  . *WizWhy* will display the Print Trends dialog box.

In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file; or MS Access to export the Trends to a database.

In the Print / Export Range box select either Current screen to print (or export) the trend in the current screen only, or All to print (or export) all the trends.

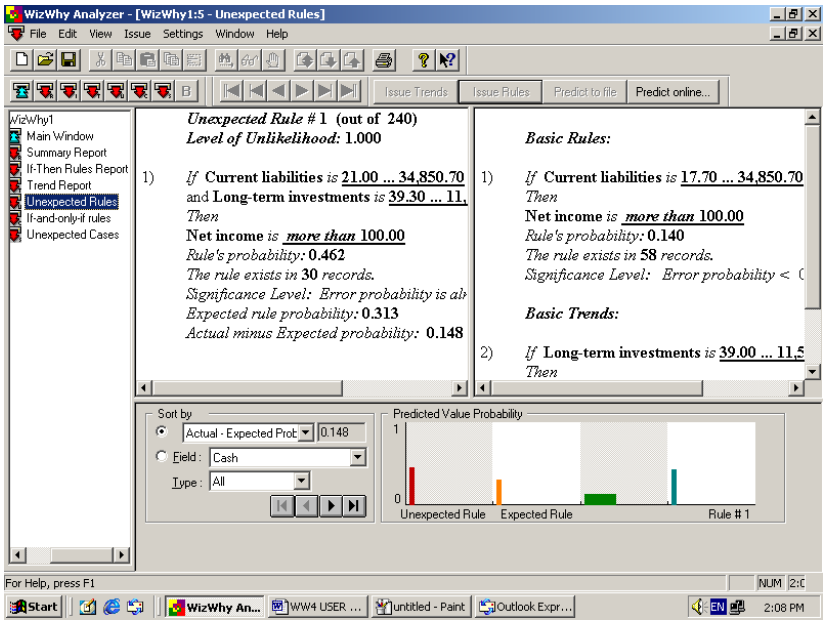
Select the Print charts to print the graphical presentations of the Trend in addition to the text.

You can enter any text in the Heading box, which will be printed as the report's title.

The Unexpected Rules report

The *unexpected rule* report displays the rules that are unexpected relative to more basic rules and trends. These unexpected rules describe the interesting phenomena in the data. The screen is divided into three panes:

- The left top pane displays an unexpected rule.
- The right top pane displays the basic rules and trends in relation to which the rule under discussion is unexpected.
- The bottom pane lets you sort the Unexpected Rules, and displays the rules graphically. By clicking on the arrows at the bottom you can move to the next or previous unexpected rule.



Tip: You can move the vertical bar between the panes to display the various data more clearly.

What is an unexpected rule?

The unexpected rule has at least two conditions. The basic rules have fewer conditions (in many cases they have one condition only), and the basic trends, by definition, have one condition only, as well. Each of the

conditions of the Unexpected Rule appears in the Basic Rules and Trends. The Unexpected rule is unlikely relative to the Basic Rules and Trends.

To see how the Unexpected Rule's level of unlikelihood is computed, consider a data set of 1000 records, where each record refers to one patient, and contains the information whether the patient shows either symptom A or B and the diagnosis (whether or not the patient suffers from the disease D). Suppose also that 30% of the patients have the disease D, and the following three rules were discovered:

1. If a patient shows symptom A, the probability that he or she suffers from the disease D is 50%. (Number of cases: 100).
2. If the patient shows symptom B, the probability that he or she suffers from the disease D is 50%. (Number of cases: 100).
3. If the patient shows both, symptom A and symptom B, the probability of the disease D is 20%. (Number of cases: 20).

In this example, rules #1 and #2 are the basic rules, and rule #3 is the unexpected rule.

WizWhy calculates what should have been the probability of suffering from the disease D among the patients showing both symptoms, A and B, on the basis of rules #1 and # 2, contrary to the actual probability in rule 3. This is the expected probability. To calculate the expected probability *WizWhy* measures the dependency between symptom A and B. Both rules 1 and 2 refer to 200 patients, of which 100 suffer from the disease. Had the symptoms been completely dependent, we would expect 200 patients showing both symptoms. Had they been completely independent, we would expect 40 such patients. In fact there are 100 patients showing both symptoms (of which 20 suffer from the disease), which means that the two symptoms are moderately independent. It turns out that the expected probability in this case is 62.5%. (Had the two symptoms been completely dependent, the expected probability would have been 50%, and had they been completely independent it would have been 70%). The difference between the expected probability, which is 62.5%, and the actual

probability, 20%, signifies how unlikely rule #3 is in regard to rules #1 and #2.

WizWhy measures this unlikelihood in an additional way. *WizWhy* calculates the conditional probability of the event described in rule 3, under the conditions described in rules 1 and 2. In the example under discussion the conditional probability is almost 0, and therefore, the level of unlikelihood, which is 1 minus the conditional probability, is almost 1.

Note that the probability of the Unexpected Rule may be much higher or much lower than the expected probability. Any significant deviation is unexpected.

Basic rules and trends

The right top pane displays the basic rules and the basic trends. The basic rules are included in the rules discovered by *WizWhy*. The rules are displayed in the same way they are presented in the rule report. (See pages 71 - 75 for an explanation about the rule presentation in the rule report).

The basic trends are one-condition relations that do not meet the requirements of a rule in regard to the minimum number of cases in a rule, or the minimum probability. The basic trends can be reviewed in the Trend Report (see pages 78 - 81 for more details).

Unexpected rule

The *unexpected rule* is one of the if-then rules discovered in the data. It is displayed as all the if-then rules together with the following four lines:

In the Unexpected Rule # *WizWhy* prints the serial number of the rule. The rules are sorted by their level of unlikelihood (see explanation below).

The Level of Unlikelihood signifies how unlikely the rule is in regard to the Basic Rules and the Basic Trends. The higher the Level of Unlikelihood, the more unlikely the rule is.

The Expected Rule Probability is the probability one would expect on the basis of the Basic Rules and Trends.

The Actual minus Expected Probability is the difference between the Rule Probability (i.e. the actual probability) and the Expected Probability.

Sorting the unexpected rules

On the left-hand side of the bottom pane, you can determine how the unexpected rules will be sorted. The default is that the unexpected rules are sorted by their level of unlikelihood. In this way you view the most unlikely rule first.

If you wish to limit the unexpected rules to those where a certain field appears in one of the conditions, select the Field line, and then select the desired field. The unexpected rules within this field will be sorted by their level of unlikelihood.

You can also limit the unexpected rules to if-then or if-then-*not* rules in the type line. The default is All, meaning that both if-then and if-then-not rules are displayed.

To move to the next or previous unexpected rule, use the arrows at the bottom.

Visualization of the unexpected rule

On the right hand side of the bottom pane, *WizWhy* displays a graphical representation of the Basic Trend and Rules together with the unexpected rule.

The bars are displayed in the same way they are presented in the Trend report, namely, the height denotes the probability of the Predicted Value, and the width signifies in how many cases the rule holds (see pages 79 - 80).

The first bar to the left refers to the unexpected rule. The next bar represents the expected rule. It has the same width as the unexpected rule, but its height denotes the expected probability. The other bars refer to the basic rules and trends. Green bars represent rules, and gray represent

trends. When you click on a bar *WizWhy* marks the matching rule in the upper panes.

You can also visualize each of the rules, by clicking in any part of the rule, and then clicking on the right mouse button and selecting Rule chart (see pages 74 - 75 for more details).

Printing and exporting the unexpected rules

To print the unexpected rules on the printer or export them to another file select the print option  . *WizWhy* will display the Print Unexpected Rules dialog box.

In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file; or MS Access to export the Unexpected Rules to a database.

In the Print / Export Range box select either Current screen to print (or export) the Unexpected Rule in the current screen only, or All to print (or export) all the Unexpected Rules. You can also enter a range of Unexpected Rule serial numbers in the Rule from… to… to print (or export) some of the Unexpected Rules. (The serial number is printed in the first line at the left top pane).

Select Every Unexpected Rule Starts at new page to avoid cases where two Unexpected Rules are printed on the same page.

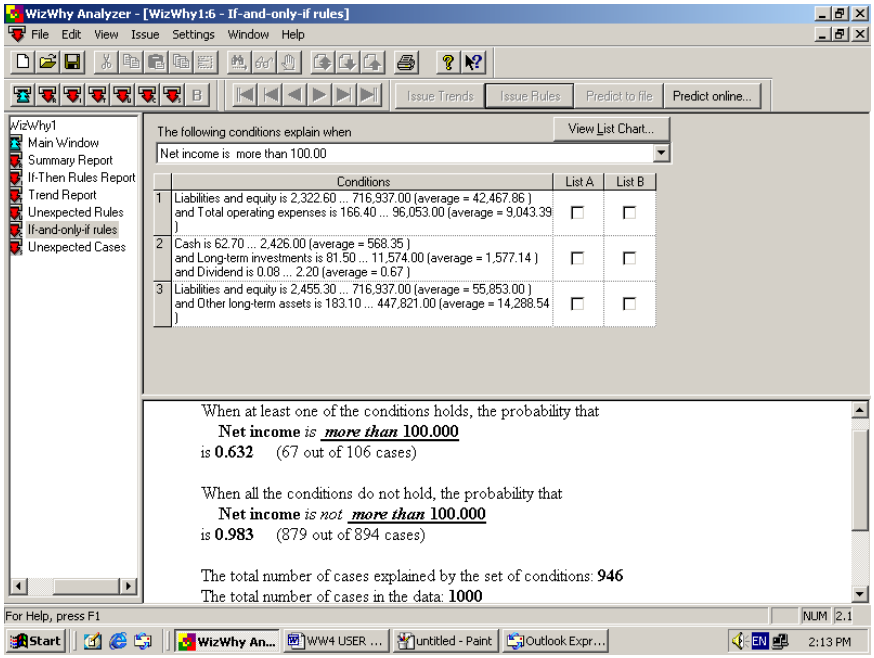
Select Print Basic Rules and Trends to print (or export) the rules and trends in text mode.

Select Print charts to print the graphical presentations of the Unexpected Rules.

You can enter any text in the Heading box, which will be printed as the report's title.

The If-And-Only-If Rules report

The If-And-Only-If Rules report displays the necessary and sufficient conditions for the predicted values. The if-then rules represent sufficient conditions: the “if” condition is a sufficient condition to the conclusion (the “then” part). The if-and-only-if rules go one step further: they represent necessary and sufficient conditions.



The upper box displays one of the predicted values. You can click on the arrow to scroll among the other values, and select the value you are interested in.

The central pane lists the necessary and sufficient conditions that explain the value in the upper box.

The bottom pane displays to what extent the conditions in the central pane explain the value in the upper box.

What is an if-and-only-if rule?

An if-and-only-if rule has the following structure: The dependent variable' value is *r*, if and only if, the value of Field B is *b*, *or* the value of Field C is *c*, etc. In other words, if the value of Field B is *b*, or the value of Field C is *c*, then the dependent variable's value is *r*; and Field B is *not-b*, and the value of Field C is *not-c*, then the dependent variable's value is *not-r*.

When the dependent variable is analyzed as Boolean, *WizWhy* searches two kinds of if-and-only-if rules. The first if-and-only-if rule lists the conditions that meet the following requirement: If at least one of the conditions holds, there is a high probability that the dependent variable's value is *r*, and if none of them holds, there is a high probability that the dependent variable's value is *not-r*. The other if-and-only-if rule lists the conditions that meet the opposite requirement: If at least one of them holds the dependent variable's value is *not-r*, and if none of them holds, the dependent variable's value is *r*.

When the dependent variable is *not* analyzed as Boolean, *WizWhy* applies the above-mentioned search for each of the dependent variable's values. In other words, for each value of the dependent variable *WizWhy* searches two sets of necessary and sufficient conditions. For example, if one of the values is S, *WizWhy* searches one set that refers to *s*, and another set that refers to *not-s*.

The conditions that compose the if-and-only-if rules are the same "if" conditions that compose the if-then rules. *WizWhy* first discovers the if-then rules, and then tries to use the conditions of the if-then rules to explain the positive and negative cases of each of the dependent variable values.

Contrary to the if-then rules, the if-and-only-if rules cover not only the positive cases, but the negative cases as well. Each set of necessary and sufficient conditions explains not only when the predicted value holds, but also when it does not hold. In this sense the if-and-only-if rules are more comprehensive than the if-then rules.

Because there might be several ways to cover (that is, explain) the cases, *WizWhy* searches for the optimal covering, that is a small the set of

necessary and sufficient conditions that explains the maximum number of positive and negative cases.

Each if-and-only-if rule meets two requirements:

- The list of conditions includes a maximum of 50. Having more than 50 conditions might turn it into being uninteresting.
- The probability that the predicted value holds (when at least one of the conditions holds) or does not hold (when none of the conditions hold) should be relatively high in comparison to the primary probability. For example, when the primary probability of the predicted value is 50%, the probability that the predicted value holds (when at least one condition holds) or does not hold (when none of the conditions hold) should be at least 75%.

If one of the above-mentioned requirements is not fulfilled in regard to a certain value of the dependent variable, *WizWhy* does not present an if-and-only-if rule for this value.

Obviously, the if-and-only-if rules are applicable for issuing predictions for new cases. However, as mentioned earlier there might be two if-and-only-if rules for each value (one that refers to the value *r* and another that refers to the value *not-r*), and predictions can also be based on the if-then rules. To resolve the possibility of inconsistent predictions, *WizWhy* compares between these methods, and for each value it finds out the most accurate method. This is the method used for issuing the predictions.

Note that the if-and-only-if rules can also be considered as another way of summarizing the data: One if-and-only-if rule can do the work of thousands of rules in explaining the data.

The if-and-only-if rule parameters

At the bottom pane *WizWhy* displays some parameters in regard to the if-and-only-if rule. These parameters can help you when considering to what extent the if-and-only-if rule explains the cases.

If the value under analysis is, for example, *r*, *WizWhy* lists the following lines:

In the first item, *WizWhy* displays the probability that *r* holds, if at least one of the conditions holds, that is, if the first condition holds, *or* the second condition holds and so on. *WizWhy* also prints the total number of the cases entailing this probability, that is the total number of cases where at least one of the conditions holds, and the total number of cases where at least one of the conditions holds, and *r* holds as well.

In the second item, *WizWhy* displays the probability that *not-r* holds if all the conditions do not hold, that if the first condition does *not* hold, *and* the second condition does *not* hold and so on. *WizWhy* also prints the relevant number of cases.

In the Total number of cases lines, *WizWhy* prints the sum of the numbers presented above: the sum of the positive and negative cases explained by the conditions, and the total number of the positive and negative cases. By dividing these two numbers *WizWhy* calculates the success rate. This number denotes to what extent the conditions correctly explain the predicted value.

The next line lists the primary positive and negative probabilities of the predicted value (note that the sum of the two probabilities is 1).

Finally, the Improvement Factor line denotes to what extent the if-and-only-if rule explains the dependent variable values relative to an explanation based on the primary probability only.

Displaying the rule behind the condition

Every condition is derived from an if-then rule that is included in the rule report. If you want to see the rule, right-click on the condition's serial number (on the left column). To restore the if-and-only-if rule's parameters, click on the right mouse button and select If-and-only-if Rule.

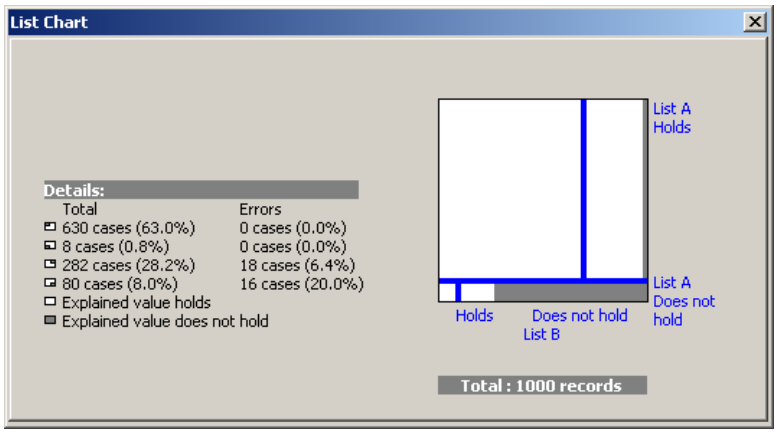
Visualizing an if-and-only-if rule

On the right side of the central pane there are two columns – List A and List B. Select one or more conditions on List A and one or more other conditions on List B. (Each condition can be in one list only). Now click

on the View List Chart button. *WizWhy* will display a chart containing four rectangles (limited by a blue border line) representing the following four combinations:

1. The conditions of both List A and List B hold.
2. List A's conditions hold but List B's conditions do not hold.
3. List A's conditions do not hold, but list B's conditions hold.
4. The conditions of both List A and List B do not hold.

The white area in each rectangle indicates the cases where the dependent variable's value holds, while the gray area indicates the cases where this value does not hold.



Most of the area in the first three rectangles should be white (because in these rectangles at least one of the conditions holds) while most of the area in the fourth rectangle should be gray (because in this rectangle all the conditions do not hold). Therefore, in the first three rectangles the gray area indicates errors (that is, cases where the conditions do not explain the dependent variable's value), while in the fourth rectangle the white area indicates the errors.

By re-issuing this chart you can see the contribution of each condition (or group of conditions) to the explanations of the cases.

Printing and exporting the if-and-only-if rules

To print the if-and-only-if rules on the printer or export them to another file select the print option  . *WizWhy* will display the Print If-and-only-if Rules dialog box.

In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file; or MS Access to export the Unexpected Rules to a database.

In the Print / Export Range box select either Current screen to print (or export) the if-and-only-if rule in the current screen only, or All to print (or export) all the if-and-only-if rules. You can also enter a range of If-and-only-if Rule serial numbers in the from… to… to print (or export) some of the if-and-only-if rules.

Select Every If-and-only-if Rule Starts at new page to avoid cases where two if-and-only-if rules are printed on the same page.

Select Print Basic Rules print (or export) the if-then rules where the conditions are included.

Select Print charts to print the graphical presentations of the if-and-only-if rule.

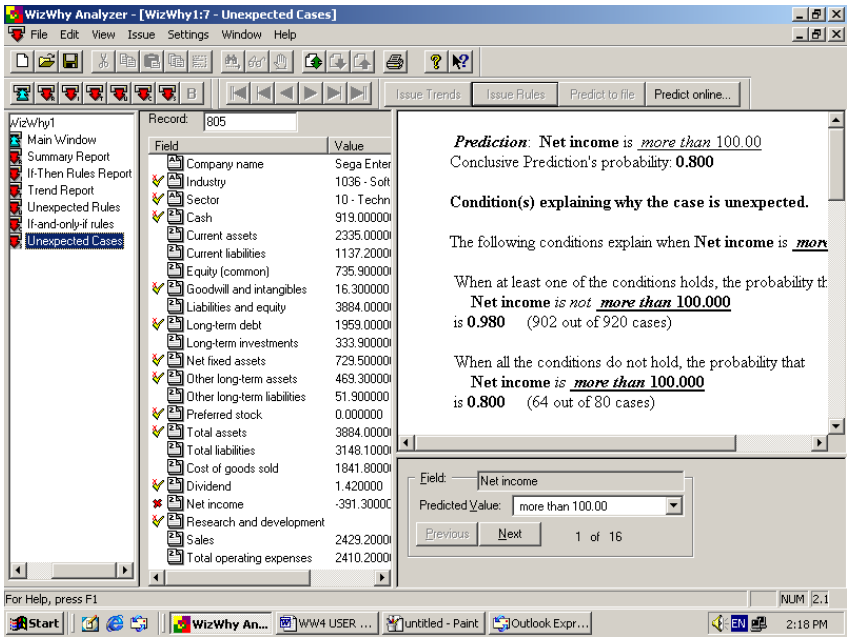
You can enter any text in the Heading box, which will be printed as the report's title.

The unexpected cases report

The *unexpected cases* report lists the cases deviating from the rules. As mentioned earlier in this chapter (see pages 68 - 69), *WizWhy* uses the if-then rules and if-and-only-if rules to issue predictions in regard to the records of the data set that was analyzed. In some cases the actual value in

the record might be different from the value predicted by *WizWhy*. The Unexpected Rule Report lists these cases.

Each unexpected case, that is a record deviating from the rules, is presented on a different screen.



The unexpected cases are grouped by the predicted value, that is the value that should have been in the dependent variable according to the rules. Within the predicted value, the unexpected cases are sorted as follows:

- If the dependent variable is *continuous*, the unexpected cases are sorted by the absolute difference of the predicted value from the actual one. The cases where the deviation is the largest are presented first.
- If the dependent variable is *categorical*, and the predictions are based on the if-then rules, the unexpected cases are sorted by the prediction's conclusive probability (this is a measurement of the prediction error probability – for more details see page 104). The

cases having the highest conclusive probability are presented first. If the predictions are based on if-and-only-if rules (and therefore the conclusive probability is not calculated) the cases are sorted by their serial number.

Predicted value

Select the predicted value you are interested in from the list box at the bottom left pane. As mentioned earlier, this is the value predicted by the rules, that is that value that should have been the dependent variable's value. *WizWhy* will display the contents of the first unexpected case having this predicted value.

Click the next or previous buttons to see the other cases having the same predicted value.

Record details grid

In the pane on the right of the screen, *WizWhy* displays the record details, that is the record deviating from the rules.

WizWhy displays the field names and the values in the record under analysis. The record's serial number is displayed at the top.

A red X checkmark at the far left indicates the dependent variable. A checkmark indicates the condition fields that are relevant to the prediction.

When the prediction is based on a if-and-only-if rule, where at least one of the conditions holds, *WizWhy* displays the checkmark in the fields that fulfill the conditions (because these fields imply that at least one of the conditions holds). When the prediction is based on an if-and-only-if rule, where none of the conditions hold, *WizWhy* displays the checkmark (with a small x in a superscript font) in all the fields included in the rule (because all them together imply that the conditions do not hold).

When the prediction is based on if-then rules, this checkmark can be green, yellow or white. Green indicates that *if-then* rules apply; yellow

indicates that *if-then-not* rules apply; and white indicates that both, if-then and if-then-not rules apply.

In the icon immediately to the left of the field name, an A indicates a categorical (alphanumeric) field, a 2 indicates a numeric field and a D indicates a date field.

The rules explaining the prediction

WizWhy displays the rules explaining the prediction in the right pane. Following the explanations earlier in this chapter about the way *WizWhy* issues predictions (see pages 68 - 69), the rules are either if-then or if-and-only-if rules. These rules together with the record's field values (marked by a checkmark) entail the predicted value, which is different from the dependent variable's actual value (marked by the red X).

WizWhy might not present any rule, in rare cases when the prediction is based on ruling out all the other possible values. For example, *WizWhy* might predict that the dependent variable's color is blue, not on the basis of rules positively implying this color, but on the basis of rules implying that no other color holds.

When the rules explaining the prediction are if-then rules, the number of rules displayed in the right pane is limited by what you determined in the Rule Type screen at the Maximum number of rules to be displayed on the screen line. For example, if you entered 100, *WizWhy* displays in the Rule Report just the first 100 rules (according to the sort determined in the Sort rules line). To display additional rules –

- Select the Number of displayed rules button on the toolbar  .
- Change the number in the Maximum number of rules to be displayed on the screen line, and click OK. *WizWhy* will refresh the Rule List in accordance with this number.
- You can now select the Rule Page Down button  , to display additional rules, and the Rule Page Up button  to go back.

Printing and exporting the unexpected cases

To print the unexpected cases on the printer or export them to another file select the print option  . *WizWhy* will display the Print Unexpected Cases dialog box.

In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file; or MS Access to export the Unexpected Rules to a database.

In the Print / Export Range box select either Current screen to print (or export) the unexpected cases in the current screen only, or All Cases to print (or export) all the unexpected cases. You can also limit the cases to those having the current predicted value and to further limit them by entering a range of unexpected cases serial numbers in the from… to… boxes.

Select Every Unexpected Case Starts at new page to avoid cases where two Unexpected Rules are printed on the same page.

Select Print Rules to print (or export) the if-then rules or if-and-only-if rules explaining the deviation.

You can enter any text in the Heading box, which will be printed as the report's title.

8. Issuing Predictions

On the basis of the if-then rules and the if-and-only-if rules, *WizWhy* can predict the value of the dependent variable of a new record, given the values of all or some of the condition fields (the independent variables).

WizWhy includes three types of prediction methods:

- Predictions generated using the Predict to File command. With this method, the predictions are issued for records saved in a different file.
- Predictions made using the Predict On-Line command. With this method, you enter the values of the condition fields manually.
- Predictions issued by the independent *WizWhy* Predictor application also based on field values that you enter manually.

How does *WizWhy* issue predictions?

When a new record is entered *WizWhy* applies the rules on the values of this record and calculates the predicted value. For example, one can run *WizWhy* on financial data, where the dependent variable is a field signifying whether the company went bankrupt. *WizWhy* will reveal the rules that relate between the company data and the probability of going bankrupt. When the data of a new company is entered, *WizWhy* will then apply the relevant rules, and calculate the probability of this company going bankrupt.

When issuing the predictions *WizWhy* can list the rules that entail each prediction. These rules serve as the explanations of the predictions.

> **Note**: The prediction assumes that the record(s) for prediction have been taken from a population similar to that of the data from which the rules were generated.

Validating the rules and Updating predictions to file

Based on rules issued for a given data set, *WizWhy* can make predictions for each record of another data set. *WizWhy* reads the values of the new records, applies the relevant rules, and issues a prediction for each record.

This method can be used for two purposes:

- When the new data set contains the values of the dependent variable, the prediction can be used for validating the rules. The accuracy of the predictions can be calculated by comparing the *WizWhy* predicted values against the actual values. In a typical application of this method, the data set is randomly cut into two files – one file is used as the Train file (the data set used for issuing the rules), while the other serves as the Test file (the data set used for the validation).
- When the new data set does not contain the values of the dependent variable, the prediction is used for predicting the expected values.

Validating the rules

To validate the rules –

Move to the Prediction Input tab in the Main Window.

Open the Input data set according to the explanations presented in chapter 5. This Test data set will serve as the Test file.

WizWhy will open a Save As dialog box, where you enter the path and file name of a new ASCII file. As long as you just validate the rules, the file will not be updated.

Select the prediction method from the Issue prediction by list box. This option is active only if the dependent variable is analyzed as Boolean.

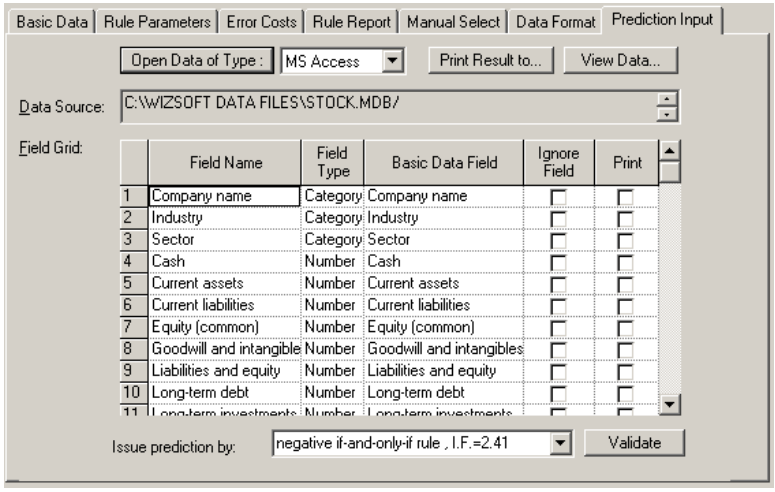
Click on the Validate button.

Up to 3 prediction methods are displayed in the Issue prediction by list box (when the dependent variable is analyzed as Boolean):

- If-then rules: the predictions will be issued on the basis of the if-then rules (the if-and-only-if rules will be ignored).
- Positive if-and-only-if rule: the predictions will be based on the if-and-only-if rule explaining when the predicted value holds. (The negative if-and-only-if rule and the if-then rules will be ignored).
- Negative if-and-only-if rule: the predictions will be based on the if-and-only-if rule explaining when the predicted value does not hold. (The positive if-and-only-if rule and the if-then rules will be ignored).

The improvement factor (I.F.) of each method is also displayed. The improvement factor denotes to what extent the rule(s) explain the predicted value relative to an explanation based on the primary probability only. The default prediction method is the one having the highest improvement factor.

The list is limited to the discovered rules. For example, if *WizWhy* did not discoverer a positive if-and-only-if rule, the list contains if-then rules and a negative if-and-only-if rule only.



Once you clicked on the Validate button, *WizWhy* issues the predictions on the Test file, and compares the predicted values with the actual ones. It then displays a window summarizing the rule explanatory power. This window contains the following lines:

Number of Misses: Cases where *WizWhy* predicts R while the actual value is *not*-R.

Number of False Alarms: Cases where *WizWhy* predicts *not*-R, while the actual value is R.

Average Error Cost (per record): The number of misses multiplied by the cost of a miss plus the number of false alarms multiplied by the cost of a false alarm, divided by the number of records.

Expected Average Error Cost (per record): The average error cost resulting from issuing predictions on the basis of the % frequency of the predicted value, the cost of a miss, and the cost of a false alarms only.

Improvement Factor: The expected error cost divided by the average error cost. This number signifies the contribution of the rules beyond what is known without the rules.

Predicting the expected values

To update predictions to a file –

Move to the Prediction Input tab in the Main Window.

Open the Input data set according to the explanations presented in chapter 5. This Test data set will serve as the Test file.

WizWhy will open a Save As dialog box, where you enter the path and file name of a new ASCII file.

Note that the process refers to 3 tables: (1) The Basic Data is the train data where the rules were issued, (2) The Prediction Input Data contains new records, in regard to which *WizWhy* will issue the predictions, (3) The Output Data is an ASCII file containing the *WizWhy* predictions.

The fields of the input data will be displayed in the field grid. Fill in the lines in the Data Grid.

The Field Name is the name, as it appears in the Prediction Input data.

The Field Type column defines whether the field type is Category, Number or Date. Category indicates categorical (alphanumeric) data; for example, Item Number, Address, Reference Number, Occupation, or Yes/No fields. Number indicates numeric data to which mathematical computations can be applied; for example, Total $, Weight, Length or Percent. Date is the date data in one of the standard formats (D-M-Y, M-D-Y, etc.).

In the Basic Data Field *WizWhy* displays the matching field name in the train data set (i.e., the data set in regard to which the rules were issued). By default, *WizWhy* matches fields having the same field names. To change or enter a field name, click in the cell and select among the Basic Data field names. Note that when you click in the cell, *WizWhy* does not list the field names that were already matched. Therefore, to change field names start by deleting the wrong matches, and only then select the right matches.

Select in the Ignore Field column to instruct *WizWhy* to disregard the data in this field when issuing the predictions. (It makes sense to do so, when you know that the data in one of the fields is corrupted or quite different from the basic data).

Select in the Print column to instruct *WizWhy* to include the data of this field in the file with the predictions. This file will include all the checked fields together with the prediction itself.

Click on the Predict to File button in the action bar. *WizWhy* will create the prediction file and notify you of the processing. The file with the prediction can be opened by any application that reads ASCII files.

The file with the predictions includes the fields that you determine together with the following two columns:

- Predicted value, which is the value predicted by *WizWhy*.
- Conclusive prediction's probability. Note that records for which no rules are applicable (and therefore, for which no prediction can be made) are assigned a -1 (minus 1) conclusive prediction's

probability. Note also that this item is not calculated when the prediction is based on if-and-only-if rules.

You can translate the conclusive probability into predictions in two ways:

You can use the decision point to determine either a positive prediction of the predicted value (when the conclusive probability is higher than the decision point), or a negative prediction, saying that the predicted value will not hold (when the conclusive probability is lower than the decision point).

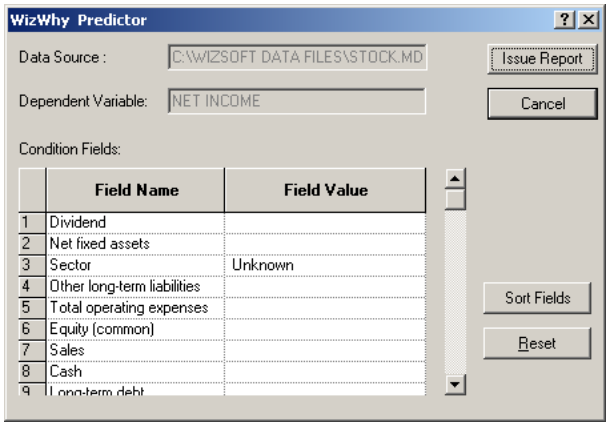
Alternatively, you can sort the records by the conclusive probability, and assign the predicted value to the top P records, where P is the primary frequency of the predicted value. For example, when analyzing company data, where the dependent variable is whether a company went bankrupt, assuming that the primary probability of going bankrupt is 1%, you can sort the records by the conclusive probability (which is the probability of going bankrupt), and assign a positive prediction to the top 1% records.

Tip: When you wish to update predictions in a file, you don't have to wait till *WizWhy* finishes issuing the rules. When you determine the analysis parameters (as explained in chapter 6), you can also enter the prediction input parameters as explained above. If you then click on the Predict to File button, *WizWhy* performs the entire process. It first reveals the rules, and then automatically continues to update the predictions in the file.

Predict on-line

To issue a prediction for a new case entered manually --

Click the Predict On-Line command button in the action bar (this button is activated only after the rules have been issued). The *WizWhy* Predictor dialog box will be displayed.



You may wish to sort the fields to facilitate your data entry. To do so, click the Sort Fields command button. A small Sort Condition Fields dialog box will be displayed.

Select the criterion for sorting the records:

- Prediction Power is an index computed from the number of rules in which the field appears and the significance level of these rules. This is the default. When the fields are sorted in this manner, the higher the field in the list, the more relevant it is for issuing predictions.
- Original Field Order indicates the order of the fields in the original data set.
- Alphabetical order simply lists the fields from A through Z.

In the Field Value column, enter values of new data in all or some of the condition fields. In categorical fields, you may select your values from the drop-down list (of which Unknown if the first choice). Only those fields and values included in the Rule Report are displayed. If the value does not appear in the list, select Unknown. In numeric (quantitative) fields, type in the value.

When you have completed your data entry, click Issue Report. The Prediction Report will be displayed in the *WizWhy* work area.

Reading the prediction report

The prediction report contains these elements:

The header includes general information, including a list of the values that you entered in each of the condition fields and the Predicted Value.

The Prediction section includes the results of applying the rules on the conditions. For example:

> Prediction's significance level: Error probability = 0.00
>
> Primary Predicted Value probability: 0.29
>
> Conclusive prediction's probability: 0.69
>
> Prediction: more than 1000

Where:

Prediction's significance level indicates the degree to which the prediction is doubtful. The calculation of this figure is based on the error probability of the if-then rules, the direction of the rules (if-then versus if-then-*not*) and the number of cases in a rule. The lower the error probability, the more certain the prediction is.

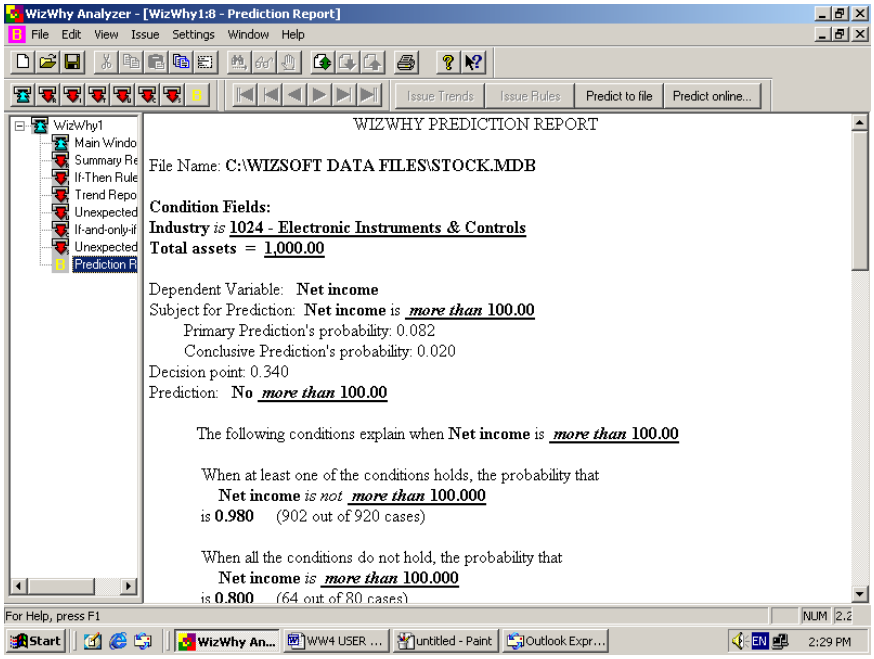
When the number of if-then rules is close to the number of if-then-not rules, the error probability rises. Such a case usually indicates that the record for prediction has been taken from a population *not* similar to that of the data from which the rules were generated.

Note that this line is not displayed when the prediction is based on an if-and-only-if rule. The error probability of predictions based on if-and-only-if rules is always very low.

Primary Predicted Value probability is the a priori probability of the predicted value in the data set from which the rules were generated. This is the probability of the predicted value, if no rule applies to the case under analysis.

Conclusive prediction's probability is actually the "*bottom line*". This figure takes into account all the other figures, and indicates the final probability of the predicted value. The higher the conclusive probability, the higher the probability that the predicted value holds in the case under

analysis. Once again, note that this line is not displayed when the prediction is based on an if-and-only-if rule.



In the Prediction line *WizWhy* prints the prediction itself, which is either a positive prediction of the Predicted Value (when the conclusive probability is higher than the decision point), or a negative prediction, saying that the predicted value will not hold (when the conclusive probability is lower than the decision point).

The Relevant Rules section lists the rules that were applied on the conditions. The rules are either if-then rules, or if-and-only-if rules (according to the set of rules selected for issuing the predictions).

When the prediction is based on the if-then rules, the number of rules displayed is limited by what you determined in the Rule Type screen at the Maximum number of rules to be displayed on the screen line. For example, if you entered 100, *WizWhy* displays in the Rule Report just the first 100 rules (according to the sort determined in the Sort rules line). To display additional rules –

- Select the Number of displayed rules button on the toolbar  .
- Change the number in the Maximum number of rules to be displayed on the screen line, and click OK. *WizWhy* will refresh the Rule List in accordance with this number.
- You can now select the Rule Page Down button  , to display additional rules, and the Rule Page Up button  to go back.

Creating a prediction application

The *WizWhy* predictor is an independent application for issuing predictions on the basis of the discovered rules. This method is ideal for limiting access to the contents of confidential data.

Using the *WizWhy* Predictor enables you to issue the rules on one computer and to transfer the application to other computers accessed by authorized *WizWhy* users.

To create a *WizWhy* Prediction application:

- Issue rules.
- Use the Save or Save As command to save the working file (*.wwr file). The rules are saved in this file.
- Copy the working file and install the *WizWhy* Predictor application in another computer.
- To issue Prediction Reports in this computer, access the *WizWhy* Predictor application directly from the Windows Start menu. The *WizWhy* Predictor enables the user to issue prediction reports like the full *WizWhy* application, but the process begins with selecting the Rule file.

Important: The other computer *must* have a legally installed version of *WizWhy* Predictor or a license to use the application.

9. Frequently Asked Questions

Q: *How do I know that the WizWhy algorithm is actually finding all the possible rules?*

A: The algorithm itself is far too complex to be explained in this guide. But you can perform an experiment: take a data set in which rules have already been defined and have *WizWhy* analyze it. See if *WizWhy* can find the rules!

Q: *How do I know that the WizWhy algorithm predicts accurately?*

A: You can check how well *WizWhy* predicts without going into the details of the *WizWhy* algorithm. Simply select a data set; cut it randomly into two parts, one part will serve as a train file, while the other part will be the test file. Select the train file for issuing the rules, and then validate the rules by issuing predictions to the test file. See pages 98 - 100 for more explanations.

Q: *How accurate WizWhy is in comparison with other tools for issuing predictions?*

A: You can compare *WizWhy* with other tools for issuing predictions by using the method described in the previous answer. Using this method one can compare the accuracy of any two methods without going into details of the mathematical algorithms behind the methods.

Q: *How does WizWhy avoid revealing redundant rules?*

A: When *WizWhy* reveals the rules, it deletes the redundant ones. For example, consider the following two rules:

(1) If Field A is *a*, the dependent variable is *r*.

(2) If Field A is *a*, and Field B is *b*, the dependent variable is *r*.

Because rule (2) is identical to rule (1), except for the additional condition in (2), its rule probability should be at least 2% higher than rule (1) probability. For example, if the rule probability of rule (1) is 70%, the rule probability of (2) should be 72% or higher. If not, (2) is considered to be a redundant rule.

Q: *Can WizWhy analyze series?*

A: *WizWhy*'s algorithm is not a time series analysis algorithm. However *WizWhy* can be applied, if you convert your data to a table. Assume that you data set contains the sales figures of the last 100 months. Convert this list into records having the following fields: The first record will contain the sales of the first and the second months, the second record will contain the figures of the second and the third months, and so on. You can add other fields such as the month name (January, February, and so on), the difference between month N and month N-1. You will end up with a table having the following fields: Sales of Month N, Sales of Month N-1, % Difference, Month N name. Select Month N as the dependent variable. *WizWhy* will reveal the rules explaining the sales in any month as a function of the sales in the previous month and the month name. These rules will present both the trend (the change from the previous month) and seasonal effect (the month).

Q: *What are the parameters that affect the prediction's accuracy?*

A: Usually *WizWhy*'s predictions are most accurate when you use the *WizWhy* defaults in the Rule Parameters dialog box. If you reduce the number of cases in a rule, *WizWhy* will reveal more rules, but the effect of overfitting might be increased as well. On the other hand if you increase the number of cases in a rule, the effect of overfitting will be decreased, but some important rules might be ignored.

Q: *How does the cost of error affect the accuracy of the predictions?*

A: If the difference between misses (*WizWhy* predicts 1, but the actual value is not 1) and false alarms (*WizWhy* predicts not 1, but the actual value is 1) matters, you have to enter the cost of errors. The cost of errors

signifies the proportion between the importance of avoiding a miss versus the importance of avoiding a false alarm. *WizWhy*'s object is to minimize the total cost of errors. If the cost of a miss is higher than a cost of false alarm, *WizWhy* predictions will include fewer misses than false alarms. When the cost of a miss is equal to the cost of a false alarm the total number of errors is minimal.

Q: *How can I decrease the number of misses and false alarms in the analyzed data?*

A: Perform one or more of the following:

- In the Rule Parameters dialog box, decrease the Minimum number of cases in a rule.
- In the Error Costs dialog box make sure that the cost of a miss is equal to the cost of a false alarm.

Q: *How can I decrease the number of rules while revealing only the most significant ones?*

A: In the Rule Parameters dialog box perform one or more of the following:

- Increase the Minimum probability of if-then rules.
- Increase the Minimum probability of if-then-*not* rules.
- Increase the Minimum number of cases of a rule.
- Decrease the Maximum number of conditions in a rule.

Q: *A Rule report was generated with no (zero) rules. Why?*

A: The parameters that you set in the Rule Parameters dialog box were not sufficient to establish rules. Readjust the parameters: you might change one or more of the following:

- Minimum probability of if-then rules
- Minimum probability of if-then-*not* rules

- Minimum number of cases in a rule

Q: *How do I increase the speed of issuing the rules?*

A: You can set a number of different parameters to influence the speed of issuing the reports. The main methods include:

Increasing the Minimum number of cases in a rule in the Rule Parameters dialog box.

Reducing the Maximum number of conditions in a rule in the Rule Parameters dialog box.

Deleting uninformative fields by selecting Ignore in the field grid.

Increasing the RAM of your computer. (Note that if not enough RAM is available, it causes *WizWhy* to use the hard drive resources to issue the calculations, decreasing its efficiency by a factor of 100).

Q: *What do I do if, when issuing the rules, WizWhy gets stuck in the middle of the calculation stage?*

A: You can click the Move Forward button. This button appears on the Progress Indicator, whenever possible. For example, if you click this button in the searching for 3-conditions rules, *WizWhy* will *WizWhy* will jump to next stage without completing the search for rules having 3 conditions (or more). You can also select the Cancel button to stop the entire process of issuing the rules.

Q: *What is the largest size of data set that WizWhy can practically analyze?*

A: The size of the data set is neither limited by the number of fields nor the number of records. However, to save time, if the data set is very large consider the following:

If there are more than 200 – 300 fields, you can start the analysis by issuing the trend report, and then ignoring all the fields having a low prediction power.

If there are more than 500,000 records, you can start by creating a representative sample of the data (using one of the statistical packages) and run *WizWhy* on this sample. Note that as rule of thumb 1000 positive cases suffice for revealing the important rules. For example, if the primary probability of the predicted value is 1%, a data set having 100,000 (where 1000 are positive examples) contains enough information to enable *WizWhy* to reveal the important rules.

Q: *How does WizWhy segment the dependent variable into intervals?*

A: When the dependent variable is continuous and is not analyzed as Boolean, *WizWhy* cuts it into up to 9 intervals. The segmentation into intervals follows two restrictions: (1) the intervals should be in accordance with the distribution of the values, and (2) the intervals should be as equal as possible.

Q: *How do I save the reports?*

A: Whenever you operate *WizWhy*, it creates a *.wwr file, which includes the parameters of the analysis and the last issued reports. You can save this file through the File - Save or File - Save As option. To open the saved file, select it through the File - Open option (or from the list of last four saved files).

Note that when you reopen a file, *WizWhy* keeps the previously issued reports until the new ones are actually issued; then it replaces them with the new saved reports. If you want to save the previous reports, use the File - Save As option to save the *.wwr file under a different name, and reissue the analysis from the beginning.

Q: *How do I use the rules in another application?*

A: You have two options:

Use the Print option to export the Rule, Trends, Unexpected Rules, If-and-only-if Rules or Unexpected Cases reports to MS Access, and then use the rules in your application.

Use the *WizWhy* ActiveX (OCX) version. This program lets you operate all the *WizWhy* commands embedded in another application. Note that the ActiveX program is not included in the *WizWhy* package, and should be purchased separately.

Q: *How do I sort the rules in the Rule report?*

A: The rules in the Rule Report are sorted by the rule significance level, the rule probability or number of cases in a rule. The criterion is determined in the Rule Parameters dialog box. If you wish to issue a report where the rules are sorted in another way, use the Print option to export the rules to MS Access, and sort the rules there.

Q: *What is the structure of the Microsoft Access table created by the WizWhy export option?*

A: *WizWhy* exports the rules to Microsoft Access in the following two formats:

Spreadsheet format: Each rule is written in one record, where each column refers to another field in the data set.

One condition in a line: Each condition (and the result) is written in another record.

Q: *How do I change the appearance of the reports?*

A: There are three ways:

Use the Data Format dialog box to change the report header and the settings for font type, style and size.

Use the Print option to export the report to Microsoft Access, and edit it there.

Use the standard Windows functions to cut, copy and paste parts of the report into a word processor (or any other Windows – compliant – text - processing application, such as PowerPoint) and edit it there.

Glossary

Association rules

A method for revealing all the if-then rules in a given data set. One of the main challenges of such a method is to validate each possible relationship, in a *reasonable time-span*. *WizWhy* employs a sophisticated association rules algorithm that reveals all the if-then rules in an astonishingly short time.

Average error cost (per record)

The total cost of errors divided by the number of records in the data. When issuing predictions *WizWhy*'s object is to minimize this number.

Basic data

The data set under analysis. When issuing predictions this is the Train data.

Basic rules and trends

The basic rules and trends are if-then statements, in regard to which the *unexpected rule* is unlikely. The conditions of both are the conditions of the unexpected rule; however, the unexpected rule contains all the conditions, while each of the basic rules and the basic trends contains some of them.

The basic rules are included in the rules discovered by *WizWhy*. The basic trends are one-condition relations that do not meet the requirements of a rule in regard to the *minimum number of cases in a rule*, or the *minimum probability*.

Boolean analysis

An analysis of information in a data set in which the *dependent variable* has only two alternative values, such as Yes or No, or 0 or 1. If the dependent variable contains more than two different values, the Boolean analysis is performed with respect to one of the values selected by the user. If the dependent variable is continuous, a Boolean analysis is performed with respect to the user-defined range. *WizWhy* can issue both Boolean and *multi-value analysis*.

Conclusive prediction's probability

When *WizWhy* issues a prediction, and the prediction is based on the *if-then rules*, *WizWhy* calculates the probability that the *predicted value* holds in the dependent variable. This probability is called the Conclusive Probability. It takes into account all the other figures, and indicates the final calculation of the probability of the predicted value. The higher the conclusive probability, the higher probability that the predicted value holds in the case under analysis.

Data mining

The search for valuable, yet hidden, patterns and relationships within a data set.

Decision point

When *WizWhy* issues a prediction, and the prediction is based on the *if-then* rules, *WizWhy* calculates the probability that the value of the *dependent variable* in the record under analysis is 1 (assuming that 1 is the *predicted value*). This probability is called the Conclusive Probability. Now, when the analysis is *Boolean*, there should be a decision point, that discriminates between the cases where the dependent variable is 1 and the cases where it is not 1. This point is the decision point: When the conclusive probability is above the decision point *WizWhy* predicts that the predicted value is 1, and if it is below the decision point, *WizWhy* predicts that the dependent variable's value is not 1.

Dependent variable

The field to be explained in relation to the other fields of the data (independent variables).

Error costs

The cost of a miss versus the cost of a false alarm.

Both misses and false alarms refer to predictions. For example, when diagnosing a patient, if one predicts that the patient does not suffer from a certain disease, and it turns out that he or she does, the diagnosis (or prediction) is considered as a miss (the diagnosis missed the disease). On the other hand, if one predicts that the patient suffers from the disease although in fact he or she does not, the diagnosis is an example of a false alarm. When issuing predictions *WizWhy* can trade off between misses and false alarms, and the cost of errors entered here tells *WizWhy* what the optimal trade off is.

Error probability

The error probability of a rule indicates the percent chance that the rule under discussion exists accidentally in the data set. The *significance level* of the rule equals 1 *minus* the error probability. It may be interpreted as the extent to which a rule presents an essential phenomenon in the data. The lower the error probability, the higher the degree the rule can be "trusted" when issuing a prediction. The concept of the error probability is similar to the concept of the α probability in classical statistical tests.

Expected average error cost (per record)

The expected average error cost is the result of issuing predictions on the basis of the % frequency of the *predicted value*, the cost of a miss, and the cost of a false alarm only. In other words, this is the expected average error cost when no rule is known. For example, if the frequency of the predicted value is 15%, and a cost of a miss is 2, while the cost of false

alarm is 1, the best way for issuing predictions when no rule is known, is to predict that the predicted value holds in *none* of the records. Such a prediction will result in misses in 15% of the records, no false alarms, and since the cost of a miss is 2, the average cost per record will be 0.3. This is the expected average error cost.

Expected probability

The conditional probability of the event described by the conditions, result and the number of cases of the *unexpected rule*, given the events described by the Basic Rules and Trends.

Expected rule

A rule having the same conditions, result and number of cases as the *unexpected rule*, while the rule's probability is *the expected probability*. This is the rule one would expect on the basis of the *basic rules and trends*.

Field type

The Field type defines whether the field type is *Category*, *Number* or *Date*

Category indicates categorical (alphanumeric) data; for example, Item Number, Address, Reference Number, Occupation, or Yes/No fields.

Number indicates numeric data to which mathematical computations can be applied; for example, Total $, Weight, Length or Percent.

Date is the date data in one of the standard formats (D-M-Y, M-D-Y, etc.).

If-and-only-if rule

A list of conditions that meet the following requirement: If at least one of the conditions holds, there is a high probability that the dependent

variable's value is *r*, and if none of the conditions hold, there is a high probability that the dependent variable's value is *not-r*.

Each one of the conditions has the following structure: The value of Field A is *a*, and the value of Field B is *b*, and so on (like the conditions of the if-then rules).

Each if-and-only-if rule can be interpreted as a list of necessary and sufficient conditions for *r*.

If-then rule

A rule having the following structure:

If the value in field A is *a*
And the value in field B is *b*
And ..
Then the value of the dependent variable is *r*

If-then-not rule

A rule having the following structure:

If the value in field A is *a*
And the value in field B is *b*
And ..
Then the value of the dependent variable is *not* r

Improvement factor

The *expected error cost* divided by the *average error cost*. This number signifies the contribution of the rules beyond what is known without the rules.

Level of unlikelihood

How unlikely the *unexpected rule* is in regard to the *basic rules and trends.* The higher the level of unlikelihood, the more unlikely the rule is.

Missing value

An empty field in a certain record. For each field you may determine if missing values are informative or not, namely, whether or not you wish *WizWhy* to look for rules such as: If Field A is empty, then the dependent variable is 1. Both numeric and categorical fields are considered empty if they contain no data. Numeric fields that contain 0 (zero) are not considered as empty.

Multi-value analysis

WizWhy can analyze the dependent variable either as Boolean (that is having two values) or as multi-value. When the dependent variable is categorical the multi-value analysis analyzes each value of the dependent variable. When the dependent variable is continuous *WizWhy* divides it into up to 9 intervals, and issues the analysis in regard to each of these intervals.

Necessary and sufficient conditions

See: if-and-only-if rules

Prediction's error probability

The degree to which the prediction is doubtful. The calculation of this figure is based on the *significance level* of the rules, the direction of the rules (if-then versus if-then-not) and the number of cases in a rule. The lower the error probability, the more certain the prediction is.

When the number of if-then rules is close to the number of if-then-not rules, the error probability rises. Such a case usually indicates that the

record for prediction have been taken from a population *not* similar to that of the data from which the rules were generated.

Prediction file

The file in which *WizWhy* updates the predictions when issuing predictions in another file.

Prediction input

The data set for issuing the predictions, when issuing predictions to another file.

Prediction power

An index calculated by *WizWhy*, that measures to what extent a given field explains the dependent variable. If one has to issue predictions by using one field only, this index designates which field is the best predictor, the second best, etc. When calculating this index, *WizWhy* takes into account the expected number of errors and the cost of errors (misses and false alarms).

Prediction's significance level

See: Prediction's error probability

Primary prediction's probability

The a priori probability of the *predicted value* in the data (from which the rules were generated). This is the probability of the predicted value, if no rule applies to the case under analysis.

Probability

See: Conclusive prediction's probability, Error probability, Expected probability, Prediction's error probability, Primary prediction's probability, Rule's probability.

Records with no relevant rules

Records where no rule applies. The field values in these records meet none of the conditions of the rules. When *WizWhy* issues predictions in regard to these records, it determines the prediction on the basis of the % frequency of the predicted value and the error costs.

The number of records with no relevant rules can usually be reduced by reducing the minimum number of cases in a rule, and the minimum probability of if-then and if-then-not rules.

Rule's probability

The percentage of cases in which both the "if" *and* the "then" sections hold, within the total number of cases in which the "if" sections holds.

Rule report

Lists the discovered rules together with an analysis of the rules explanatory power.

Significance level

See: Error Probability

Success rate

Success Rate when Predicting 1: The % success of the positive predictions. This number is calculated from the number of misses and the total number of positive predictions.

Success Rate when Predicting NOT 1: The % success of the negative predictions.

The proportion between the success rate of the positive and negative predictions is the result of the proportion between the price of a miss and the price of a false alarm.

Sufficient and necessary conditions

See: if-and-only-if rules

Trend report

Presents the one-condition relations in the data. Together with the Unexpected Rule Report it summarizes the data.

Unexpected rule report

Displays the rules that are unexpected relative to more basic rules and trends. These unexpected rules describe the interesting phenomena in the data.

Unlikelihhod

See: Level of unlikelihood

Quick Reference Guide

After you have familiarized yourself with *WizWhy*, you may wish to use this chapter to refresh your memory about how to perform certain *WizWhy* processes or how to use a certain function available through the *WizWhy* menus.

The *File* menu

This menu provides options for performing global operations on data files.

New opens a new *WizWhy* working file (*.wwr). This enables you to select a new data set for issuing rules.

Open opens a previously saved *WizWhy* working file.

Close closes the current *WizWhy* working file.

Save saves the current working file

Save As saves the current *WizWhy* working file with a new name.

Open Basic Data opens a data set for analysis. *WizWhy* will reveal the rules in this data set. The data type may be Text, dBase, MS Access, MS SQL, Oracle, ODBC or OLE DB.

Open Prediction Data opens a data set for updating the predictions.

Database Relationships enables you to edit the database connections (if several tables are selected).

Print prints the current report on the printer or exports it as an ASCII file or an MS Access table.

Print Setup enables you to change the printer settings.

Exit closes the *WizWhy* application.

The *Edit* menu

The contents of the Edit menu change according to the information currently displayed in the *WizWhy* screen. This section describes special Edit functions.

Update check boxes enables you to select / clear all the fields in the Analyze if Empty or Ignore Field column, and to change all the field types from Number to Category.

Select Rule enables you to select a rule in the Rule Report

Select Rules for SQL enables you to select rules to be exported to an SQL statement.

The *Issue* menu

This menu is used to generate reports and predictions.

Trend Report issues the Trend report only (without issuing the rules).

Rule Report issues all the reports (the Trend report, the Rule report, the Unexpected Rule report, the If-and-only-if Rules report, and the Unexpected Cases report).

Predict On Line issues a Prediction report on the screen based on conditions that you enter manually.

SQL Statement builds an SQL statement to be used in any database application (such as MS Access) to select records. To use this option, first use Edit - Select Rules to choose the rules to be exported.

Update Prediction to a File applies the rules on the records of the Prediction Input data set, and builds a new ASCII file that contains the predictions.

Appendix: The Mathematics behind *WizWhy*

This appendix briefly presents the formulas used to calculate some of the figures presented by *WizWhy*. The appendix is aimed at those who are interested in mathematics. If you are not, simply skip it. You don't have to be a mathematician in order to use *WizWhy* effectively.

The error probability of an if-then rule

We will use the following notations:

- m is the number of cases in the if-then rule;
- n is the number of records satisfying the rule's condition;
- N is the total number of records in the investigated file;
- M is the number of records where the dependent variable R is r;
- α is an error probability of this rule.

α is calculated as follows:

$$\alpha = \sum_{k=m}^{n} P_{N,M}(n,k),$$

where

$$P_{N,M}(n,k) = \frac{\binom{k}{M}\binom{n-k}{N-M}}{\binom{n}{N}}$$

The error probability α for the if-then-not rule is calculated as follows:

$$\alpha = \sum_{k=0}^{m} P_{N,M}(n,k)$$

Unexpected rules

Definition: A rule containing q-conditions ($q > 1$) in the "if" part, *R is r* in the "then" part, and having probability P that *R is r* is called *unexpected* if at least one of the following two requirements is fulfilled:

(1) There is a rule containing *R is not r* in the "then" part and q_1 -conditions ($q_1 < q$) in the "if" part such that the set of 1-conditions entered in the q_1-condition is a subset of the set of 1-conditions entered in the q-condition.

(2) There is no rule containing *R is r* in the "then" part and q_1-conditions ($q_1 < q$) in the "if" part (where the set of 1-conditions entered in the q_1-conditions is a subset of the set of 1-conditions entered in the q-conditions) for which the inequality $p << P$ is not fulfilled, where p denotes a probability of such a rule. (If $r = 1$, p is a probability that R is 1; if $r = 0$, p is a probability that R is 0.)

The second requirement means that the probability that R is r for any discovered rule with the above-defined q_1-conditions in the "if" part and R is r in the "then" part must be much less than P.

The unexpected rule parameters

Unlike other if-then rules, each Unexpected Rule revealed by *WizWhy* contains the following additional parameters:

1. Expected rule's probability
2. Level of unlikelihood

We will use the following notations:

N is the total number of records in the investigated file;
M is the total number of records where the Field to Predict R is r ;
p_a is the a priori probability that $R = r$, that is, $p_a = \frac{M}{N}$;
m_i is the number of records satisfying the ith 1-condition entered in the q-condition of the unexpected rule, $i = 1,\ldots,q$;

S_i is the number of records satisfying both the ith 1-condition and the condition R is r , $i = 1,\ldots,q;$;

p_i is the probability that R is r under the ith 1-condition, $p_i = \frac{s_i}{m_i}$

K is the number of records satisfying the q-condition of the unexpected rule;

S is the number of records satisfying both the q-condition of the unexpected rule and the condition R is r ;

$P_{\exp}$ is the expected probability for the considered unexpected rule;

U is the level of unlikelihood.

The *expected probability* $P_{\exp}$ is calculated as follows:

$$P_{\exp} = \frac{P_{dep} - P_{ind}}{k_{dep} - k_{ind}} \cdot (K - k_{ind}) + P_{ind} \ , \qquad \text{if} \quad k_{ind} \le K \le k_{dep} \ ;$$

$$P_{\exp} = P_{ind} \ , \qquad \text{if} \quad K < k_{ind} \ ,$$

where:

$$P_{dep} = \frac{\sum_{i=1}^{q} p_i}{q} \ ;$$

$$k_{dep} = \min_{i=1,\ldots,q} m_i \ ;$$

$$P_{ind} = \frac{1}{1 + \left(\frac{p_a}{1 - p_a} \right)^{q-1} \prod_{i=1}^{q} \frac{1 - p_i}{p_i}} \ ;$$

$$k_{ind} = \frac{\prod_{i=1}^{q} m_i}{N^{q-1}}$$

The *level of unlikelihood* U is calculated as follows:

$$U = \max_{i=1,\ldots,q} (1 - U_i) \ ,$$

where

$$U_i = \frac{\binom{S}{s_i} \cdot \binom{K-S}{m_i - s_i}}{\binom{K}{m_i}}$$

Prediction

When *WizWhy* issues a prediction on the basis of the if-then rules (and not the if-and-only-if rules), and the dependent variable was analyzed as Boolean, *WizWhy* lists the following parameters.

1. Weighted average rule probability
2. Prediction's significance level: error probability
3. Primary prediction's probability
4. Conclusive prediction's probability

All the rules applicable to the record under analysis are considered for calculating these parameters. Assume that n such rules have been found: the first k rules are of the "if-then" type, and the remaining $n-k$ rules are of the "if-then-not" type. For rule i let us denote the rule's probability, the error probability, and the number of cases in the rule by p_i, α_i, and m_i respectively, $i = 1,\ldots,n$.

The *weighted average rule probability* is calculated as follows: The significance level of rule i *is* taken as a weight for this rule's probability. The significance level of rule i is equal to $1-\alpha_i$. However, if the significance level is equal to 0.5, then it may be assumed that a measure of our confidence in the rule's validity is near to zero. Therefore, to evaluate the weight of the rule's probability, it is worthwhile to map the set of significance level's values changed within the segment [0.5 , 1] into segment [0 , 1]. Assume that this mapping is linear, and for $\alpha_i = 0$, the weight is equal to 1; for $\alpha_i = 0.5$ the weight is equal to 0. Then the weight

for the rule's probability with a priori probability α_i is equal to $1-2\alpha_i$. Thus, the weighted average rule probability P is:

$$P = \frac{\sum_{i=1}^{k}(1-2\alpha_i)\cdot p_i + \sum_{i=k+1}^{n}(1-2\alpha_i)\cdot(1-p_i)}{\sum_{i=1}^{n}(1-2\alpha_i)}$$

The *primary prediction's probability* is the average probability that the value of the dependent variable R is the predicted value r. The primary probability is presented to enable the user to evaluate the difference between the prediction's probability and the average probability. Let M be the frequency of value r of field R within the investigated file. Let N be the number of the investigated file's records. Then the primary probability P_a is equal to $\frac{M}{N}$.

The *prediction's significance level* means the measure of confidence that the prediction's probability P is true. The significance level is equal to $1-\alpha$, where α is the prediction's *error probability*. Error probability α is calculated as follows.

Assume that $P \geq P_\alpha$. The probability α_{pos} that at least one of k if-then rules is not erroneous is first calculated.

$$\alpha_{pos} = 1 - \prod_{i=1}^{k}\alpha_i$$

The probability α_{neg} that at least one of $n-k$ if-then-not rules is not erroneous is calculated as follows:

$$\alpha_{neg} = 1 - \prod_{i=k+1}^{n}\alpha_i$$

We assume that all the discovered rules are independent; that is, the complete set of events (cases) defined by all the rules is a union of non-intersecting sets, each of which is the set of events (cases) of a certain rule. Then, the error probability for the prediction is calculated as follows:

$$\alpha = 1 - \frac{\alpha_{pos} \cdot \sum_{i=1}^{k} n_i \; + (1 - \alpha_{neg}) \cdot \sum_{i=k+1}^{n} n_i}{\sum_{i=1}^{n} n_i} \; ,$$

where $n_i = \frac{m_i}{p_i}$.

If $P \geq P_a$, then α_{pos} is calculated for if-then-not rules, and α_{neg} is calculated for if-then rules.

The *conclusive prediction's probability* is the mathematical expectation *MP* of a probability *P* for the prediction. It is calculated as follows:

$$MP = (1 - \alpha) \cdot P + \alpha \cdot P_{\alpha} \, ,$$

where:

P is the weighted average rule probability
α is the prediction's error probability
P_a is the primary prediction's probability

The deduction of the formula can be explained as follows.

Consider the aggregation of all the relevant rules as the certain "generalized" rule. Let *A* be the following event: according to the given values of condition fields, this "generalized" rule has been found. The only parameter describing this event is the probability *P* that *R* = *r* (weighted average rule probability). We will consider this parameter as a random variable since we have found not only the value *P* of the parameter, but also the probability (equal to the "generalized" rule's significance level $1-\alpha$) of this value.

Consider the complete space of events as follows:

A is the event that the "generalized" rule has been found;

$\bar{A}$ is the event that the "generalized" rule has not been found.

The value of the above parameter (the probability that *R is r*) for the event $\bar{A}$ can be determined as follows: If the "generalized" rule was not found,

then this parameter's value is equal to the primary probability P_a. Since the event $\bar{A}$ is the complement of event A with respect to the whole space of events, the probability of value P_a is equal to α. The formula follows from the definition of a mathematical expectation of the considered random variable.

Index