Data Mining and the Case for Sampling

Solving Business Problems Using SAS® Enterprise Miner™ Software

A SAS Institute

Best Practices Paper

SAS Institute

Table of Contents

Figures

Abstract

Industry analysts expect the use of data mining to sustain double-digit growth into the 21st century. One recent study, for example, predicts the worldwide statistical and data mining software market to grow at a compound annual growth rate of 16.1 percent over the next five years, reaching $1.13 billion in the year 2002 (International Data Corporation 1998 #15932).

Many large- to mid-sized organizations in the mainstream of business, industry, and the public sector already rely heavily on the use of data mining as a way to search for relationships that would otherwise be "hidden" in their transaction data. However, even with powerful data mining techniques, it is possible for relationships in data to remain hidden due to the presence of one or more of the following conditions:

- data are not properly aggregated
- data are not prepared for analysis
- relationships in the data are too complex to be seen readily via human observation
- databases are too large to be processed economically as a whole.

All of these conditions are complex problems that present their own unique challenges. For example, organizing data by subject into data warehouses or data marts can solve problems associated with aggregation.[1] Data that contain errors, missing values, or other problems can be cleaned in preparation for analysis.[2] Relationships that are counter-intuitive or highly complex can be revealed by applying predictive modeling techniques such as neural networks, regression analysis, and decision trees as well as exploratory techniques like clustering, associations and sequencing. However, processing large databases en masse is another story — one that carries along with it its own unique set of problems.

This paper discusses the use of sampling as a statistically valid practice for processing large databases by exploring the following topics:

- data mining as a part of the "Business Intelligence Cycle"
- sampling as a valid and frequently-used practice for statistical analyses
- sampling as a best practice in data mining
- a data mining case study that relies on sampling.

For those who want to study further the topics of data mining and the use of sampling to process large amounts of data, this paper also provides references and a list of recommended reading material.

[1] Accessing, aggregating, and transforming data are primary functions of *data warehousing*. For more information on data warehousing, see the "Recommended Reading" section in this paper.

[2] *Unscrubbed data* and similar terms refer to data that are not prepared for analysis. Unscrubbed data should be *cleaned* (*scrubbed*, *transformed*) to correct errors such as missing values, inconsistent variable names, and inconsequential outliers before being analyzed.

The Overabundance of Data

In the past, many businesses and other organizations were unable or unwilling to store their historical data. Online transaction processing (OLTP) systems, rather than decision support systems, were key to business. A primary reason for not storing historical data was the fact that disk space was comparatively more expensive than it is now. Even if the storage space was available, IT resources often could not be spared to implement and maintain enterprise-wide endeavors like decision support systems.

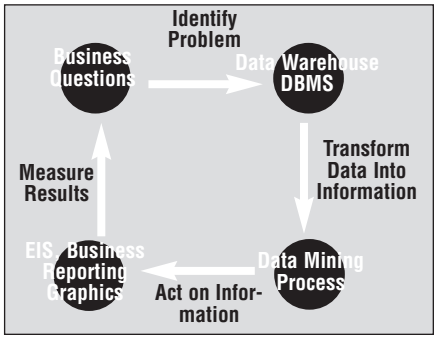
Times have changed. As disk storage has become increasingly affordable, businesses have realized that their data can, in fact, be used as a corporate asset for competitive advantage. For example, customers' previous buying patterns often are good predictors of their future buying patterns. As a result, many businesses now search their data to reveal those historical patterns.

To benefit from the assets bound up in their data, organizations have invested numerous resources to develop data warehouses and data marts. The result has been substantial returns on these kinds of investments. However, now that affordable systems exist for storing and organizing large amounts of data, businesses face new challenges. For example, how can hardware and software systems sift through vast warehouses of data efficiently? What process leads from data, to information, to competitive advantage?

While data storage has become cheaper, CPU, throughput, memory management, and network bandwidth continue to be constraints when it comes to processing large quantities of data. Many IT managers and business analysts are so overwhelmed with the sheer volume, they do not know where to start. Given these massive amounts of data, many ask, "How can we even begin to move from data to information?" The answer is in a data mining process that relies on sampling, visual representations for data exploration, statistical analysis and modeling, and assessment of the results.

Data Mining and the Business Intelligence Cycle

During 1995, SAS Institute Inc. began research, development, and testing of a data mining solution based on our world-renowned statistical analysis and reporting software — the SAS System. That work, which resulted in the 1998 release of SAS Enterprise Miner™ software, taught us some important lessons.[3] One lesson we learned is that data mining is a process that must itself be integrated within the larger process of business intelligence. Figure 1 illustrates the role data mining plays in the business intelligence cycle.



Integrating data mining activities with the organization's data warehouse and business reporting systems enables the technology to fit within the existing IT infrastructure while supporting the organization's larger goals of

- identifying business problems,
- transforming data into information,
- acting on the information, and
- assessing the results.

Figure 1 : The Data Mining Process and the Business Intelligence Cycle

[3] According to the META Group, "The SAS Data Mining approach provides an end-to-end solution, in both the sense of integrating data mining into the SAS Data Warehouse, and in supporting the data mining process. Here, SAS is the leader" (META Group 1997, file #594).

The SEMMA Methodology

SAS Institute defines data mining as the process used to reveal valuable information and complex relationships that exist in large amounts of data. Data mining is an iterative process — answers to one set of questions often lead to more interesting and more specific questions. To provide a methodology in which the process can operate, SAS Institute further divides data mining into five stages that are represented by the acronym SEMMA.

Beginning with a statistically representative sample of data, the SEMMA methodology — which stands for Sample, Explore, Modify, Model, and Assess — makes it easy for business analysts to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. Here is an overview of each step in the SEMMA methodology:

- **Sample** the data by creating one or more data tables.[4] The samples should be big enough to contain the significant information, yet small enough to process quickly.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- **Modify** the data by creating, selecting, and transforming the variables to focus the model selection process.
- **Model** the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.

SEMMA is itself a cycle; the internal steps can be performed iteratively as needed. Figure 2 illustrates the tasks of a data mining project and maps those tasks to the five stages of the SEMMA methodology. Projects that follow SEMMA can sift through millions of records[5] and reveal patterns that enable businesses to meet data mining objectives such as:

- Segmenting customers accurately into groups with similar buying patterns
- Profiling customers for individual relationship management
- Dramatically increasing response rate from direct mail campaigns
- Identifying the most profitable customers and the underlying reasons
- Understanding why customers leave for competitors (attrition, churn analysis)
- Uncovering factors affecting purchasing patterns, payments and response rates
- Increasing profits by marketing to those most likely to purchase

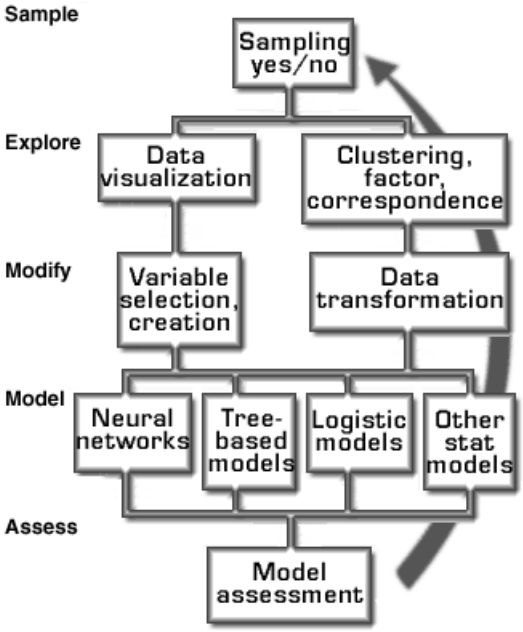


Figure 2 : Steps in the SEMMA Methodology

[4]The terms *data table* and *data set* are synonymous.

[5]*Record* refers to an entire row of data in a data table. Synonyms for the term *record* include *observation*, *case*, and *event*. *Row* refers to the way data are arranged horizontally in a data table structure.

- Decreasing costs by filtering out those least likely to purchase
- Detecting patterns to uncover non-compliance.

How Large is "A Large Database"

To find patterns in data such as identifying the most profitable customers and the underlying reasons for their profitability, a solution must be able to process large amounts of data. However, defining "large" is like trying to hit a moving target; the definition of "a large database" is changing as fast as the enabling technology itself is changing. For example, the prefix *tera*, which comes from the Greek *teras* meaning "monster," is used routinely to identify databases that contain approximately one trillion bytes of data. Many statisticians would consider a database of 100,000 records to be very large, but data warehouses filled with a terabyte or more of data such as credit card transactions with associated demographics are not uncommon. Performing routine statistical analyses on even a terabyte of data can be extremely expensive and time consuming. Perhaps the need to work with even more massive amounts of data such as a petabyte (2^{50}) is not that far off in the future (Potts 1997, p. 10).

So what can be done when the volume of data grows to such massive proportions? The answer is deceptively simple; either

- try to process the entire database, or
- process only a sample of it.

Processing the Entire Database

To move from massive amounts of data to business intelligence, some practitioners argue that automated algorithms with faster and faster processing times justify processing the entire database. However, no single approach solves all data mining problems. Instead, processing the entire database offers both advantages and disadvantages depending on the data mining project.

Advantages

Although data mining often presupposes the need to process very large databases, some data mining projects can be performed successfully when the databases are small. For example, all of the data could be processed when there are more variables[6] than there are records. In such a situation, there are statistical techniques that can help ensure valid results in which case, an advantage of processing the entire database is that enough richness can be maintained in the limited, existing data to ensure a more precise fit.

In other cases, the underlying process that generates the data may be rapidly changing, and records are comparable only over a relatively short time period. As the records age, they might lose value. Older data can become essentially worthless. For example, the value of sales transaction data associated with clothing fads is often short lived. Data generated by such rapidly changing processes must be analyzed often to produce even short-term forecasts.

Processing the entire database also can be advantageous in sophisticated exception reporting systems that find anomalies in the database or highlight values above or below some threshold level that meet the selected criteria.

---

[6] *Variable* refers to a characteristic that defines records in a data table such as a variable B_DATE, which would contain customers' birth dates. *Column* refers to the way data are arranged vertically within a data table structure.

If the solution to the business problem is tied to one record or a few records, then to find that subset, it may be optimal to process the complete database. For example, suppose a chain of retail paint stores discovers that too many customers are returning paint. Paint pigments used to mix the paints are obtained from several outside suppliers. Where is the problem? With retailers? With customers? With suppliers? What actions should be taken to correct the problem? The company maintains many databases consisting of various kinds of information about customers, retailers, suppliers, and products. Routine anomaly detection (processing that is designed to detect whether a summary performance measure is beyond an acceptable range) might find that a specific store has a high percentage of returned paint. A subsequent investigation discovers that employees at that store mix pigments improperly. Clearer instructions could eliminate the problem. In a case like this one, the results are definitive and tied to a single record. If the data had been sampled for analysis, then that single, important record might not have been included in the sample.

Disadvantages

Processing the entire database affects various aspects of the data mining process including the following:

Inference/Generalization

The goal of inference and predictive modeling is to apply successfully findings from a data mining system to new records. Data mining systems that exhaustively search the databases often leave no data from which to develop inferences. Processing all of the data also leaves no *holdout data* with which to test the model for explanatory power on new events. In addition, using all of the data leaves no way to validate findings on data unseen by the model. In other words, there is no room left to accomplish the goal of inference.

Instead, holdout samples must be available to ensure confidence in data mining results. According to Elder and Pregibon, the true goal of most empirical modeling activities is "to employ simplifying constraints alongside accuracy measures during model formulation in order to best generalize to new cases" (1996, p. 95).

Occasionally, concerns arise about the way sampling might affect inference. This concern is often expressed as a belief that a sample might miss some subtle but important niches — those "hidden nuggets" in the database. However, if a niche is so tiny that it is not represented in a sample and yet so important as to influence the big picture, the niche can be discovered: either by automated anomaly detection or by using appropriate sampling methods. If there are pockets of important information hidden in the database, application of the appropriate sampling technique will reveal them and will process much faster than processing the whole database.

Quality of the Findings

Using exhaustive methods when developing predictive models may actually create more work by revealing spurious relationships. For example, exhaustive searches may "discover" such intuitive relationships as the fact that people with large account balances tend to have higher incomes, that residential electric power usage is low between 2 a.m. and 4 a.m., or that travel-related expenditures increase during the holidays. Spending time and money to arrive at such obvious conclusions can be avoided by working with more relevant subsets of data. More findings do not necessarily mean quality findings and sifting through the findings to determine which are valid takes time and effort.

In addition to unreliable forecasts, exhaustive searches tend to produce several independent findings, each of which needs a corresponding set of records on which to base inferences. Faster search algorithms on more data can produce more findings but with less confidence in any of them.

Speed and Efficiency

Perhaps the most persistent problems concerning the processing of large databases are speed and cost. Analytical routines required for exploration and modeling run faster on samples than on the entire database. Even the fastest hardware and software combinations have difficulty performing complex analyses such as fitting a stepwise logistic regression with millions of records and hundreds of input variables. For most business problems, there comes a point when the time and money spent on processing the entire database produces diminishing returns wherein any potential modest gains are simply not worth the cost. In fact, even if a business were to ignore the advantages of statistical sampling and instead choose to process data in its entirety (assuming multiple terabytes of data), no time-efficient and cost-effective hardware/software solution that excludes statistical sampling yet exists.

Within a database, there can be huge variations across individual records. A few data values far from the main cluster can overly influence the analysis, and result in larger forecast errors and higher misclassification rates. These data values may have been miscoded values, they may be old data, or they may be outlying records. Had the sample been taken from the main cluster, these outlying records would not have been overly influential.

Alternatively, little variation might exist in the data for many of the variables; the records are very similar in many ways. Performing computationally intensive processing on the entire database might provide no additional information beyond what can be obtained from processing a small, well-chosen sample. Moreover, when the entire database is processed, the benefits that might have been obtained from a pilot study are lost.

For some business problems, the analysis involves the destruction of an item. For example, to test the quality of a new automobile, it is torn apart or run until parts fail. Many products are tested in this way. Typically, only a sample of a batch is analyzed using this approach. If the analysis involves the destruction of an item, then processing the entire database is rarely viable.

Processing a Sample

Corporations that have achieved significant return on investment (ROI) in data mining have done so by performing **predictive data mining**. Predictive data mining requires the development of accurate predictive models that typically rely on sampling in one or more forms. ROI is the final justification for data mining, and most often, the return begins with a relatively small sample.[7]

Advantages

Exploring a representative sample is easier, more efficient, and can be as accurate as exploring the entire database. After the initial sample is explored, some preliminary models can be fitted and assessed. If the preliminary models perform well, then perhaps the data mining project can continue to the next phase. However, it is likely that the initial modeling generates additional, more specific questions, and more data exploration is required.

[7]Sampling also is effective when using *exploratory* or *descriptive* data mining techniques; however, the goals and benefits (and hence the ROI) of using these techniques are less well defined.

In most cases, a database is logically a subset or a sample of some larger population.[8] For example, a database that contains sales records must be delimited in some way such as by the month in which the items were sold. Thus, next month's records will represent a different sample from this month's. The same logic would apply to longer time frames such as years, decades, and so on. Additionally, databases can at best hold only a fraction of the information required to fully describe customers, suppliers, and distributors. In the extreme, the largest possible database would be all transactions of all types over the longest possible time frame that fully describes the enterprise.

Speed and Efficiency

A major benefit of sampling is the speed and efficiency of working with a smaller data table that still contains the essence of the entire database. Ideally, one uses enough data to reveal the important findings, and no more. Sufficient quantity depends mostly on the modeling technique, and that in turn depends on the problem. "A sample survey costs less than a complete enumeration, is usually less time consuming, and may even be more accurate than a complete enumeration," as Saerndal, Swensson, and Wretman (1992, p. 3) point out. Sampling enables analysts to spend relatively more time fitting models and thereby less time waiting for modeling results.

The speed and efficiency of a process can be measured in various ways. Throughput is a common measure; however, when business intelligence is the goal, business-oriented measurements are more useful. In the context of business intelligence, it makes more sense to ask big picture questions about the speed and efficiency of the entire business intelligence cycle than it does to dwell on smaller measurements that merely contribute to the whole such as query/response times or CPU cycles.

Perhaps the most encompassing business-oriented measurement is one that seeks to determine the cost in time and money to go from the recording of transactions to a plan of action. That path — from OLTP to taking action — includes formulating the business questions, getting the data in a form to be mined, analyzing it, evaluating and disseminating the results, and finally, taking action.

Visualization

Data visualization and exploration facilitate understanding of the data.[9] To better understand a variable, univariate plots of the distribution of values are useful. To examine relationships among variables, bar charts and scatter plots (2-dimensional and 3-dimensional) are helpful. To understand the relationships among large numbers of variables, correlation tables are useful. However, huge quantities of data require more resources and more time to plot and manipulate (as in rotating data cubes). Even with the many recent developments in data visualization, one cannot effectively view huge quantities of data in a meaningful way. A representative sample gives visual order to the data and allows the analyst to gain insights that speed the modeling process.

Generalization

Samples obtained by appropriate sampling methods are representative of the entire database and therefore little (if any) information is lost. Sampling is statistically dependable. It is a mathematical science based on the demonstrable laws of probability, upon which a large part of statistics is built.[10]

---

[8] Population refers to the entire collection of data from which samples are taken such as the entire database or data warehouse.

[9] Using data visualization techniques for exploration is Step 2 in the SEMMA process for data mining. For more information, see the sources listed for data mining in the "Recommended Reading" section of this paper.

[10] For more information on the statistical bases of sampling, see the section "The Statistical Validity of Sampling" in this paper.

Economy

Data cleansing (detecting, investigating, and correcting errors, outliers, missing values, and so on) can be very time-consuming. To cleanse the entire database might be a very difficult and frustrating task. To the extent that a well-designed data warehouse or mart is in place, much of this cleansing has already been addressed. However, even with "clean" data in a warehouse or mart, additional pre-processing may be useful for the data mining project. There may still be missing values or other fields that need to be modified to address specific business problems. Business problems may require certain data assumptions, which indicate the need for additional data preparation. For example, a missing value for DEPENDENTS actually could be missing, or the value could be "zero." Rather than leave these values as missing, for analytical purposes, you may want to impute a value based on the available information.

If a well-designed and well-prepared data warehouse or mart is in place, less data pre-processing is necessary. The remaining data pre-processing is performed as needed for each specific business problem, and it is much more efficiently done on a sample.

Data augmentation (adding information to the data such as demographics, credit bureau scores, and so on), like data cleaning, will be less expensive if applied only to a sample. If the analysis is complicated and computationally intensive, it may be cost-effective to run a pilot study to see if further analysis is warranted. If additional data are to be purchased to augment the currently available data, then a test study on one part of the data might be an excellent approach. For example, the additional data for one geographic region could be purchased. Then as a pilot study, the full analysis performed on that one geographic region. After assessing the pilot study results, the decision could be made about purchasing additional data.[11]

In general, if a small-scale pilot study (based on a representative sample) provides useful results, then the costs of performing a larger, more comprehensive study may be a worthwhile investment. If the pilot study does not provide useful results, then it may be better to stop this line of analysis and go to the next problem.

Disadvantages

Not all samples are created equal. To be representative, a sample should reflect the characteristics of the data. Business analysts must know the data well enough to preserve the important characteristics of the database. In addition, the technology must be robust enough to perform various sampling techniques, because different sampling techniques are appropriate in different situations.

The Statistical Validity of Sampling

Statistics is a branch of applied mathematics that enables researchers to identify relationships and to search for ways to understand relationships. Modern statistical methods rely on sampling. Analysts routinely use sampling techniques to run initial models, enable exploration of data, and determine whether more analysis is needed. Data mining techniques that

[11] SAS Institute Inc. has agreements with a number of data providers including Claritas Inc., Geographic Data Technologies, and Acxiom Corporation.

use statistical sampling methods can reveal valuable information and complex relationships in large amounts of data — relationships that might otherwise be hidden in a company's data warehouse.

Size and Quality Determine the Validity of a Sample

In data mining, the sample data is used to construct predictive models, which in turn, are used to make predictions about an entire database. Therefore, the validity of a sample, that is, whether the sample is representative of the entire database, is critically important.

Whether a sample is representative of the entire database is determined by two characteristics — the *size* of the sample and the *quality* of the sample. The importance of sample size is relatively easy to understand. However, understanding the importance of quality is a bit more involved, because a quality sample for one business problem may not be a quality sample for another problem.

How the Size of Samples Affects Validity

As the size of the sample grows, it should with increasing clarity reflect the patterns that are present in the entire database. When we compare a graphical representation of a sample with one of a database, we readily can see how samples are able to reflect the patterns or *shapes* of databases.

For example, Figure 3 graphs customer income levels. Graph A represents the entire database (100 percent of the records), and reveals that most of the customers are of middle income. As you go toward the extremes — either very low or very high income levels, the number of customer records declines. Graph B, which is a sample of 30 percent of the database, reveals the same overall pattern or shape of the database. Statisticians refer to this pattern as the *distribution* of the data.

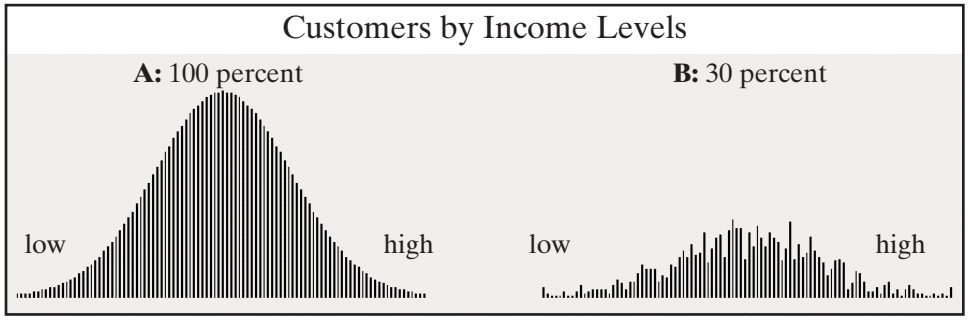


Figure 3 : How Sampling Size Affects Validity

If we were to increase the size of the sample, it would continue to take on the distribution of the database. Taken to a logical extreme, if we read the entire database, we have the largest, most comprehensive "sample" available — the database itself. The distribution would, of course, be the same, and any inferences we make about that "sample" would be true for the entire database.[12]

[12]For more information about the size of samples, see the section "Sampling as a Best Practice in Data Mining: Determining the Sample Size" in this paper.

How the Quality of Samples Affects Validity

Quality, in the context of statistical sampling techniques, refers to whether the sample captures the characteristics of the database that are needed to solve the business problem. Is the sample in fact an ideal representation of the data at large? The highest quality sample would be an exact miniature of the database; it would preserve the distributions of individual variables and the relationships among variables. At the other extreme, the lowest quality sample would be so unrepresentative or biased in some direction as to be of no use. In practice, a sample should be unbiased enough to be at least typical of the database.

But how is it possible to construct unbiased samples? How can we ensure that the samples used in statistical analyses such as data mining projects are representative of the entire database? The answer lies in the procedures used to select records from the database. The sampling selection procedures determine the likelihood that any given record will be included in the sample. To construct an unbiased sample, the procedure must be based on a proven, quantifiable selection method — one from which reliable statistics can be obtained. In the least sophisticated form of sampling, each record has the same probability of being selected.

Randomness: the Key to Quality Samples

The concept of randomness and the role it plays in creating unbiased samples can be seen in a technique that is fundamental to statistical analysis. That technique is *simple random sampling*. In this context, *simple* means that the records are selected from the database in their "simplest form;" that is, the records are not preprocessed or grouped in some manner prior to sampling (Phillips 1996, p. 91).

Random in this context does not mean haphazard or capricious. Instead, random refers to a lack of bias in the selection method. In a random sample, each record in the database has an equal chance of being selected for inclusion in the sample. Random sampling in a sense levels the playing field for all records in a database giving each record the same chance of being chosen, and thereby ensuring that a sample of sufficient quantity will represent the overall pattern of the database. For example, to create a simple random sample from a customer database of one million records, you could use a selection process that would assign numbers to all of the records — one through one million. Then the selection process could randomly select numbers (Hays 1973, pp. 20-22).

Random Samples Reveal the Distribution of the Data

When we randomly select records from a database and obtain a sample of sufficient size, the sample reflects characteristics of the database as a whole. As shown previously in Figure 3, as the sample size gets larger, the distribution (shape) of the sample reflects the distribution of the entire database. In Figure 3, the data have a symmetric and bell-shaped distribution. This is a commonly occurring and well-understood type of shape and is referred to as a *normal distribution.*

Not all data are distributed normally. Graphical representations of databases often reveal that the data are skewed in one direction or the other. For example, if in Figure 3, most of the customers had relatively lower incomes, then the distribution would be skewed to the right. If these data were randomly sampled, then as the sample size gets larger, its distribution would reflect better the distribution of the entire database.

For example, assume that we have a customer database of one million records from a national chain of clothing stores. The chain specializes in athletic wear with team logos. As a result of that specialization, 2 out of 3 of the records in the database are for customers between ages 8 and 28, who prefer that style of clothing.

Figure 4 shows the distribution of the database and a sample.

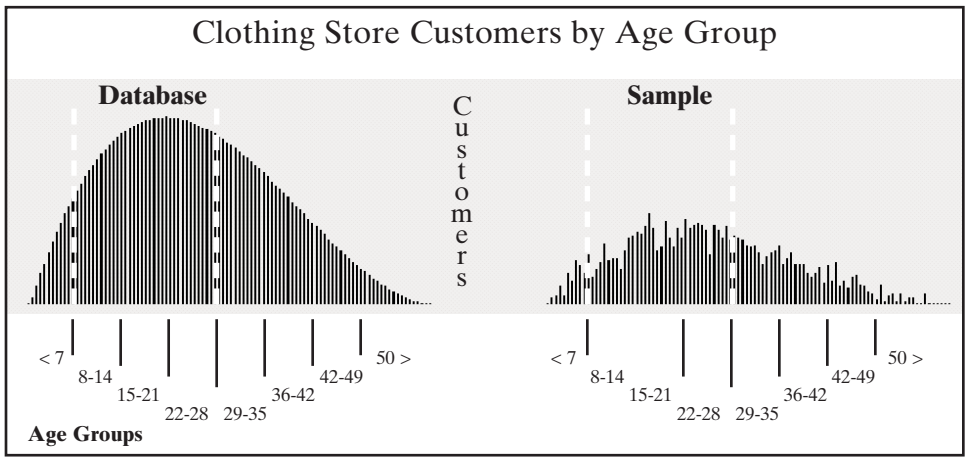


Figure 4 : How Samples Reveal the Distribution of Data

As you might expect, when you randomly select records from the database, you are more likely to obtain records for customers age 8 to 28 than you are likely to obtain other records simply because, proportionately, there are more 8- to 28-year old customers. Perhaps you might not obtain a record for that age group the first time or even the second or third time, but as more records are selected, your sample would eventually contain the 2 out of 3 ratio.[13]

If we extend the logic behind random samples to a more complex data mining scenario, we can see how random samples are the foundation for the exploration, analysis, and other steps in the SEMMA methodology. For example, if

- 59 percent of the 8- to 28-year-old customers were males, and
- 23 percent of them bought at least one jacket and a copy of a sports-related magazine as a part of the same purchase, and
- they paid for their purchases in cash, then

you could expect a random sample of sufficient size to reveal those tendencies as well.

Random Samples Reveal Other Characteristics in the Data

The distribution of the data is not the only characteristic revealed by sample data. When mathematical formulas based on the laws of probability are applied to a sample, the sample data can reveal other characteristics including summary statistics such as the mean (statistical average), median (the midpoint), and mode (the most frequently occurring value). In addition, other formulas can be applied to the sample to obtain measures of the spread or variability of the data, including the *standard deviation* upon which *confidence levels* are based.

[13]The laws of probability are based in part on the notion of randomly sampling an infinite number of times. Hays provides a good explanation of this basis in the section titled “In the Long Run” (1981, pp. 22-25).

Stated simply, *standard deviation* is the average distance of the data from the mean. *Confidence level* refers to the percentage of records that are within a specified number of standard deviations from the mean value. For a normal (symmetric, bell-shaped) distribution, approximately 68 percent of all records in a database will fall within a range that is 1 standard deviation above and below the mean. Approximately 95 percent of all records will fall within a range that is 2 standard deviations above and below the mean (Hays 1981, p. 209).

In summary, when we have a database and use an unbiased selection process to obtain records, then at some point as more and more records are selected, the sample will reveal the distribution of the database itself. In addition, by applying mathematical formulas based on laws of probability, we can determine other characteristics of the entire database. These sampling techniques and their underlying principles hold true for all populations regardless of the particular application. It does not matter whether we are rolling dice, blindly pulling red and white marbles from a barrel, or randomly selecting records from a year's worth of customer transaction data in order to construct a sample data table for use in a data mining project.[14]

Current Uses of Sampling

Sampling has grown into a universally accepted approach for gathering information, and it is widely accepted that a fairly modest-sized sample can sufficiently characterize a much larger population. Sampling is used extensively by governments to identify problem areas and national trends, as well as to measure their scope and scale. For example, sampling is used in the United States to measure and to characterize unemployment and the size of the labor force, industrial production, wholesale and retail prices, population health statistics, family incomes and expenditures, and agricultural production and land use. As Cochran (1977, p. 3) points out in *Sampling Techniques*, sampling associated with the U.S. national census was introduced in 1940 and its use in survey questions has grown steadily:

> Except for certain basic information required for every person for constitutional or legal reasons, the whole census was shifted to a sample basis. This change, accompanied by greatly increased mechanization, resulted in much earlier publication and substantial savings.

Today, the U.S. government publishes numerous reports based on sample data obtained during censuses. For example, sampling conducted during the 1990 census was used as the basis for a variety of reports that trace statistics relating to social, labor, income, housing, and poverty (U.S. Bureau of the Census 1992).

Sampling is also used by local governments and by commercial firms. Television and radio broadcasters constantly monitor audience sizes. Marketing firms strive to know customer reactions to new products and new packaging. Manufacturing firms make decisions to accept or reject whole batches of a product based on the sample results. Public opinion and elections polls have used sampling techniques for decades.

Sampling of an "untreated" group can provide a baseline for comparing the "treated" group, and hence for assessing the effectiveness of the treatment. Researchers in various fields routinely conduct studies that examine wide-ranging topics including human behavior and health; sampling is often an integral part of these analyses.

---

[14]The techniques used in sampling (and the mathematical formulas that statisticians use to express those techniques) grew out of the study of probability. The laws of probability grew out of the study of gambling — in particular, the work of the 17th century mathematicians Blaise Pascal and Pierre de Fermat who formulated the mathematics of probability by observing the odds associated with rolling dice. Their work is the basis of the theory of probability in its modern form (Ross 1998, p. 89).

When Sampling Should Not Be Used

Sampling is useful in many areas of research, business, and public policy, but there are a few areas in which sampling is not a recommended practice. For example, sampling is not recommended under the following conditions:

- When exact dollar and cents accounting figures are required. For example, in systems that track asset/liability holdings, each individual account and transaction would need to be processed.
- When the entire population must be used. For example, the U.S. Constitution states that an "actual enumeration" of the U.S. population is to be performed (Article 1, Section 2, Paragraph 3).[15]
- When the process requires continuous monitoring. For example, for seriously ill medical patients, in precision manufacturing possesses, and for severe weather over airports.
- When performing auditing in which every record must be examined such as in an audit of insurance claims to uncover anomalies and produce exception reports.

Myths about Sampling

Despite sampling's scientific background, many people still have reservations about the validity of sampling. One well-known incident often cited as an example of the unreliability of sampling is the newspaper-published predictions of the outcome of the 1948 U.S. presidential campaign in which the Chicago Daily Tribune used samples of voters to report that Dewey had beaten Truman. Of course, Truman proved the "experts" wrong by winning the election. The photograph of Truman smiling at a victory celebration while holding on high the headlines that read, "Dewey Defeats Truman" was seen by millions then, and that image persists today. But what's wrong with that picture of sampling?

The problem was two-fold. First, in 1948 the science of using sampling in political polls was still in its infancy. Instead of using random sampling techniques, which would have given each voter equal opportunity to be polled, the pollsters at that time constructed their sample by trying to match voters with what they assumed was America's demographic makeup. This "technique" led the pollsters to choose people on the basis of age, race, and gender thereby skewing the results (Gladstone 1998).

The second error was one of timing that affected the quality of the sample data. Specifically, the surveys upon which the Dewey/Truman prediction was made ended at least two weeks before Election Day. Given the volatile political landscape of the late 1940's, those two weeks were more than enough time for a major shift in votes from Dewey to Truman. (McCullough 1992, p. 714).

Despite the now well-documented problems of the 1948 poll, myths about the validity of sampling persist. When evaluating the applicability of sampling to business intelligence technology such as data mining, the skepticism is often expressed in one of the following ideas:

---

[15]The Constitution reads, "The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct." As the year 2000 approached, the legality and political ramifications of using statistical sampling for the Decennial Census were debated publicly and in court. For example, in its report to the President of the United States, the American Statistical Association has argued in favor of using sampling to mitigate the inevitable undercount of the population (American Statistical Association 1996). However, an August 24, 1998 ruling by a U.S. federal court upheld the constitutional requirement for enumeration (U.S. House of Representatives v. U.S. Department of Commerce, et al. 1998). The federal court ruling has been appealed to the U.S. Supreme Court (Greenhouse 1998).

- Sampling misses important information.
- Sampling is difficult.
- Lots of hardware lets you avoid sampling.

Myth 1: Sampling Misses Important Information

The concern that sampling misses important information stems from the fact that data usually contain *outlying records* (values that are far from the main cluster). Consider a proposed mail campaign in which we want to randomly sample the database, build a model, and then *score*[16] the database. If the database contains one extraordinary customer who spends far more than any other customer, then there are implications for the model's scoring ability depending on whether the extraordinary customer is included in the sample.

If the extraordinary customer's record is included, the model might be overly optimistic about predicting response to the mailing campaign. On the other hand, if the record is not included, the model may yield more realistic prediction for most customers, but the score for the extraordinary customer may not reflect that customer's true importance.

In fact, by not sampling, important information can be missed because some of the most interesting application areas in data mining require sampling to build predictive models. Rare-event models require enriched or weighted sampling to develop models that can distinguish between the event and the non-event.[17] For example, in trying to predict fraudulent credit card transactions, the occurrence of fraud may be as low as 2 percent; however, that percentage may represent millions of dollars in write-offs. Therefore, a good return on the investment of time and resources is to develop a predictive model that effectively characterizes fraudulent transactions, and helps the firm to avoid some of those high-dollar write-offs.

There are many sophisticated modeling strategies from which to choose in developing the model. A critical issue of these strategies involves which one to use: A sample? An enriched sample? Or the entire database? If a simple random sample of the database is used, then it is likely that very few of the rare events (fraudulent transactions) will be included in the sample. If there are no fraudulent transactions in the sample that is used to develop the model, then the resulting model will not be able to characterize which transactions are fraudulent and which are not.

If the entire database is used to *train*[18] a model and the event of interest is extremely rare, then the resulting model may be unable to distinguish between the event and the non-events. The model may correctly classify 99.95 percent of the cases, but the 0.05 percent that represent the rare events are incorrectly classified.

Also, if the entire database is used to train the model and the event of interest is not rare, then it may appear to be trained very well, in fact it may be "over trained" (or over fitted). An over-trained model is trained not only to the underlying trends in the data, but unfortunately, it is also trained to the specific variations of this particular database. The model may predict this particular database very well, but it may be unable to correctly classify new

---

[16] *Scoring* is the process of applying a model to new data to compute outputs. For example, in a data mining project that seeks to predict the results of the catalog mailing campaign, scoring the database might predict which recipients of the catalog will purchase which goods and in what amounts.

[17] Also referred to as *case-control* sampling in biometrics and *choice-based* sampling in econometrics. For more information on the use of sampling in biometrics and econometrics, see Manski and McFadden, (1981) and Breslow and Day (1980) respectively.

[18] *Training* (also known as *fitting*) a model is the mathematical process of calculating the optimal parameter values. For example, a straight line is determined by two parameters: a slope parameter and an intercept parameter. A linear model is trained as these parameter values are calculated.

transactions (those not currently in the database). A serious problem of using the entire database to develop the model is that no records remain with which to test or refine the model's predictive capabilities.

By contrast, if an enriched sample is used to develop the model, then a larger percentage of the rare fraudulent transactions are included in the sample, while a smaller percentage of the non-fraudulent transactions are included. The resulting sample has a larger percentage of fraudulent cases than the entire database. The resulting model may be more sensitive to the fraudulent cases, and hence very good at characterizing fraudulent transactions. Moreover, the model can also be tested and refined on the remaining records in the database (those not used to develop the model). After the model is fully developed, then it can be field-tested on new transaction data.

Another concern about sampling is that a technique may be inappropriately applied. The problem is if an inappropriate sampling technique is applied to a data mining project, then important information needed to solve the business problem may be overlooked and thereby excluded from the sample. To ensure that all relevant information is included, business analysts must be familiar with the data as well as the business problem to be solved. To illustrate, again consider a proposed mailing campaign. If a simple random sample were selected, then important information like geographic region may be left out. There may be a strong interaction between timing of when the catalog is to be mailed and where it is to be mailed. If the proposed catalog is to contain cold weather goods and winter apparel, then a simple random sample of the national database may be inappropriate. It may be much better to stratify the random sampling on region, or even to exclude the extreme south from the mailing.

Myth 2: Sampling is Difficult

It is sometimes argued that sampling is too difficult due to the size and other characteristics of large databases. For example, some argue that no one can completely understand which variables are important and what interactions are present in massive amounts of data.

Without the use of modern software technologies, this argument has merit. However, software that is designed to enable modern statistical sampling techniques can assist business analysts in understanding massive amounts of data by enabling them to apply the most effective sampling techniques at the optimal time in the data mining process. For example, easy-to-use but powerful graphical user interfaces (GUIs) provide enhanced graphical capabilities for data visualization and exploration. GUIs built on top of well-established yet complex statistical routines enable practitioners to apply sampling and analytical techniques rapidly and make assessments and adjustments as needed.

Myth 3: Lots of Hardware Lets You Avoid Sampling

Another common misconception about sampling is that if you have enough hardware resources, you can avoid sampling altogether. This notion is based in part on the fact that improvements in technology have been substantial in recent years. In particular, disk space and memory have become vastly more affordable than in years past. However, certain technological constraints remain including the following:

- network bandwidth
- throughput
- memory management
- load balancing of CPUs.

From a perspective of the resources required, data mining is the process of selecting data (network bandwidth and throughput), exploring data (memory), and modeling data (memory management and CPU) to uncover previously unknown information for a competitive advantage.

Another sampling myth related to hardware resources is the idea that parallel processing is a requirement for data mining. In fact, in its worst incarnation, this myth states that parallel processing is a panacea for the problems of mining massive amounts of data. Although parallel processing can increase the speed with which some data processing tasks are performed, simply applying parallel processing to a data mining project ignores the fact that data mining is not merely a technical challenge. Instead, along with hardware and software challenges, the massive data processing tasks known collectively as "data mining" have as their impetus logical, business-oriented challenges. From the business perspective, data mining is a legitimate investment — one that is expected to provide healthy ROI — because of the way it supports the business goals of the organization.

To support business goals, data mining must itself be understood and practiced within a logical process such as the SEMMA methodology. Simply addressing a portion of the technical challenge by adding parallel processors ignores the fact that many of the constraints on a data mining project can recur throughout the process. For example,

- sampling can be I/O intensive as can variable selection
- some exploratory steps can be memory-intensive
- the modeling steps can be very CPU-intensive.

Often, the entire process or a sub-process is repeated sequentially. As a result, simply adding parallel processors will not deliver faster results. In terms of processing overhead, the most restrictive constraints in the business intelligence cycle are the input/output (I/O) constrained operations, not the central processing unit (CPU) constrained operations. Therefore, threading or "paralleling" the I/O operations may achieve more efficient processing. However, adding more CPUs might provide improvement only in selected steps of the data mining process. For example, the modeling step might benefit from additional CPUs because fitting a large, complex statistical model is a CPU-intensive operation.

Paralleling the CPU operations also should raise some other concerns. In particular, more threads can create conflicting demands for critical system resources such as physical memory. Beyond the physical memory problem is the problem of matching the workloads for the various threads in the parallel environment. If the workload per thread is not properly matched, then the parallel algorithm simply uses more CPU time and does not reduce the elapsed time to deliver the results.

Sampling as a Best Practice in Data Mining

Sampling is not new to businesses that rely on analytics. For decades, organizations in business, industry, government, and academia have relied on sampling for statistical analyses. As Westphal and Blaxton (1998, p. 85) point out, extracting data from a database for the purpose of data mining is based on the sampling techniques routinely used in surveys:

> What you are doing with an extraction is taking a representative sample of the data set. This is similar to the way in which statistical sampling traditionally has

been performed on large populations of observations. When surveyors acquire information from a general population, they sample only enough of that population to get a good approximation. You do not have to identify every single occurrence of a pattern within a data set in order to infer that the pattern exists. Once you lock onto a pattern you can get a feel for how extensive the pattern is throughout the entire data set through alternative reporting methods. Remember that you are performing data mining, not generating reports. Do not feel that you need to process the entire data set at one time. There will usually be more than enough results from the segmented data to keep you busy. We have found important patterns using as little as several hundred records. It is not the size of your data set that counts, but the way in which you use it. Keep in mind that in a well-constructed data mining environment, you will have access to all of the data that you need by making iterative extractions in a series of steps.

In the comparatively new discipline of data mining, the use of statistical sampling poses some new questions. This section addresses several of the more common concerns about how best to use sampling in data mining projects. In particular, this section addresses the following questions:

- What needs to be done to prepare data for sampling?
- What are the common types of sampling that apply to data mining?
- How should the size of the sample be determined?
- What are some general sampling strategies that can be used in data mining?
- Should multiple samples be taken for special purposes such as validation and testing?
- What considerations exist when sampling small data tables?

Preparing the Data for Sampling

Prior to sampling data and analyzing business problems, most businesses create some form of central data storage and access facility, which is commonly referred to as a data warehouse. Employees in various departments in the business will expect to access the data they need quickly and easily. Data warehouses enable many groups to access the data, facilitate updating the data, and improve efficiency of checking the data for reliability and preparing the data for analysis and reporting.

For example, if the data mining problem is to profile customers, then all of the data for a single customer should be contained in a single record. If you have data that describes a customer in multiple records, then you could use the data warehouse to rearrange the data, prior to sampling.

Common Types of Sampling

Types of sampling commonly used in data mining projects include the following:

Simple Random Sampling

Each data record has the same chance of being included in the sample.

N-th Record Sampling

Each *n*-th record is included in the sample such as every 100th record. This type of sampling is also called *systematic sampling*. For structured data tables, only a portion of the structure may be captured.

First *N* Sampling

The first *n* records are included in the sample. If the database records are in random order, then this type of sampling produces a random sample. If the database records are in some structured order, then the sample may capture only a portion of that structure.

Cluster Sampling

Each cluster of database records has the same chance of being included in the sample. Each cluster consists of records that are similar in some way. For example, a cluster could be all of the records associated with the same customer, which indicates different purchases at various times.

Stratified Random Sampling

Within each stratum, all records have the same chance of being included in the sample. Across the strata, records generally do not have the same probability of being included in the sample.

Stratified random sampling is performed to preserve the strata proportions of the population within the sample. In general, categorical variables are used to define the strata. For instance, gender and marital status are categorical variables that could be used to define strata. However, avoid stratifying on a variable with too many levels (or too few records per level) such as postal codes, which can have many thousands of levels.

Determining the Sample Size

To help determine the appropriate size of a sample given a particular data table, statisticians have developed mathematical formulas.[19] The formulas are designed to help the researcher select the optimal sample size by addressing questions such as

- What is the target variable?
- Which variables should be in the model?
- What is the functional form of the model (linear, linear with interaction terms, nonlinear, and so on)?
- What is an acceptable level of accuracy for the results?

If the answers to these questions are known, then sampling theory may be able to provide a reasonably good answer to the required sample size. The less confidence you have in the answers to these questions, the more you are into exploration of the data and iterating through the SEMMA process.

The specifics of the statistical formulas used to determine optimal sample sizes can be complex and difficult for the layperson to follow, but it is possible to generalize about the factors one needs to consider. Those factors are

- the complexity of data,
- the complexity of model, and
- the appropriateness of the data to the model.

[19]For sources that include formulas for determining sample sizes, see Cochran (1977, pp. 72 ff) and Snedecor and Cochran (1989, pp. 52-53, and 438-440).

Complexity of Data

The first step to understanding the complexity of the data is to determine what variable or variables are to be modeled. Very different models can be developed if the target variable is a continuous variable (such as "amount of purchase"), rather than a two-level variable (such as one that indicates "purchase" or "no purchase"). In a well-defined analysis, the researcher will likely have some prior expectations and some confidence in those expectations.

Depending on the question or questions being asked, the outliers may be the most informative records or the least informative records. Even in the worst case, you should have some idea of what the target variable is when doing predictive data mining. If the target consists of a rare event and a commonly occurring event, then you can stratify on it. If there are multiple records, you can perform cluster sampling, and so on. You should explore the data well enough to identify the stratification variables and outlying records. It may be important to stratify the sample using some category such as customer gender or geographic region. For some business problems, such as fraud detection, the outlying records are the most informative and should be included in the sample. However, for other problems, such as general trend analysis, the outlying records are the least informative ones.

While some level of complexity exists in most data tables, many business problems can be effectively addressed using relatively simple models. Some solutions necessitate limiting the model complexity due to outside constraints. For example, regulators might require that the values of the model parameters be no more complex than necessary so that the parameters can be easily understood.[20]

The Complexity of the Model

Model complexity can range from the simple to the very intricate.[21] Models such as regressions, decision trees, and neural networks can range from simple, almost intuitive designs to designs that are complex and difficult to grasp. Linear regression models can include many variables that are linearly related to the target variable. Each linearly related variable has an associated slope parameter that has to be estimated. Enough records have to be included so that the parameters can be estimated.

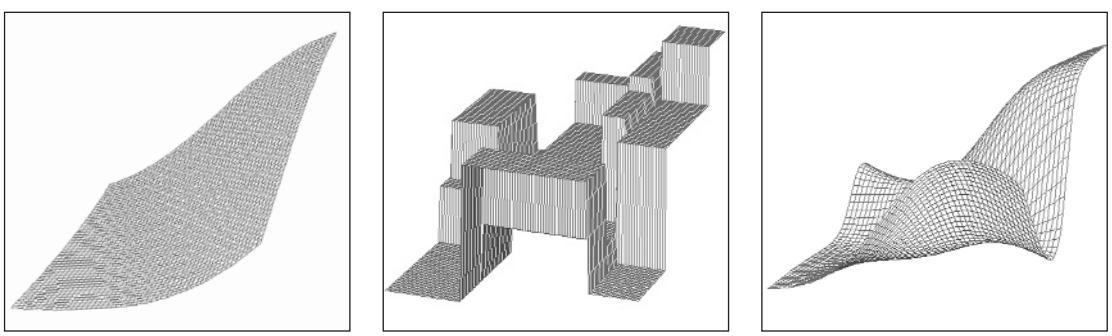


Figure 5 : Example Surface Plots for Fitted Models; Regression, Decision Tree, and Neural Network

[20]Modeling can benefit from the application of "Ockham's Razor," a precept developed by the English logician and philosopher William of Ockham (circa. 1285 to 1349), which states that "entities ought not to be multiplied except of necessity." (Gribbon 1996, p. 299).

[21]A simple regression model has one input variable linearly related to an output variable. $Y = a + bX$. More complex regression models include more variables, interaction terms, and polynomials of the input variables. A simple neural network having no hidden layers is equivalent to a class of regression models. More complex neural networks include nonlinear transformations of the input variables, hidden layers of transformations, and complex objective functions. A simple decision tree has only a few branches (splits) of the data before reaching the terminal nodes. More complex decision trees have many splits that branch through many layers before reaching the terminal nodes.

Decision tree analysis and clustering of records are iterative processes that sift the data to form collections of records that are similar in some way.

Neural networks are still more complex with nonlinear relationships, and many more parameters to estimate. In general, the more parameters in the model, the more records are required.[22]

The Appropriateness of Data to the Model

Different statistical models are appropriate for different types of data. For example, the business question might be: "How much can we expect the customer to purchase?" This question implies the need for a continuous target variable, which can contain a wide range of monetary values. Linear regression models might predict customer purchases quite accurately, especially if the input variables are linearly related to the target variable. If the input variables are nonlinearly related to the target variable, and they have complex interrelationships among themselves, then a neural network — or possibly a decision tree — might make more accurate predictions. If there are many missing values in the data table, then decision trees might provide the most accurate predictions.

Some modeling questions are easier to answer after exploring a sample of the data. For example, are the input variables linearly related to the target variable? As knowledge about the data is discovered, it may be useful to repeat some steps of the SEMMA process. An initial sample may be quite useful for data exploration. Then, when the data are better understood, a more representative sample can be use for modeling.

General Sampling Strategies

Some sampling strategies are routine in nature and can be used as general guidelines. For example, if you only know the target variable, then take an initial sample for exploratory purposes. A simple random sample of 1 percent of a large database may be sufficient. If this sample proves acceptable, then the sample analysis results should generalize to the entire database. If the data exploration reveals different response rates among the levels of variables, then the variables can be used to create strata for stratified random sampling.

The structure of the records may also influence the sampling strategy. For example, if the data structure is *wide* (data containing more variables for each record than individual records), then more variables have to be considered for stratification, for inclusion in the model, for inclusion in interaction terms, and so on. A large sample may be needed. Fortunately, some data mining algorithms (CHAID and stepwise regression, for example) automatically assist with variable reduction and selection.

By contrast, if the data structure is *deep* (data containing a few variables and many individual records), then as the sample size grows, more patterns and details may appear. Consider sales data with patterns across the calendar year. If only a small number of records are selected, then perhaps, only quarterly patterns appear; for example, there is a winter sales peak and a summer sales slump. If more records are included in the sample, perhaps monthly patterns begin to appear followed by weekly, daily, and possibly even intra-day sales patterns. As a general sampling strategy, it may be very useful to first take a relatively wide sample to search for important variables (inputs), and then to take a relatively deep sample to model the relationships.

[22]For more information on neural networks, see Sarle 1997.

Using Sample Data for Training, Validation, and Testing

An especially beneficial sampling practice is to *partition* (split) the sample into three smaller data tables that are used for the following purposes:

- training
- validation
- testing.

The training data table is used to train models, that is, to estimate the parameters of the model. The validation data table is used to fine tune and/or select the best model. In other words, based on some criteria, the model with the best criteria value is selected. For example, smallest mean square forecast error is an often-used criterion. The test data table is used to test the performance of the selected model. After the best model is selected and tested, it can be used to score the entire database (Ripley 1996, p 354).

Each record in the sample can appear in only one of the three smaller data tables. When you partition the sample data table, you may want to use the simple random sampling technique again, or stratified random sampling may be appropriate. So first, you might randomly select a small fraction of the records from a 5-terabyte database, and in general, simpler models require smaller sample sizes and more complex models require larger sample sizes. Then secondly, you might use random sampling again to partition the sample: 40 percent training, 30 percent validation, and 30 percent test data tables.

Sampling and Small Data Tables

Sampling small data tables requires some special considerations. For example, if data are scarce, then simple sampling techniques may be inappropriate. More complex sampling techniques may be required. However, in data mining projects, lack of data is usually not an issue. In addition, cross-validation is often better than partitioned sample validation. When cross-validation is used, the data are partitioned several ways, and a new model is trained for each resulting data table. For example, the data are partitioned k ways, and k models are trained. For each iteration of training, a different subset of the data is left out. This left-out subset is often called a *holdout sample*. The holdout sample can be used for validation such as to compute the error criteria.

Finally, when working with small data tables, *bootstrapping* may be appropriate. Bootstrapping is the practice of repeatedly analyzing sub-samples of the data. Each sub-sample is a random sample with replacement from the full sample.

Case Study: Using Sampling in Churn Analysis

To illustrate sampling techniques, this section presents a business problem example from the telecommunications industry. Like many other businesses, telecommunications firms want to investigate why customers *churn* or switch to a competitor. More specifically, firms want to determine what is the probability that a given customer will churn. Are there certain characteristics of customers who are likely to move to a competitor? By identifying customers who are likely to churn, telecommunications firms can be better prepared to respond to the offers of competing firms. For example, the firm might want to counter with a better offer such as a lower rate or a package of service options for those who are predicted to churn and are potentially profitable.

The case study used SAS Enterprise Miner™ software running on Windows NT Server™ with Windows 95™ clients, and processed approximately 11 million records (approximately 2-gigabytes).[23] We sampled incrementally starting at 100,000 records (0.03 percent sample), and then 1 percent and 10 percent samples concluding with 100 percent or all 11 million records. We also sampled repeatedly (100 runs for each sample size with 1 run at 100 percent) to show that sampling can yield accurate results. Plotting the classification rate for the sample sizes shows the resulting accuracy. (See Figure 19, "Incremental Sample Size and the Correct Classification Rates.")

Enterprise Miner provides multiple sampling techniques and multiple ways to model churn. Figure 6 outlines the steps in the customer churn project and the nodes used to perform those steps.

Step	Action	Enterprise Miner Node
1.	Access the data	Input Data Source node
2.	Sample the data	Sampling node
3.	Partition the data	Data Partition node
4.	Develop a model	Regression node
5.	Assess the model classification results	Assessment node

Figure 6 : Churn Analysis — Steps, Actions, and Nodes

We begin by using the graphical user interface (GUI) of Enterprise Miner software. By dragging and dropping icons onto the Enterprise Miner workspace, we build a process flow diagram (PFD) that fits the requirements of our data mining project. The icons that we drag and drop on the workspace represent nodes, which open to tabbed dialogs where a variety of statistical tasks can be performed. Using the mouse, we connect the nodes in the workspace in the order in which they are needed in the process flow. Information flows from node to node along this path. In addition to running the PFD as a part of the churn analysis, we can save the diagram for later use, and export it to share with others who have interest in the project.

Figure 7 shows the process flow diagram for the customer churn analysis project.

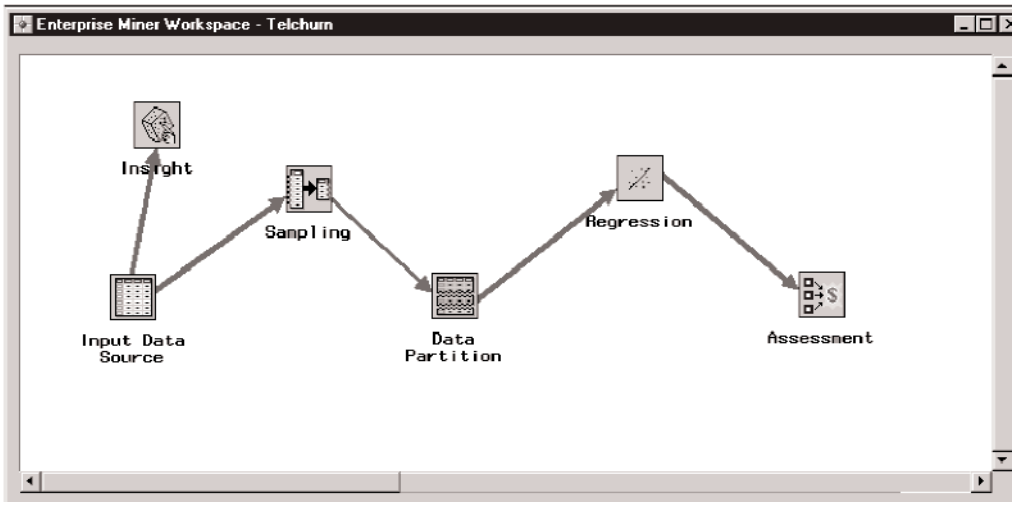


Figure 7 : Process Flow Diagram for the Customer Churn Project

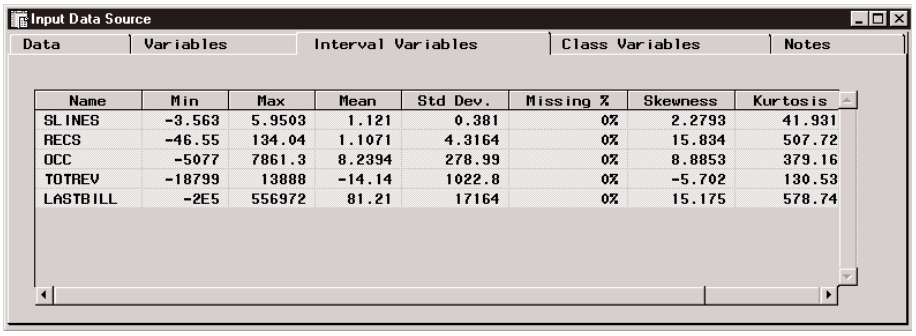
[23]The Windows NT Server™ 4.0 with Service Pack 3 was configured as follows: 4-way Pentium Pro™, 200MHz, 512k L2 Cache, 4GB Physical RAM, Ultra SCSI controllers, internal and external RAID 5 arrays, 100MBps Ethernet™.

Step 1: Access the Data

Enterprise Miner is built around the SEMMA methodology. The first step in this methodology is to access the data base in order to sample it. To access the data, we use the Input Data Source Node, which reads data sources and defines their attributes for subsequent processing. (See Figures 8 and 9.) *Metadata* (information about the data) is automatically created for each variable when we import our churn data table with the Input Data Source node. Initial values are set for the measurement level and the model role for each variable. The metadata derives from a *meta-sample*.[24] The metadata describe the data, and let us do some initial exploration of our churn data table by variable type.

Interval Variables

Each row in the table represents an interval variable. The variables are listed in the column labeled **Name**, and the remaining columns show (for each variable) the minimum, maximum, mean, standard deviation, number of missing values, skewness coefficient, and kurtosis coefficient.

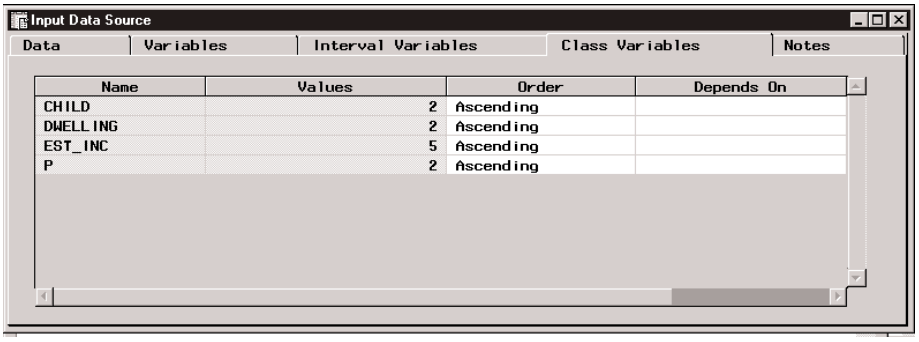


Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
SLINES	-3.563	5.9503	1.121	0.381	0%	2.2793	41.931
RECS	-46.55	134.04	1.1071	4.3164	0%	15.834	507.72
OCC	-5077	7861.3	8.2394	278.99	0%	8.8853	379.16
TOTREV	-18799	13888	-14.14	1022.8	0%	-5.702	130.53
LASTBILL	-2E5	556972	81.21	17164	0%	15.175	578.74

Figure 8 : Input Data - Interval Variable

Class Variables

For class variables, we see the number of levels, order, and hierarchical dependencies that may be in the data as shown in Figure 9.



Name	Values	Order	Depends On
CHILD	2	Ascending	
DWELLING	2	Ascending	
EST_INC	5	Ascending	
P	2	Ascending	

Figure 9 : Input Data - Class Variables

Each row represents a class variable. The column **Values** indicates the number of levels for each class variable. The column **Order** indicates how the levels are ordered. For example, the variable CHILD has two levels, and has the levels in ascending order.

[24] The meta-sample can be set to any size and is used *only* to get information about the data. All calculations are done on either the entire input data source or the training data table — not the meta-sample. The automatically generated metadata can be changed if you choose to override the default best guess.

The variable **P** represents whether a customer has churned (switched firms). **P** has two levels: 0 indicating churn and 1 indicating no churn. (Enterprise Miner software enables you to model either the event or the non-event regardless of whether the event is indicated by 0, 1, or some other means.)

Figure 10 shows the percentage for each level of the variable **P**. In this case, there is a low occurrence of churn (represented by the 0 column). Therefore, we should stratify any subsets of data we create to ensure that sufficient records are represented. By stratifying, we maintain in the sample the same percentage of records in each level that are present in the entire database. Otherwise, we might obtain a sample with very few or even no customers, who had switched firms.

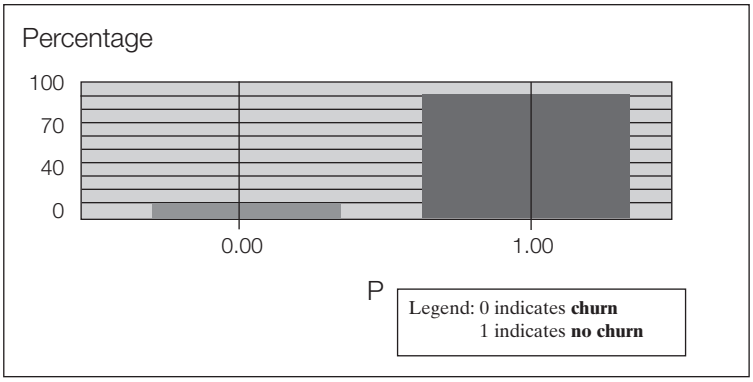


Figure 10 : Percentage of Churn and No Churn

For enriched sampling, prior probabilities can be used to weight the data records for proper assessment of the prediction results. One weighting scheme is to use the proportion in each level to assign weights to the records in the assessment data table. For example, given a variable with two levels — event and non-event — having the prior probabilities of .04 and .96, respectively, and the proportion of the levels in the assessment data table are .45 and .55, respectively; then the sampling weight for the event level equals .04/.45 and the sampling weight for the non-event level equals .96/.55.

Step 2: Sample the Data

After accessing the data and performing some initial exploration, we can take a sample. The Sampling node enables us to easily extract a sample from the input data source. For purposes of this case study, samples of different sizes were taken to illustrate that characteristics identified in the entire database also are present in the sample. In fact, sample sizes ranging from overly large to a small percentage of the data table preserve the characteristics of the entire database. There is no significant loss of information by using small, but representative samples.

Figure 11 shows the General Dialog page of the Sampling node.

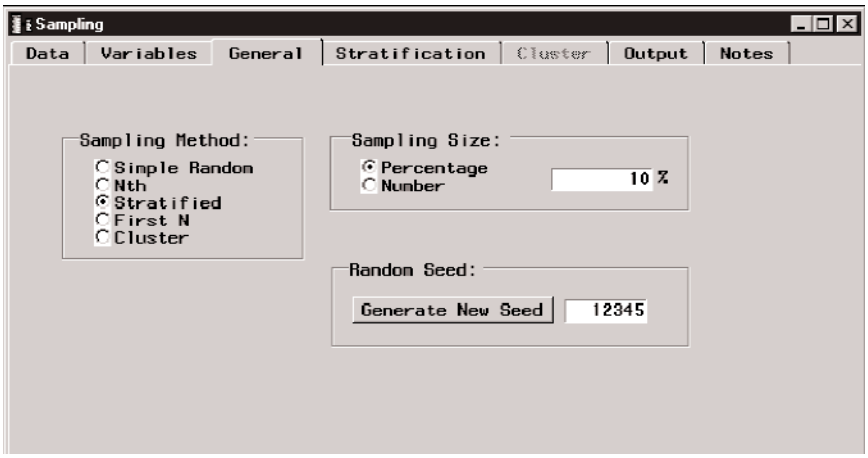


Figure 11 : General Dialog Page

Sampling Methods

As Figure 11 shows, the Sampling node supports the following types of sampling:

Simple Random

By default, every record in the data table has the same probability of being selected for the sample.

Sampling Every *N*th Record

Sampling every *n*th record is also known as *systematic sampling*.[25] This setting computes the percentage of the population that is required for the sample, or uses the percentage specified in the General tab. It divides 100 percent by this percentage to come up with a number. This setting selects all records that are multiples of this number.

Stratified Sampling

In stratified sampling, one or more categorical variables are specified from the input data table to form strata (or subsets) of the total population. Within each stratum, all records have an equal probability of being selected for the sample. Across all strata, however, the records in the input data table generally do not have equal probabilities of being selected for the sample. We perform stratified sampling to preserve the strata proportions of the population within the sample. This may improve the classification precision of fitted models, which is why we have chosen this method for our case study.

Sampling the First *N* Records

This type of sampling selects the first *n* records from the input data table for the sample.[26] We can specify either a percentage or an absolute number of records to sample in the General tab.

Cluster Sampling

This method builds the sample from a cluster of records that are similar in some way. For example, we want to get all the records of each customer from a random sample of customers.

---

[25]Sampling every *n*th record can produce a sample that is not representative of the population, particularly if the input data table is not in random order. If there is a structure to the input data table, then the *n*th-record sample may reflect only a part of that structure.

[26]Sampling the first *n* records can produce a sample that is not representative of the population, particularly if the input data table is not in random order.

Sample Size

We can specify sample size as a percentage of the total population, or as an absolute number of records to be sampled. The default percentage is 10 percent and the default number of records is 10 percent of the total records. The actual sample size is an approximate percentage or number. For example, a 10 percent random sample of 100 records may contain 9, 10, or 11 records. For the case study, we chose to specify both percentages and exact numbers. For the smaller samples, 100 thousand records were used (less than 1 percent of the data table); for larger samples, 1 percent and 10 percent were used.

Random Seed

The Sampling node displays the seed value used in the random number function for each sample. The default seed value is set to 12345. In our case study, we change the seed repeatedly to look at hundreds of different samples of different sizes. The Sampling node saves the seed value used for each sample so those samples can be replicated exactly. (If we set the seed to 0, then the computer clock at run time is used to initialize the seed stream. Each time we run the node, a new sample will be created.)

Stratification

Stratified random sampling is appropriate in this example because the levels of the categorical data could be easily under- or over-represented if simple random sampling was used. We need to stratify on the variable P, the variable that indicates whether or not the customer has churned.

In general, to perform stratified random sampling, we first select the variable(s) on which we want to stratify. Next we choose the option settings on how the stratification is to be performed to achieve a representative sample.

Variables

The Variables sub-page (Figure 12) contains a data table that lists the variables that are appropriate for use as stratification variables. Stratification variables must be categorical (binary, ordinal, or nominal); ours is binary — churn (**P**) is either 0 or 1.

Name	Status	Model Role	Measurement	Type	Format	Label
CHILD	don't use	input	binary	char	$1.	
DWELLING	don't use	input	binary	char	$1.	DWELLING*TYPE
EST_INC	don't use	input	nominal	char	$1.	ESTIMATED*INCOME
P	use	target	binary	num	BEST12.	

Figure 12 : Stratification Variables Dialog Page

Options

For our example, we are using the default settings for options. However, the Options sub-page (Figure 13) allows us to specify various details about the stratification.

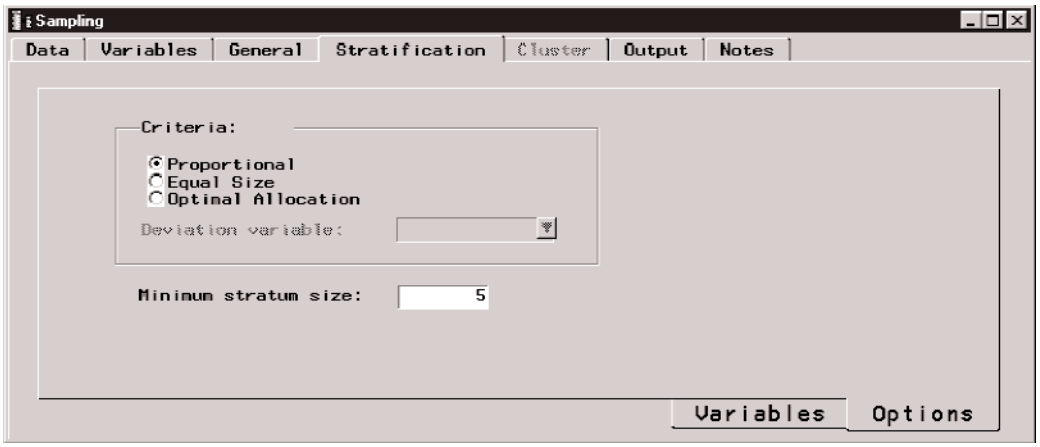
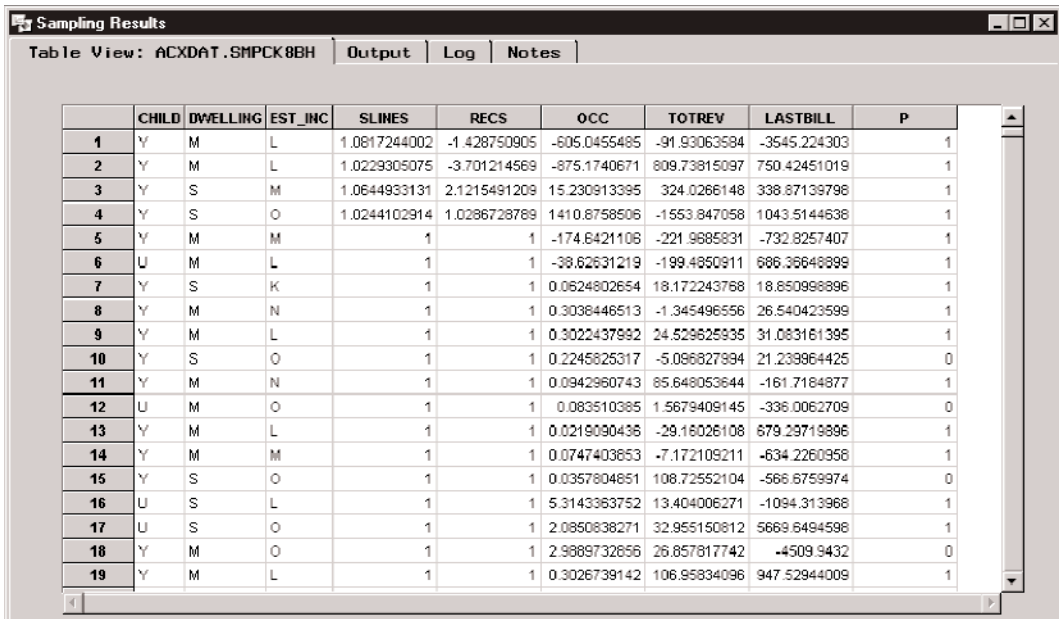


Figure 13 : Stratification Criteria Dialog Page

The Sampling node supports three criteria for stratified sampling:

- proportional sampling
- equal size[27]
- optimal allocation[28].

All stratified sampling methods in the Sampling node use computations that may require rounding the number of records within strata. To take a stratified sample on churn, we use proportional stratified sampling, whereby the proportion of records in each stratum is the same in the sample as it is in the population. We can review the results of the sample in Figure 14, the Sampling Results Browser:



	CHILD	DWELLING	EST_INC	SLINES	RECS	OCC	TOTREV	LASTBILL	P
1	Y	M	L	1.0817244002	-1.428750905	-605.0455485	-91.93063584	-3545.224303	1
2	Y	M	L	1.0229305075	-3.701214569	-875.1740671	809.73815097	750.42451019	1
3	Y	S	M	1.0644933131	2.1215491209	15.230913395	324.0266148	338.87139798	1
4	Y	S	O	1.0244102914	1.0286728789	1410.8758506	-1553.847058	1043.5144638	1
5	Y	M	M	1	1	-174.6421106	-221.9685831	-732.8257407	1
6	U	M	L	1	1	-38.62631219	-199.4850911	686.36648899	1
7	Y	S	K	1	1	0.0624802654	18.172243768	18.850998896	1
8	Y	M	N	1	1	0.3038446513	-1.345496556	26.540423599	1
9	Y	M	L	1	1	0.3022437992	24.529625935	31.083161395	1
10	Y	S	O	1	1	0.2245825317	-5.096827894	21.238964425	0
11	Y	M	N	1	1	0.0942960743	85.648053644	-161.7184877	1
12	U	M	O	1	1	0.083510385	1.5679409145	-336.0062709	0
13	Y	M	L	1	1	0.0219090436	-29.16026108	679.29719896	1
14	Y	M	M	1	1	0.0747403653	-7.172109211	-634.2260956	1
15	Y	S	O	1	1	0.0357804851	108.72552104	-566.6759974	0
16	U	S	L	1	1	5.3143363752	13.404006271	-1094.313968	1
17	U	S	O	1	1	2.0850838271	32.955150812	5669.6494598	1
18	Y	M	O	1	1	2.9889732656	26.857817742	-4509.9432	0
19	Y	M	L	1	1	0.3026739142	106.95834096	947.52944009	1

Figure 14 : Sampling Results Browser

[27]The equal size criterion specifies that the Sampling node sample the same number of records from each stratum.

[28]With optimal allocation, both the proportion of records within strata and the relative standard deviation of a specified variable within strata are the same in the sample as in the population.

Step 3: Partition the Data

We use the Data Partition node (Figure 15) to partition the sample data table into three separate, mutually exclusive data tables: training, test, and validation. Note that we also use stratification to create these partitions; we stratify on the variable **P**. By partitioning the sample, we now can perform split-sample validation, which SAS Institute recommends as a "best practice" in data mining.

The training data table is used for preliminary model fitting. The validation data table is used to monitor and tune the model weights during estimation. The validation data table also is used for model assessment. The test data table is an additional holdout data table that we can use for further model assessment. In the same way that we used stratified sampling to pull the original sample, we stratify on churn again to partition the data into train, test and validation data tables to ensure sufficient cases of churn in each of them.

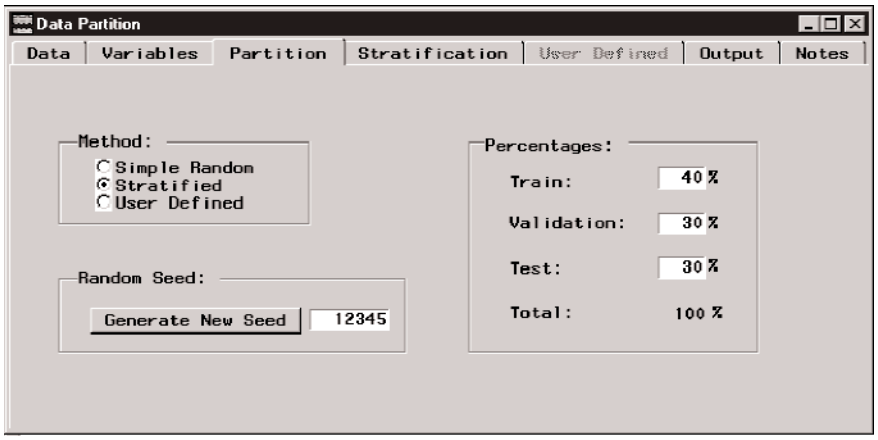


Figure 15 : Data Partition Dialog Page

Step 4: Develop a Model

Now that the data have been partitioned into training, test and validation data tables, we can model the likelihood of a customer churning. With Enterprise Miner, we could choose among a number of modeling approaches including various tree-based models as well as various neural networks. For the purposes of this example, we have chosen a logistic regression model. With the developed model, we learn which variables were important in indicating the likelihood of customers churning.

The Regression node automatically checks the data, and when a binary target variable is detected, it uses a logistic regression model. If a continuous target variable is detected, then a linear regression model is used. The Neural Network node and the Decision Tree node have similar data checking capabilities.

Figure 16 shows the Regression node Results Browser, which provides part of the estimation results from fitting the logistic regression model.

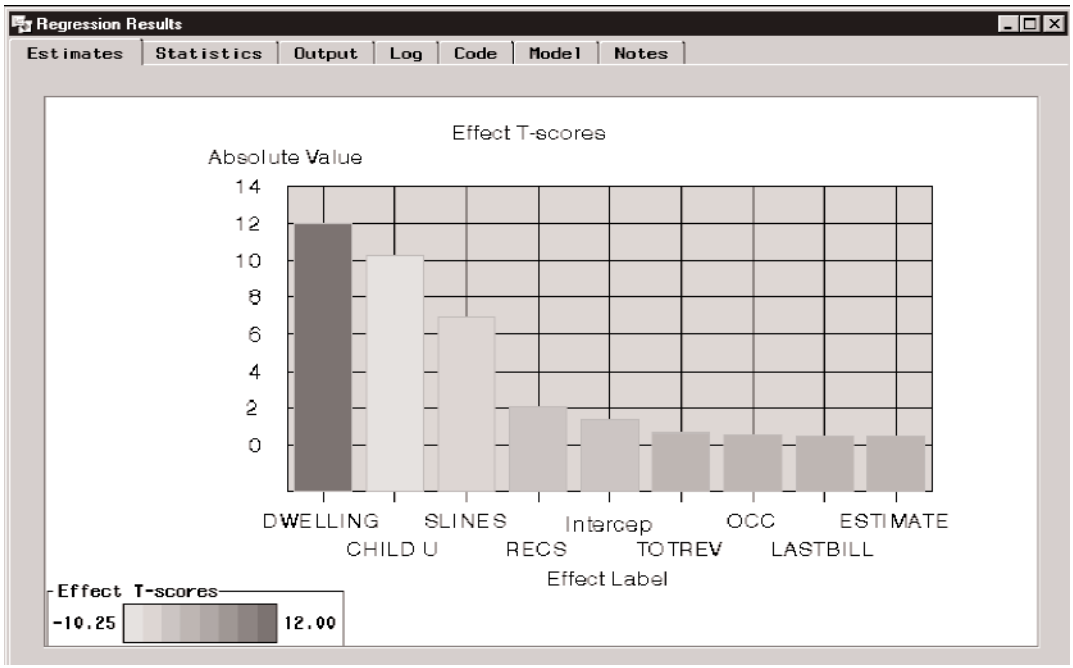


Figure 16 : Regression Results

The bar chart shows the variables that were most important in explaining the likelihood of a customer churning. The variables are sorted by their t-score, a statistic that indicates the importance of the variables in the model. For this model, the variable DWELLING was the most important variable.

The variable DWELLING has two levels, M for multi-family and S for single-family dwellings (see Figure 9). Further analysis of DWELLING indicates that customers in multi-family dwellings are more likely to churn than those in single-family dwellings.

The other variables lend themselves to similar interpretations. We can also develop decision tree models and neural networks that further identify the characteristics of customers who are likely to switch. For the purposes of this case study, we proceed using the fitted logistic regression model.

Step 5: Assess the Results

In the Assessment node, we can assess how well the logistic regression model predicted whether or not the customer would churn. The fitted model was used to predict the behavior of customers in the validation (holdout) data table and test data table.

In Figures 17 and 18, the actual values are listed on the left (1 for NO CHURN, 0 for CHURN), and the model's predictions at the bottom (1 for NO CHURN, 0 for CHURN). The vertical axis shows the number (frequency) of records in each category. The model predicted the correct classification for about 79 percent of the customers.

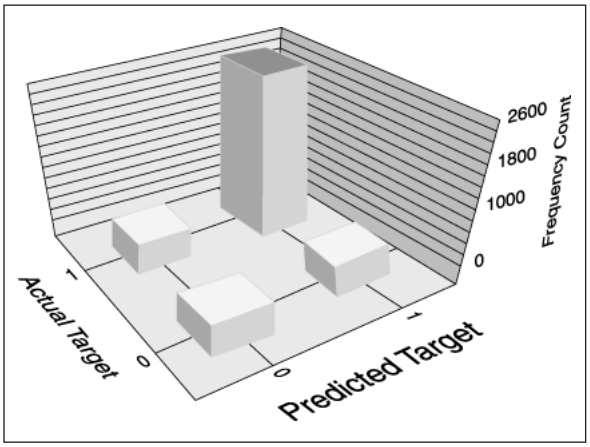


Figure 17 : Diagnostic Chart for Validation Data

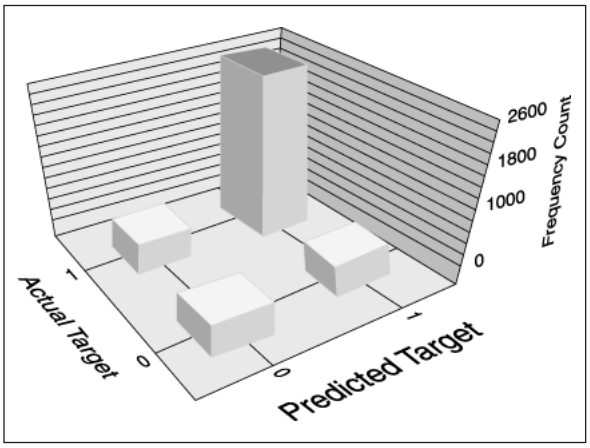


Figure 18 : Diagnostic Chart for Test Data

Next, we created samples of varying sizes. There were 100 different samples for each of the following percentages of the entire data table: 0.03 percent, 1 percent, 10 percent, and 20 percent. Note that each 0.03 percent sample has 100,000 records or 5 megabytes, and that the entire data table has 11,000,000 records or 2 gigabytes.

Each of these 400 samples was partitioned, and then separate logistic regressions were run for each of the training data tables. The logistic regression models were used to predict the behavior of customers in the respective validation and test data tables. This resulted in a distribution of correct classification rates for each group.

To assess the accuracy of the classifications, the correct classification rates for each group were averaged, and upper and lower 95 percent confidence limits were calculated. The average (or mean) value could be compared with the actual classifications, which are obtained from the entire data table. Did the samples preserve the characteristics of the entire data table? Figure 19 plots a comparison of the correct classification rates among the sample sizes.

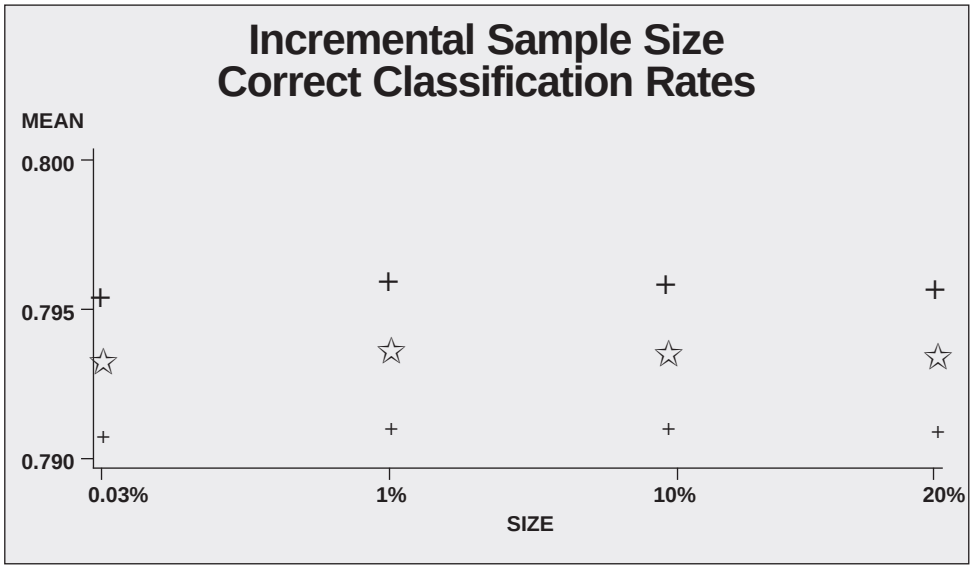


Figure 19 : Incremental Sample Size and the Correct Classification Rates

The mean value of the correct classification rates is shown on the vertical axis. The sample size is shown on the horizontal axis. The stars plot the mean value of each group, and the plus signs (+) plot the upper and lower 95 percent confidence limits. Note the consistency in the mean values and in the confidence limits. On average, the different sized-samples provide almost exactly the same accuracy in the classification rates.

In Figure 19, the plot of the correct classification rates looks impressive, but are the results really that good? What about the differences in the classification rates between the sample sizes? To answer these questions, we performed a statistical test on the difference of the means. Figure 20 compares each sample size to the other sample sizes. The first set of three rows uses the 0.03 percent sample size as the base, and then subtracts the mean of the 1 percent, 10 percent, and 20 percent samples. The second set of three rows uses the 1 percent sample as the base, and so on. The column labeled **Difference Between Means** shows that the difference in ability to correctly classify is in the fourth decimal place, or in hundredths of a percent.

Sample Size Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit
0.03%-1%	-0.0008941	-0.00039	0.00011
0.03%-10%	-0.0007792	-0.00028	0.000225
0.03%-20%	-0.0016483	-0.00016	0.001337
1% -0.03%	-0.0001102	0.00392	0.000894
1% -10%	-0.0003872	0.000115	0.000617
1% -20%	-0.0012563	0.000236	0.001729
10% -0.03%	-0.0002252	0.000277	0.000779
10% -1%	-0.0006171	-0.00011	0.000387
10% -20%	-0.0013713	0.000121	0.001614
20% -0.03%	-0.0013367	0.000156	0.001648
20% -1%	-0.0017286	-0.00024	0.001256
20% -10%	-0.0016136	-0.00012	0.001371

Figure 20 : Comparison of Sample Sizes

There are two additional columns in the table that show the 95 percent upper and lower confidence limits on the differences between the means. These values are small, and the confidence limits are "overlapping," that is, the mean for the 0.03 percent sample size falls within the confidence limits for the 20 percent sample.

If a statistically significant difference did exist, then the difference in mean values would be outside of the confidence interval (upper and lower 95 percent confidence limits). In every case, the differences are between the confidence limits. We conclude that there are no statistically significant differences between the mean values of the classification rates. Models fit using each group of samples have almost the exact same accuracy in predicting whether customers are likely to churn.

Summary of Case Study Results

The case study used samples of different sizes to train models, which were used to predict the behavior of customers. The predicted behavior was compared with the actual behavior to calculate classification rates. The classification rates had essentially the same level of accuracy for models trained using samples having 0.03 percent, 1 percent, 10 percent, and 20 percent of the data. These results confirm that properly performed sampling can maintain precision. Moreover, sampling saves processing time and enables you to validate and test the results.

SAS Institute: A Leader in Data Mining Solutions

For more than 20 years, SAS Institute Inc. has distinguished itself by providing world-class technology and the technical know-how to solve customers' business problems around the globe.

For data mining, the Institute has developed SAS Enterprise Miner™ software, which enables business leaders to make better business decisions by controlling the process that moves from data, to information, to business intelligence. However, moving from data to business intelligence requires taking logical steps that will ensure the results of your data mining efforts are reliable, cost efficient, and timely.

In data mining, *reliability* means the results are based on modern analytical practices. Thus, inferences you make are quantifiable; that is, you can say with a specific degree of confidence — for example a 95 percent confidence level — that the output of your data mining project is between the values x and y.

Cost efficiency refers to the balance between, on the one hand, the reliability of the results and, on the other hand, the human and computing resources that are expended obtaining the results.

Timeliness refers to the need to obtain reliable, cost-effective results when you need them; that is, in time to act proactively in a fast-moving, highly competitive global economy.

References

American Statistical Association, Blue Ribbon Panel on the Census, (1996 September). *Report of the President's Blue Ribbon Panel on the Census*. Retrieved September 8, 1998 from the World Wide Web: http://www.amstat.org/census.html

Breslow, N.E. and W. Day, (1980), *Statistical Methods in Cancer Research. Volume 1 - The Analysis of Case-Control Studies*, Lyon: IARC Scientific Publication No. 32.

Cochran, William G., (1977), *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Elder, John F., IV and Daryl Pregibon, (1996), Advances in Knowledge Discovery & Data Mining, "A Statistical Perspective on KDD."

Gladstone, Brooke, (1998), "Error Naming Campaign Winner," NPR Online, Morning Edition, November 3, 1998, Retrieved November 6, 1998 from the World Wide Web: http://www.npr.org/programs/morning/archives/1998/981103.me.html

Greenhouse, Linda, (1998), "High Court to Hear Appeal of Census Ruling," *The New York Times on the Web*, (September 11), Retrieved via site search on "census" September 11, 1998 from the World Wide Web: http://www.nytimes.com

Gribbon, John, (1996), *Companion to the Cosmos*, New York: Little, Brown and Company.

Hays, William L., (1981), *Statistics*, New York: CBS College Publishing.

International Data Corporation, (1998 June), "Information Access Tools: 1998 Worldwide Markets And Trends," IDC #15932, Volume: 1.

Manski, C. F. and Daniel McFadden, (1981), *Structural Analysis of discrete data with Applications*, Cambridge, Mass: MIT Press.

McCullough, David, (1992), *Truman*, New York: Simon & Schuster.

META Group, (1997 August 19), "Data Mining the SAS Data Warehouse," *Application Delivery Strategies*, File #594.

Phillips, John L., (1996), *How to Think About Statistics*, Fifth Edition, New York: W.H. Freeman and Company.

Potts, William J. E., (1997), *Data Mining Using SAS Enterprise Miner Software*. Cary, North Carolina: SAS Institute Inc.

Ripley, B.D., (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Ross, Sheldon, (1998), *A First Course in Probability*, Fifth Edition, Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Saerndal, Carl-Erik, Bengt Swensson, and Jan Wretman, (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sarle, W.S., ed. (1997), *Neural Network FAQ*, periodic posting to the Usenet newsgroup comp.ai.neural-nets, URL: ftp://ftp.sas.com/pub/neural/FAQ.html

Snedecor, George W. and William G. Cochran, (1989), *Statistical Methods*, Eighth Edition, Ames, Iowa: The Iowa State University Press.

U.S Bureau of the Census, (1992), *Statistical Abstract of the United States*, 112th edition, Washington, DC.

U.S. House of Representatives v. U.S. Department of Commerce, et al, (1998), Civil Action No. 98-0456 (Three Judge Court) (RCL, DHG, RMU), Opinion filed August 24, 1998 by Circuit Judge Douglas H. Ginsburg, and District Court Judges Royce C. Lamberth and Ricardo M. Urbina.

Westphal, Christopher and Teresa Blaxton, (1998), *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, John Wiley and Sons.

Recommended Reading

Data Mining

Berry, Michael J. A. and Gordon Linoff, (1997), Data Mining Techniques, New York: John Wiley & Sons, Inc.

SAS Institute Inc., (1997), SAS Institute White Paper, Business Intelligence Systems and Data Mining, Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1998), SAS Institute White Paper, Finding the Solution to Data Mining: A Map of the Features and Components of SAS Enterprise Miner™ Software, Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1998), SAS Institute White Paper, From Data to Business Advantage: Data Mining, The SEMMA Methodology and the SAS® System, Cary, NC: SAS Institute Inc.

Weiss, Sholom M. and Nitin Indurkhya, (1998), Predictive Data Mining: A Practical Guide, San Francisco, California: Morgan Kaufmann Publishers, Inc.

Data Warehousing

Inmon, W. H., (1993), Building the Data Warehouse, New York: John Wiley & Sons, Inc.

SAS Institute Inc., (1995), SAS Institute White Paper, Building a SAS Data Warehouse, Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1998), SAS Institute White Paper, SAS Institute's Rapid Warehousing Methodology, Cary, NC: SAS Institute Inc.

Singh, Harry, (1998), Data Warehousing Concepts, Technologies, Implementations, and Management, Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Statistics

Hays, William L. (1981), Statistics. New York: Holt, Rinehart and Winston.

Hildebrand, David K. and R. Lyman Ott, (1996), Basic Statistical Ideas for Managers, New York: Duxbury Press.

Mendenhall, William and Richard L. Scheaffer, (1973), Mathematical Statistics with Applications. North Scituate, Massachusetts.

Phillips, John L., (1996), How to Think About Statistics, Fifth Edition, New York: W.H. Freeman and Company.

Credits

Data Mining and the Case for Sampling was a collaborative work. Contributors to the development and production of this paper included the following persons:

Consultants

SAS Institute Inc.

John Brocklebank, Ph.D.
Craig DeVault
Padraic G. Neville, Ph.D.

Writers

SAS Institute Inc.

Anne H. Milley
James D. Seabolt, Ph.D.
John S. Williams

Technical Reviewers

SAS Institute Inc.

Brent L. Cohen, Ph.D.
Gerhard Held
Kristin Nauta, M.Stat.
Udo Sglavo
R. Wayne Thompson, Ph.D.

American Management Systems

Donald Rosenthal, Ph.D.

Copyeditor

SAS Institute Inc.

Rebecca Autore