

stone
analytics

Automated Analytics and Predictive Modeling

A White Paper by Stone Analytics

3665 Ruffin Road, Suite 300
San Diego, CA 92123
(858) 503-7540
www.stoneanalytics.com

Automated Analytics and Predictive Modeling A White Paper by Stone Analytics

Why Read This Paper?

According to a recent IDC study, “Business analytics implementations generated an average 5-year return on investment of 431%.” The study goes on to say organizations that have “successfully implemented and utilized analytic applications have realized returns ranging from 17% to more than 2000%, with a median ROI of 112%.”

These are some pretty extraordinary claims, and there is no doubt that individual results are sure to vary. At the same time, however, there is also no doubt that analytics has a great deal to offer the decision-making process and that it is rapidly becoming a tool employed by more and more businesses of every size and category.

Summary

Though long used by the Fortune 1000 and other large organizations, predictive analytics has been both difficult and expensive to employ. Until very recently, it required the time and effort of one or more highly-trained and highly-paid statisticians. To be reliable, it also requires large stores of data, the kind of data that in the past was kept, or purchased, by only the largest companies. With the advent of digital technology and the Internet, however, data of all kinds is becoming more and more abundant everyday. In addition, there are now statistical modules that automate the statistical process itself, making predictive modeling an easy and cost-effective addition to the decision-making process for small and medium sized companies. Decision-makers at all levels now have the ability to apply advanced analytics to a much broader range of business and organizational tasks, including site selection, fund-raising, and target marketing.

Introduction

Analytics is defined by Webster’s as “the method of logical analysis.” To a large extent, it is synonymous with statistical analysis, which has long been an essential tool in both scientific and financial settings. More recently, analytics has also become an essential tool for solving complex logistical problems, such as those associated with manufacturing and delivery schedules. In all of these areas, it is the ability to identify and apply the relationships between, available information and a question of interest whose answer is uncertain that gives analytics its unique value.

In spite of what it has to offer, however, analytics has never been cheap or easy to employ. Until recently, it required one or more highly-trained and highly-paid statisticians. To justify the cost, there had to be vast sums of money involved. At the same time, even for those willing to pay the price, data—an essential ingredient of any statistical analysis—was not always easy to come by. Few organizations possessed the kind of data needed to perform an accurate analysis. That has changed. Companies of all sizes and description now gather and maintain data about their customers and clients. From gasoline to diapers, practically every purchase made in the United States is recorded in some electronic manner. Data of all kinds are now readily available—data about income, homeownership, education, and a thousand other demographic categories. As a result, analytics is finding new applications, particularly in the area of predictive modeling,

which can be used in tasks ranging from site selection and fraud detection to fund raising and target marketing.

What is Predictive Modeling?

Think about the processes you use every day to make judgments and decisions—given past experiences and a variety of observations about the question at hand, you weigh the facts and make a determination. You are, in effect, modeling. Statistical modeling is simply the standardization of this process. Predictive models, in particular, seek to explain the relationships between a critical variable and a group of factors that help predict its outcome.

One obvious example of an area in which this is true is target marketing. Target marketing refers to the practice of ranking potential customers based on the analysis of available customer information. By narrowing the marketing focus to the most promising customer leads, target marketing substantially reduces marketing costs associated with selling products and services. From telephone companies and credit card issuers battling “churn,” to supermarkets and cable companies looking to sell additional products or services to their existing customers, target marketing is no longer confined to book and record clubs. Companies from AT&T and Sprint to Citibank and Prudential have substantial target marketing programs in place.

Another example where an important business decision has to be made on the basis of limited information is site selection. Each year, chain stores and franchises have to consider and choose from among hundreds of potential locations. It is simply not possible to visit and report on each one separately. A similar situation exists for retailers regarding staffing decisions during the holiday season. To meet manpower requirements, department stores and other retailers hire tens, even hundreds, of employees—so many that it would be impossible for HR personnel to go through every application in detail.

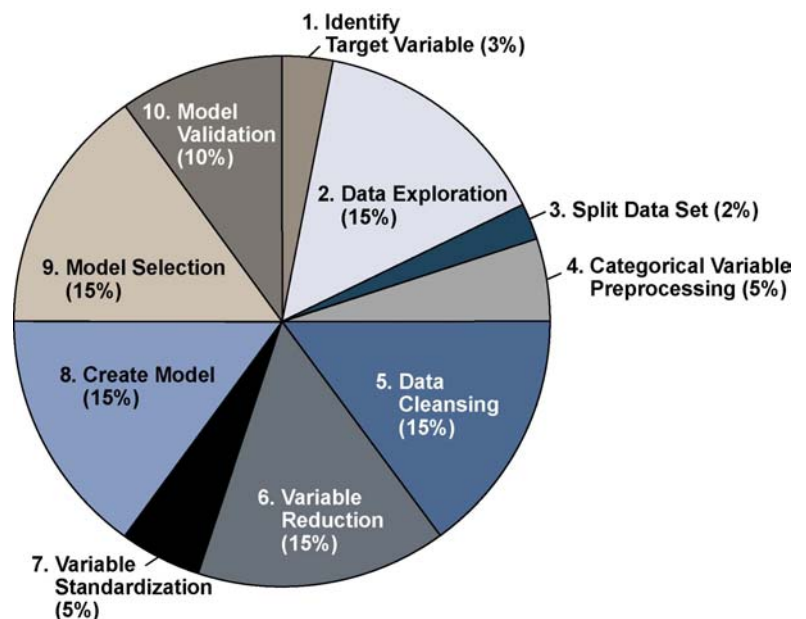
Still, while it may be limited, in each of these examples some information, or data, does exist. In terms of target marketing, it might be the Recency, Frequency, and Monetary Value of a customer’s purchase history, otherwise known as RFM. For site selection it might be the demographics of the surrounding neighborhoods or proximity to public transit, while for personnel hiring it might be the education level or previous employment of the applicant. The information may be limited or incomplete, and in an ideal world, one would hold off on any decisions until a more complete body of data could be assembled and examined. In the real world, however, that luxury rarely exists. Regardless of the data, at a certain point, the decision has to be made—the campaign has to be mailed, the site chosen, or the person hired. In each of these situations, the decision-maker must leverage the data at hand to the best advantage, and when the stakes are high enough, companies have long been willing to pay for a team of analysts in order to get this statistical advantage.

The Modeling Process

As mentioned above, in addition to data constraints, in the past there have been substantial cost constraints as well. Depending on the scope, modeling projects can cost anywhere from \$25,000 to \$100,000, and take weeks or even months to complete. Below are just some of the steps involved:

- Identify Target Variable: The analyst must select or, in many cases create, the target variable, which is the question that is being addressed.
- Data Exploration: The analyst examines the data, computing and analyzing various summary statistics regarding the different variables contained in the data set.
- Split Data Set: The analyst may randomly split the data into two sets, one of which will be used to build, or train, the model, and the other of which will be used to test the quality of the model once it is built.

The Modeling Process:
Percentage of time spent on each step



- Categorical Variable Preprocessing: Categorical variables are variables such as gender and marital status that possess no natural numerical order. These variables must be identified and handled separately from continuous numerical variables such as age and income.
- Data Cleansing: The data must be cleansed of missing values and outliers. Missing values are, quite literally, missing data. For example, there may be a number of records in a database for which the age is a null value. Handling missing data may be tricky, since the information may be unavailable for a number of different reasons. For example, if the data were collected via a survey, age might be missing because the respondent elected not to reveal his or her age. On the other hand, the response may have been originally present but was lost during processing. Outliers, on the other hand, are data records so different from the rest that they can skew the results of calculations. For example, there might be a household income entry of \$150,000 among a group where the rest of the records range

from \$30,000 to \$50,000. As with missing values, the analyst must decide whether to include, exclude, or replace them with a more typical value for that variable.

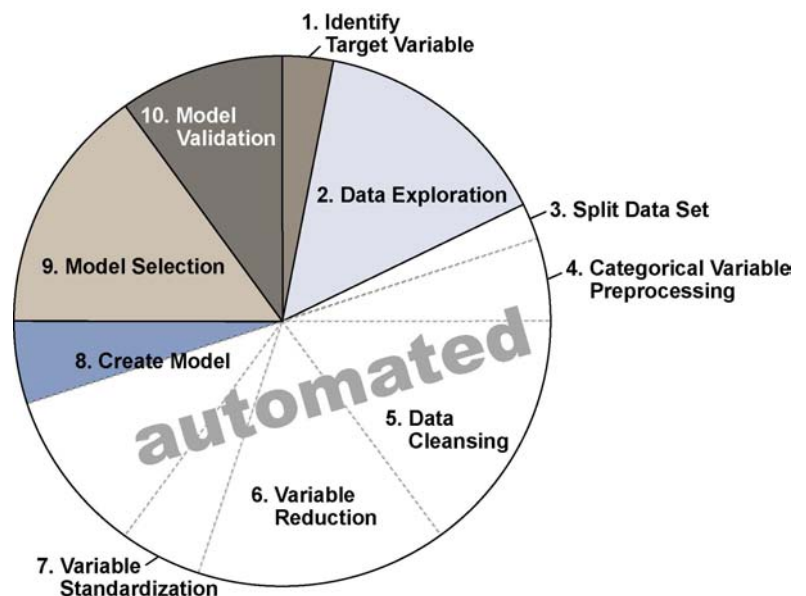
- Variable Reduction: The variables in the data set are examined in terms of their correlation with the target variable. Those that are redundant or highly correlated with each other are removed. For example, a data set may include monthly and yearly household income. In some situations, including both could be problematic.
- Variable Standardization: The remaining variables are often re-scaled so that they can be more efficiently analyzed.
- Create Model: Capture variables and weights that yield the best estimate of the target variable using the training data taken from the original data set.
- Model Selection: Several competing models are often considered.
- Model Validation: Run the model using the test data taken from the original data set.

Automated Analytics

In the past, the above steps required the expertise not only of one or more analysts, but programmers as well. It is this exact process that Stone Analytics has automated in its *Decision Science™* product line. With *Decision Science*, decision-makers of all levels can use advanced analytics to guide their critical business decisions at a fraction of the cost and time. *Decision Science* analytic engines are easy-to-use software tools that enable business professionals to build and implement powerful predictive models directly from their desktops, making it possible to apply analytics to a much broader range of business and organizational tasks.

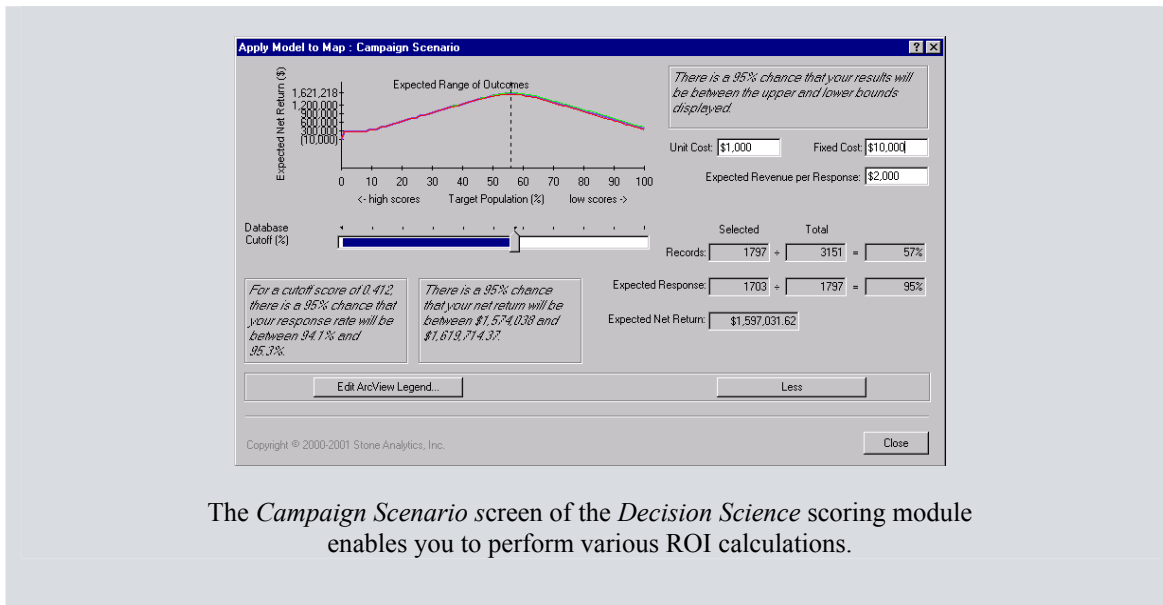
The Modeling Process with *Decision Science*

By automating steps 3 through 7, as well as elements of steps 2 and 8, *Automatic Analyst* can cut modeling time by upwards of 50%.



Decision Science enables users to make predictions of two types; (1) predictions about quantities, such as number of units sold annually, and (2) predictions about the outcome of Yes or No questions, such as whether or not a potential customer is likely to purchase your product. To do

this, it takes data with a known outcome and constructs a statistical model that relates the way in which individuals previously responded with behavioral and demographic characteristics. For example, using the results from a previous marketing promotion, *Decision Science* produces a model that gauges response to the promotion with an individual's unique characteristics. The model could then be applied to future marketing campaigns to estimate the probability of each individual in a prospect list responding in a positive manner. *Decision Science* automatically ranks, or scores, the entire list, making it easy to quickly identify the top 10 or 20 percent of likely customers, or, just as easily, exclude the bottom third.

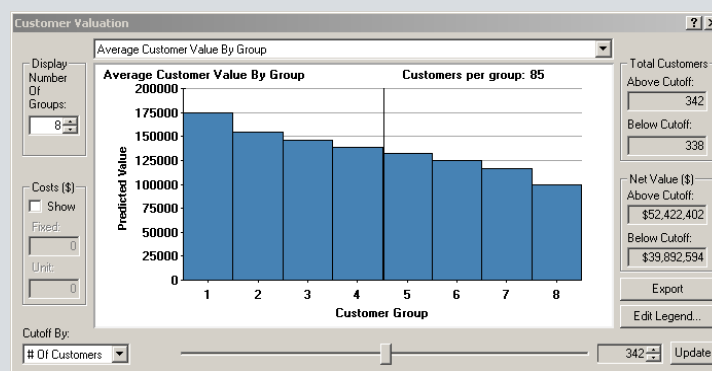


Decision Science's proprietary *Automatic Analyst*[™] technology automates much of the modeling process, enabling users with little or no statistical experience to perform statistical analysis quickly and easily. Data exploration and preprocessing, which typically takes 50 to 80% of an analyst's time during the modeling process, is handled automatically by the software. At the touch of a button, *Automatic Analyst* scans an entire data set and

- Automatically distinguishes between continuous numeric variables such as income and age, and categorical variables such as gender and marital status.
- Automatically handles problem data, such as missing values and outliers.
- Automatically partitions the data into random test and train subsets, to protect against sample bias in the data.
- Automatically examines the relationship between each potential variable and the question you are trying to answer to find the most predictive variables.

- Automatically uses these variables to build a statistical model that optimizes the decision you are about to make.
- Automatically evaluates the accuracy of the models it creates.

Problem solving is one of the core functions of any successful business. The problem might be handling the return of a purchase or insuring repayment of a loan without default. Automated analytic modules such as *Decision Science* enable business managers to identify the critical behavioral patterns within their customer and client databases, so that instead of subjectivity and conjecture, they can now base their decisions on proven analytical methods.



The *Customer Valuation* screen of the *Decision Science* valuation module enables you to sort prospective customers, store locations, or anything else by predicted value.

About Stone Analytics

Stone Analytics develops, markets, and supports analytical products that add value to any data-rich application environment. Stone Analytics' statistical modeling products are easy for software developers to integrate, and easy for customers to use. These tools bring point-and-click simplicity to the tasks of rank ordering, grouping, and similar everyday business decision-making tasks. Stone Analytics employs the latest statistical and information management techniques to provide customers with state-of-the-art solutions that add statistical modeling functionality to their applications.

Stone Analytics supplements its products with a wide range of professional services customized according to customer needs. These include integration assistance, custom model building, and consulting in the areas of data management and analytics.