GenIQ, Ltd

Genetic Modeling & Data Mining Software **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581
516.791.3544 ~ fax 516.791. 5075

br@dmstat1.com
www.GenIQModel.com

Data Mining Quiz

Bruce Ratner, Ph.D.

Data mining is a high-concept: having elements of fast action in its development, glamour as it stirs the imagination for the unconventional and unexpected, and a mystic that appeals to a wide audience that knows curiosity feeds human thought. Conventional wisdom, in the DM space, has it that everyone knows what data mining is. [1] Everyone does it – that's what they say. I don't believe it. I know that everyone talks about it; but only a small self-seeking group of data analysts genuinely does data mining. I make this empirical assertion based on my experience as a statistical modeler, data miner, and genetic-analyst consultant for many years. As a pseudo-proof of my assertion, I present a *data-mining quiz.*

Consider Table 1, below, with ten individuals whose IDs range from 1 to 10. There are two variables curiously named with double X, XX1 and XX2. The dependent/target variable assumes five Yeses and five Noes (or Nos, *Unabridged* Merriam-Webster, 2002). No information about the three variables is provided, as it is not needed for the quiz. (Why?) Validation for the quiz is moot. (Why?) If many excellent models are submitted to me, then a data-mining quiz II will address validation.

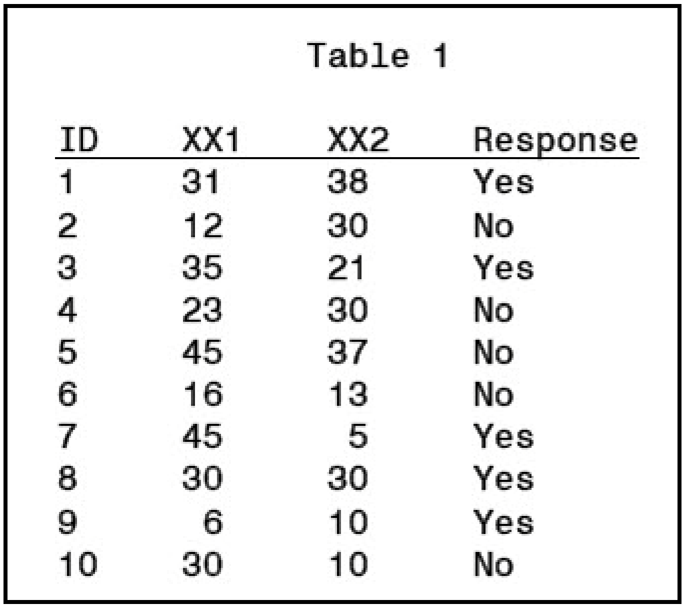


Table 1

ID	XX1	XX2	Response
1	31	38	Yes
2	12	30	No
3	35	21	Yes
4	23	30	No
5	45	37	No
6	16	13	No
7	45	5	Yes
8	30	30	Yes
9	6	10	Yes
10	30	10	No

Objective of Data Mining Quiz

The objective of the data mining quiz is to build a binary (Response = Yes) model with XX1 and XX2 such that its solution ranks first the Yeses, followed by the Noes, irrespectively of IDs, as close to, or exactly as in Table 2, below, indicating a perfect solution. My solution yields Table 2, and its every feature is in Figure 1 on page 3. I do not present in any detail my approach in this article, as it would require an unsparing amount of a ream of paper (i.e., 500, formerly 480 sheets of paper, *Concise Oxford English Dictionary*), equivalent to a ninety-minute presentation. If you are interested in a presentation of my solution, in a webcast format, please contact **me** to set-up a mutually convenient date and time.

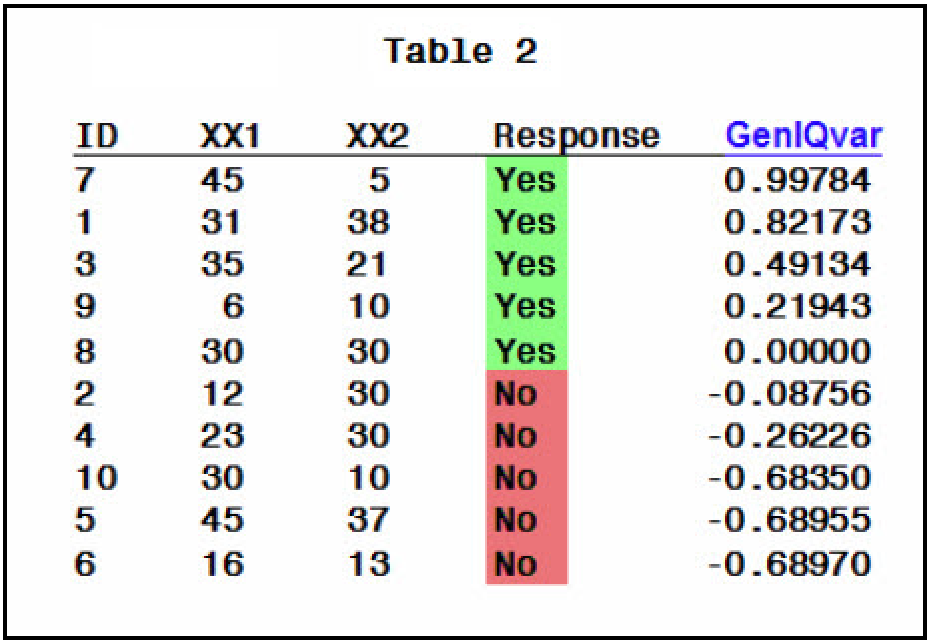
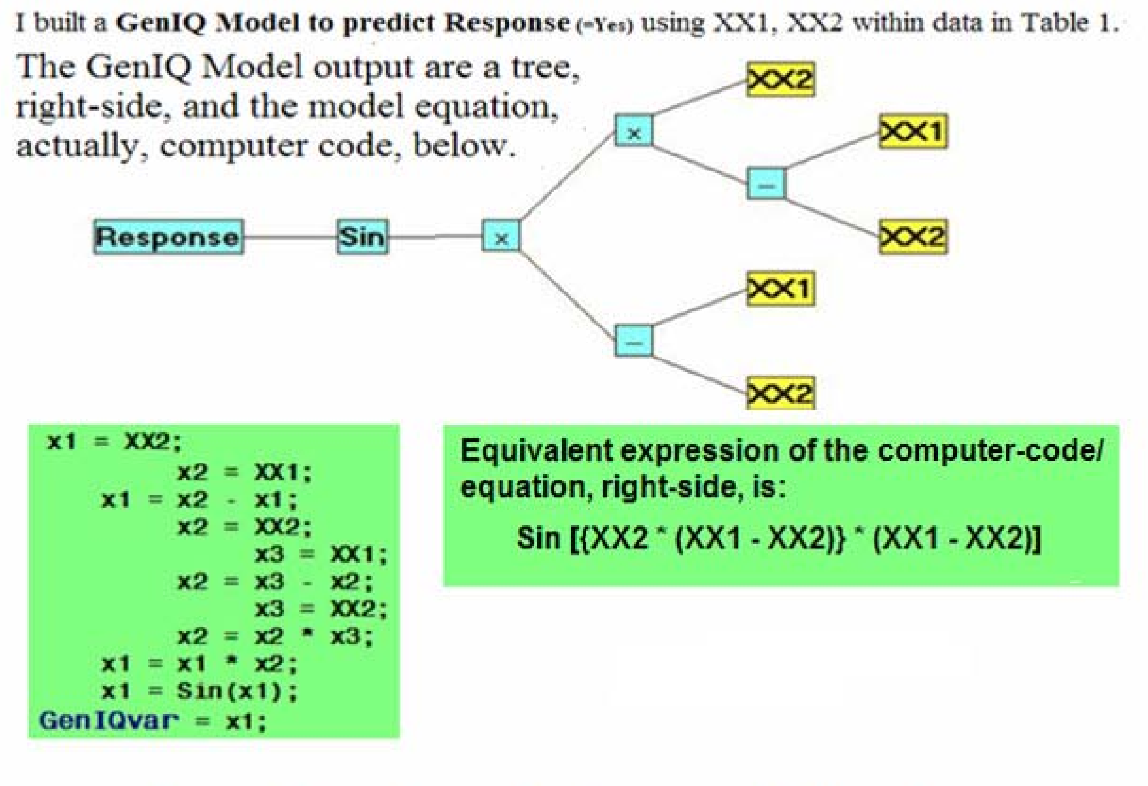


Table 2

ID	XX1	XX2	Response	GenIQvar
7	45	5	Yes	0.99784
1	31	38	Yes	0.82173
3	35	21	Yes	0.49134
9	6	10	Yes	0.21943
8	30	30	Yes	0.00000
2	12	30	No	-0.08756
4	23	30	No	-0.26226
10	30	10	No	-0.68350
5	45	37	No	-0.68955
6	16	13	No	-0.68970

My solution with all the particulars is in Figure 1, below.



I built a **GenIQ Model to predict Response** (=Yes) using XX1, XX2 within data in Table 1.

The GenIQ Model output are a tree, right-side, and the model equation, actually, computer code, below.

```
x1 = XX2;
          x2 = XX1;
    x1 = x2 - x1;
          x2 = XX2;
                x3 = XX1;
          x2 = x3 - x2;
                x3 = XX2;
          x2 = x2 * x3;
    x1 = x1 * x2;
    x1 = Sin(x1);
GenIQvar = x1;
```

Equivalent expression of the computer-code/ equation, right-side, is:

Sin [{XX2 * (XX1 - XX2)} * (XX1 - XX2)]

Figure 1: Bruce Ratner's Solution to the Data Mining Quiz

Well, now it’s time for you to share your data-mining solution to that self-seeking group of data miners. Please forward it to **me**.

I *really* hope you care to share.

Regards,

Bruce