



Powerhouse and the Data Analysis Process

PTI's (Powerhouse Technologies Inc) Powerhouse™ software is a radically new yet customer-proven approach to high-performance business analytics software. It finds persistent, real, often small effects that may be statistically insignificant but can have significant commercial value. The Powerhouse approach and technology deliver two main business benefits:

1. Eliminates costly intelligence bottlenecks by standardizing and automating key analytics processes. These bottlenecks are processes that are manual, repetitive, require scarce, highly skilled personnel to execute, and add little to no value to subsequent projects or analyses.
 - These bottlenecks include data preparation, variable selection, data analysis, and data diagnostics
 - IDC states that 70% of the data analysis process is consumed by data preparation alone.
2. Rapidly delivers the most quantifiably accurate and reliable results possible from the data provided. Using an innovative information mapping approach, Powerhouse eliminates the need for “learning algorithms” and the associated guesswork that is entailed in using them. The resulting productivity benefits translate into delivery of actionable findings and results to business managers within minutes or hours, not the traditional days or weeks.

To understand the significance of the capabilities and benefits that Powerhouse offers, it is imperative to review the data analysis process. Whenever data is to be analyzed using modeling techniques, there are several stages that an analyst has to complete. Those areas are:

- | | |
|--|---|
| 1. Data assembly (partial applicability) | 4. Data survey |
| 2. Data preparation | 5. Variable selection (Feature selection) |
| 3. Data assay | 6. Model building and explanations |

Data Assembly

The analyst acquires the datasets to be analyzed. To build models, these datasets must be assembled into row and column format (one row represents a single occurrence of the events of interest, and each column represents the state or value of some feature of interest for the event of interest). This often involves combining a number of datasets from a variety of sources (both internal and external) and in different formats into one dataset for modeling. Also, the analyst typically designs additional new variables that represent the business objects of interest.

Current Practice: Assembling the source datasets into one dataset for modeling requires the use of data handling tools and techniques such as FoxPro, Excel, Visual Basic, “C”, SAS, SQL, Java, or Perl. Testing newly-created variables for their potential performance is currently done iteratively by building models using these variables. This process is largely dependent on the skill of the analyst.

-more-

Powerhouse Solution: Like all modeling tools, Powerhouse requires a fully denormalized dataset. The area in data assembly in which Powerhouse can be used is after the analyst has created new variables. Powerhouse can quickly test the fitness of the newly-created variables without any modeling (see Variable Selection section in this document) before proceeding through the rest of the process. Assembling datasets is done with data handling tools outside of Powerhouse.

Data Preparation (also known as data cleansing)

An analyst prepares, or cleans the data in order to solve problems caused by missing values, nulls, erratic distributions, outliers, low prevalence of the value to be predicted and many other problems. Data needs to be prepared before modeling, usually to overcome limitations of learning algorithms, in order to get the best results possible. Without skilled and careful preparation, even the possibilities revealed by the Data Survey (see Data Survey section in this document) may be impossible to achieve.

Current Practice: After the source datasets are assembled into one dataset for modeling, data preparation can take up to 60-90% of total modeling project time. It usually requires application of many computationally intensive algorithms by a skilled analyst in order to achieve good results. Also, different algorithms usually require different preparation methods.

Powerhouse Solution: This entire process is performed “under the covers”. Powerhouse performs any necessary problem handling *automatically* as the data is loaded. Powerhouse either automatically handles or is completely impervious to missing values, nulls, erratic distributions, outliers, low prevalence of the value to be predicted, and many other problems which beset many other data analytic algorithms and tools.

Data Assay

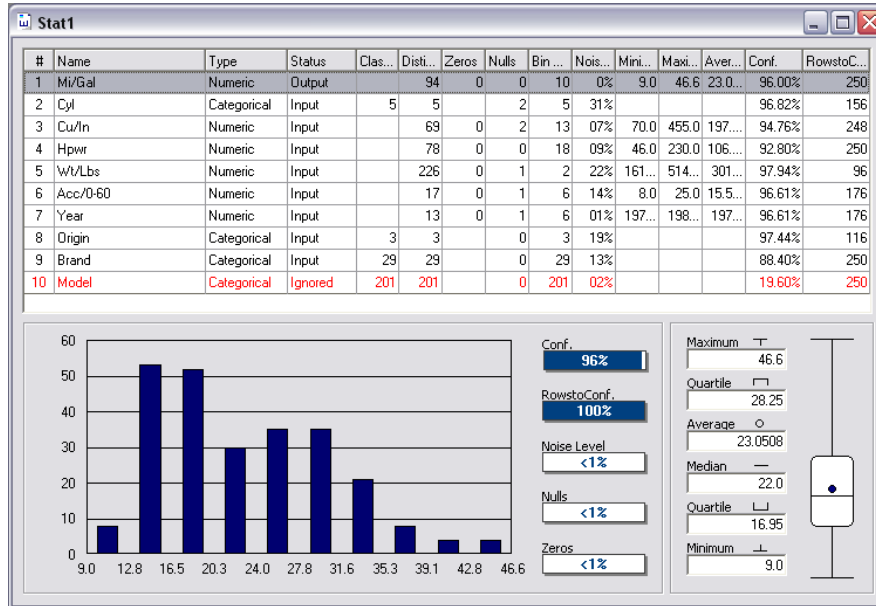
An analyst reviews and tests the individual variables in an assembled dataset to determine if there are any problems with them.

Current Practice: This diagnostic activity is accomplished typically by an analyst iteratively building models.

Powerhouse Solution: Once the assembled dataset is loaded, the analyst is provided with a comprehensive array of useful statistics about a dataset’s individual variables in one comprehensive window (**the Stat window, see Figure 1**), explicitly including their relevance for any model the analyst may wish to build. Powerhouse generates these statistics automatically while loading the data. No modeling is required to produce these statistics.

-more-

Figure 1



Assaying Data (no modeling required)

Powerhouse shows everything you need to know about your variables in one window.

You get an initial “sanity check” of each variable. Are they what you expect? Do they carry reasonable values? Are the values erroneous? Are the variables valid to use for further analysis?

Functions shown in **Figure 1** include:

- The amount of noise (garbage) in each variable
- The confidence level of each variable
- Intuitive visual representations of important measures.

Data Survey

An analyst examines the assembled dataset as a whole to determine if it can be used to build a model on that dataset, and how applicable the model would be to another, similarly structured dataset.

Current Practice: Typically, an analyst accomplishes this diagnostic activity by iteratively building models in an attempt to improve performance. Only when analysts feel that they cannot get any better results do they “declare victory” and deploy the model.

- There is no measure or statistic that indicates if the best possible model has actually been discovered.
- There is no definition of what “best” means.
- Any determination of “best” usually depends entirely on the analyst believing that a better model is unlikely to be discovered given the time and resources available.

Powerhouse Solution: The analyst gets a comprehensive array of useful statistics about the dataset in one comprehensive window (**the Entropy Map, see Figure 2**). Before investing time and resources in building models, an analyst can use this map to very quickly:

- Determine whether it is worth going forward with the analysis, or
- Whether the dataset available, even in principle, cannot provide sufficient insight.

This is where the analyst determines if more or better data is required. Powerhouse also provides a measure that can be used to define the best possible model available from the data. No modeling is required to produce these statistics. Powerhouse generates them automatically while loading the data.

-more-

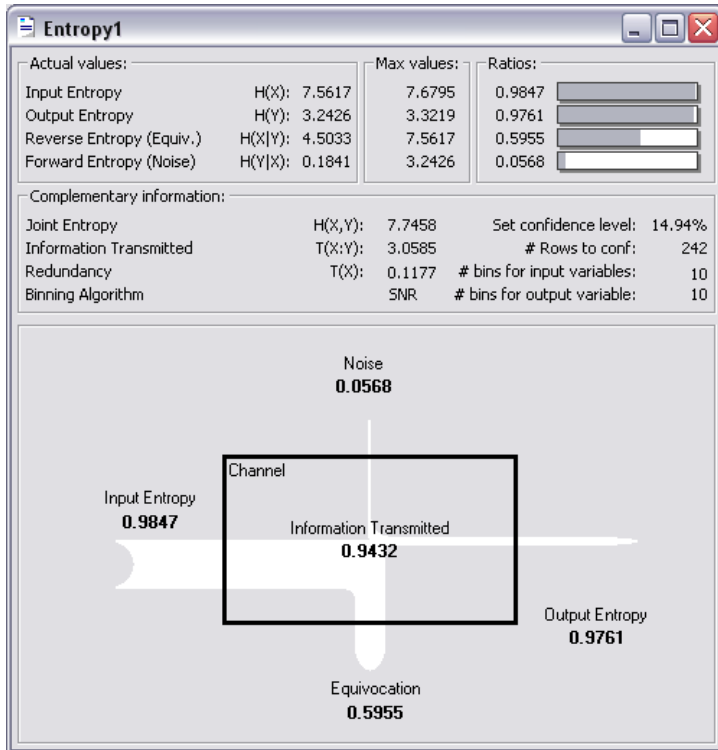


Figure 2

Evaluate Your Dataset (no modeling required)

Information Theory principles form the basis of Powerhouse. Two tenets of Information Theory, information and entropy, are at play in Figure 2.

Information can be measured as a quantity, and quantities of information can be compared and described. Any time there is uncertainty in the quality of the information, or its relevance to a model being developed, it is regarded as entropy.

Figure 2 provides a view of the general shape of a dataset, such as whether it contains enough information to build a model.

- Direct measurement of data quality (confidence level)
- Maximum possible amount of information for any model.

If the results can be used to build models for the outcome of interest, then no further discovery of the dataset's usefulness is needed.

Variable Selection (or Feature Selection)

Given an assembled and prepared data set with many variables, an analyst usually attempts to reduce the variable count by "feature selection" or "variable selection" to select some small set of variables that will produce a satisfactory model. The need for variable selection is: 1) too many variables included in a model make the modeling results hard or impossible to explain, and 2) noisy, irrelevant or "garbage"-filled variables included in a model can damage model performance and the results the model produces.

Current Practice: The analyst iteratively creates multiple models with various selections of variables that intuition suggests may be useful or worthwhile. Some tools automate part of this process in that the tools iteratively make the selections and create multiple models automatically. These techniques lead to a suggested set of variables which may (or may not) be used to create the most generally applicable model. Such techniques are:

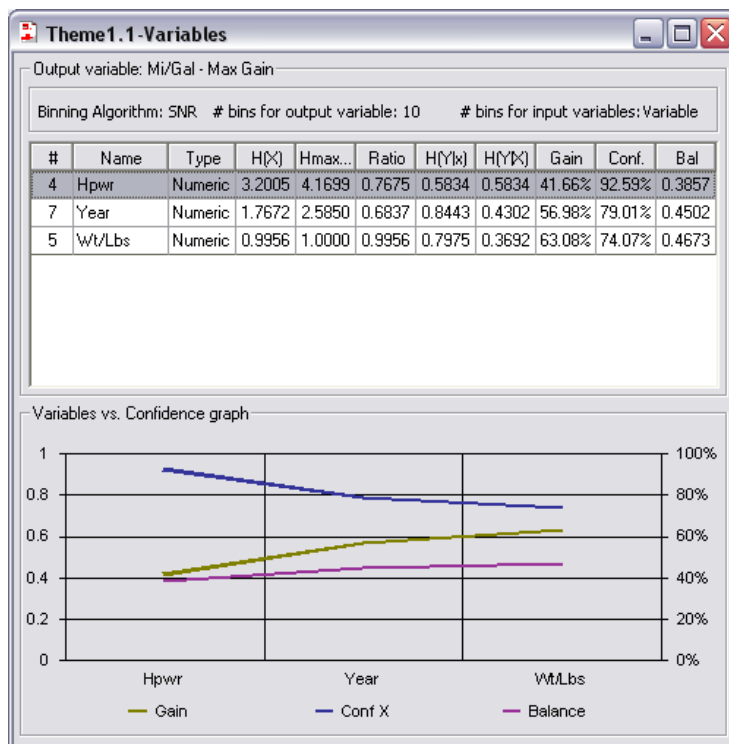
- Highly sensitive to the modeling algorithm chosen
- Highly sensitive to the way the data is prepared
- Highly dependent on analyst skill
- There is no metric by which to objectively judge the "best" selection of variables
- There is no quantitative definition of "best".

-more-

Powerhouse Solution: A “one-button” optimal variable selection tool provides a numerical measure of the quality of the variables selected – and guarantees that the first selection will always have the highest balance between information content and representativeness. The tool examines all the variables available in a few seconds or minutes (at most) and selects variables such that:

- They carry the most noise-free (low garbage) information about the output (outcomes to be predicted), and
- They carry the most generally representative information possible (so it works as consistently as possible in all datasets).

Powerhouse will measure any selection of variables for comparison, including a user-generated selection.



Variable (Feature) Selection (no modeling required)

Powerhouse automatically selects the variables that transmit the most amount of noise-free information at the touch of a button.

The variable selection made by Powerhouse, called a “Theme,” forms the best explanation of the data. If the dataset allows, it’s possible to select multiple Themes

In **Figure 3**, Powerhouse displays:

- Variables added in order of information relevance
- Information and garbage contributed by each variable
- Graphical illustration of key variable quality metrics.

Figure 3

Model Building and Explanations

Predictive Modeling Techniques

When the long process of data assembly, data preparation and variable (or feature) selection is completed as well as skill, experience and time permit, the analyst builds a model that produces the required prediction, score, probability, clustering – or the output form that the project requires.

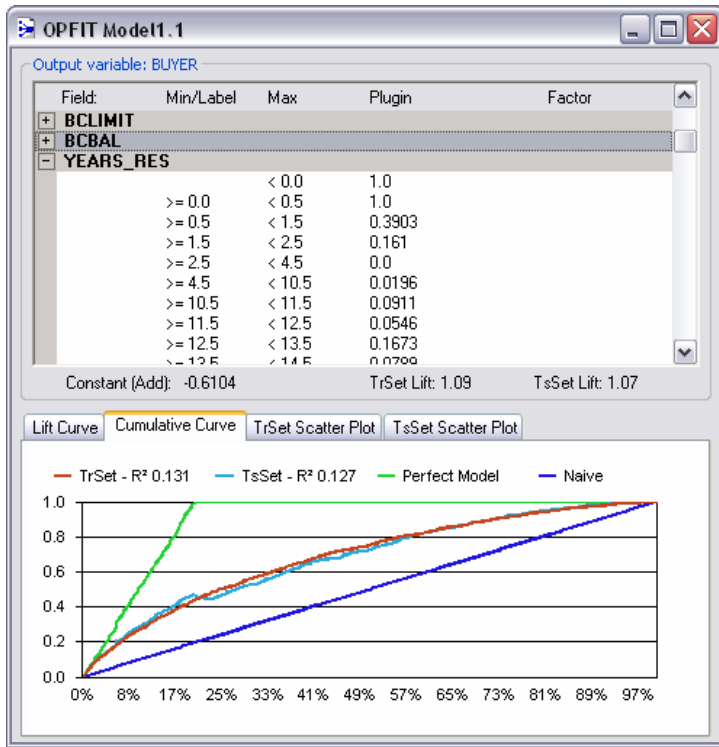
Current Practice: The analyst will select one or more algorithms with which to attempt the final model. (Each algorithm may require a differently prepared dataset.) Each algorithm has more or less parameters that have to be adjusted using analyst experience and intuition and which can have enormous effect on the quality of the final model.

The analyst will attempt many models with different parameter settings (or the tool will iteratively try multiple models with a variety of settings) in an attempt to discover an algorithm and parameter configuration that produces an acceptable model. Ultimately the analyst will select a model that intuition suggests is “best”.

Powerhouse Solution: The Powerhouse modeling tool has no parameters to adjust. In fact, it uses no learning algorithm and so requires no training cycles. Requesting a model by pressing the “Create Model” button prompts Powerhouse to read the internal information map and write out the relationships represented directly in a form simple to interpret and apply. The analyst can choose one of two representations:

- The OPFIT™ model (example results, see Figure 4) that uses a linear transfer function to map input values to predicted values, and
- The MAXIT™ model that uses a set of rules to map input values to predicted values.

As with other Powerhouse tools, the model results (predicted values or scores) are measurably the most representative and as accurate as possible given the training data available.



Build Predictions & Scores

Powerhouse OPFIT and MAXIT models provide:

- The performance of high-powered non-linear techniques such as support vector machines and neural networks
- With the speed of linear techniques such as linear regression.
- Results from a two-window wizard
- Right-click in a window to copy-and-paste statistics and graphs to MS Word, Excel, and PowerPoint.

Figure 4 shows the Powerhouse OPFIT model.

- Finds relationships between variables
- Identifies general effects, or trends, in data.

Figure 4

-more-

Cluster Modeling and Explanation Techniques

Current Practice: Many clustering algorithms require the analyst to start by asserting an arbitrary number of clusters for the tool to create. The tool then forces the data to fit into the asserted number of clusters. Various clustering algorithms have different metrics by which to judge cluster effectiveness, and many tools leave it to analyst intuition to determine when the clustering is good enough. Depending on the clustering method, there may be more or less parameters to adjust, further complicating the process.

Powerhouse Solution: The Powerhouse method reads the internal information map to directly discover the inherent structure present in the data, and reveals the optimal number of clusters. No programming or manipulation, nor arbitrary selection of the number of clusters to be created, is required to build the OPTICL™ model. The OPTICL™ (Optimum Information Clustering) cluster model, allows the user to fully explore and characterize the information relationships through simple visualizations.

For valid business or other reasons, the analyst may require more or less clusters than the optimal number. Powerhouse allows the analyst to explore the inherent structure in the data at various levels of granularity (various numbers of clusters). Also, unlike other analytics tools, Powerhouse provides an integrated approach to cluster discovery and visualization, all by pressing a single button.

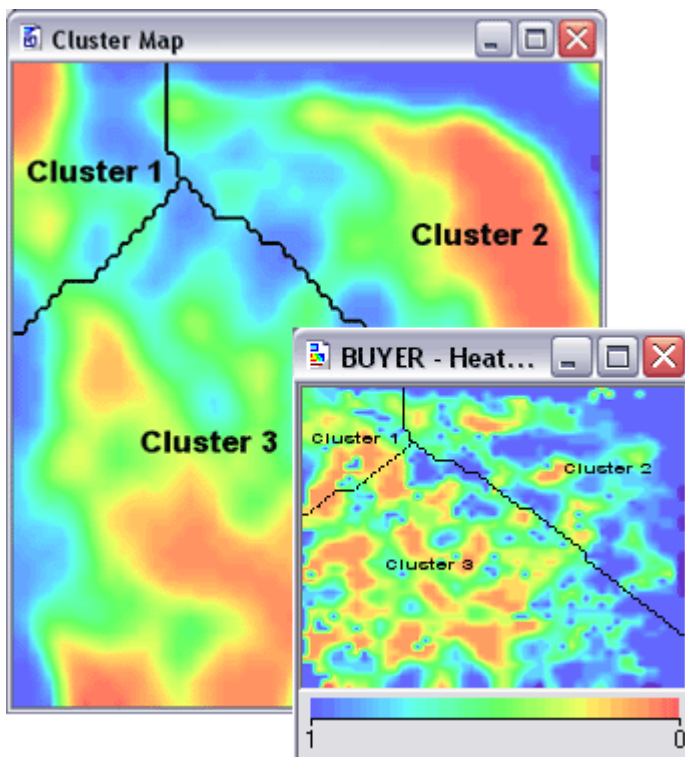


Figure 5

Cluster & Explain the Data

(no programming or manipulation required)

Cluster maps allow users to easily understand “what’s in this data anyway?”

Powerhouse finds the naturally occurring clusters in a dataset using all of the loaded variables -- regardless of type.

Records positions are organized by a clustering process. Records sharing similar information are close together, and those that are dissimilar are farther apart.

Heat Maps map variable values to color (red equals high density/value, blue is low density/value).

The Heat Map scale shows which colors correspond to which values (numeric) and which categories (categoricals) on the map.

Each variable gets its own Heat Map, and exploration tools reveal detailed attributes of each variable.

###

Oct. 17, 2005

All trademarks, service marks and company names are the property of their respective owners.