

## Data Mining Quiz – II

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 *data mining quiz – II*. (See first [Data Mining Quiz](#)).

Consider Table 1, below, with ten individuals whose IDs range from 1 to 10. There are two predictor variables curiously named with double X, XX1 and XX2. The continuous dependent/target variable *Profit* assumes ten values ranging from 10 to 1. (How can Profit be considered a continuous variable when it assumes ten discrete integer values?) 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 – III will address validation.

ID	XX1	XX2	Profit
1	45	5	10
2	32	33	9
3	33	38	8
4	32	23	7
5	10	6	6
6	46	38	5
7	25	12	4
8	23	30	3
9	5	5	2
10	12	30	1

## Objective of Data Mining Quiz – II

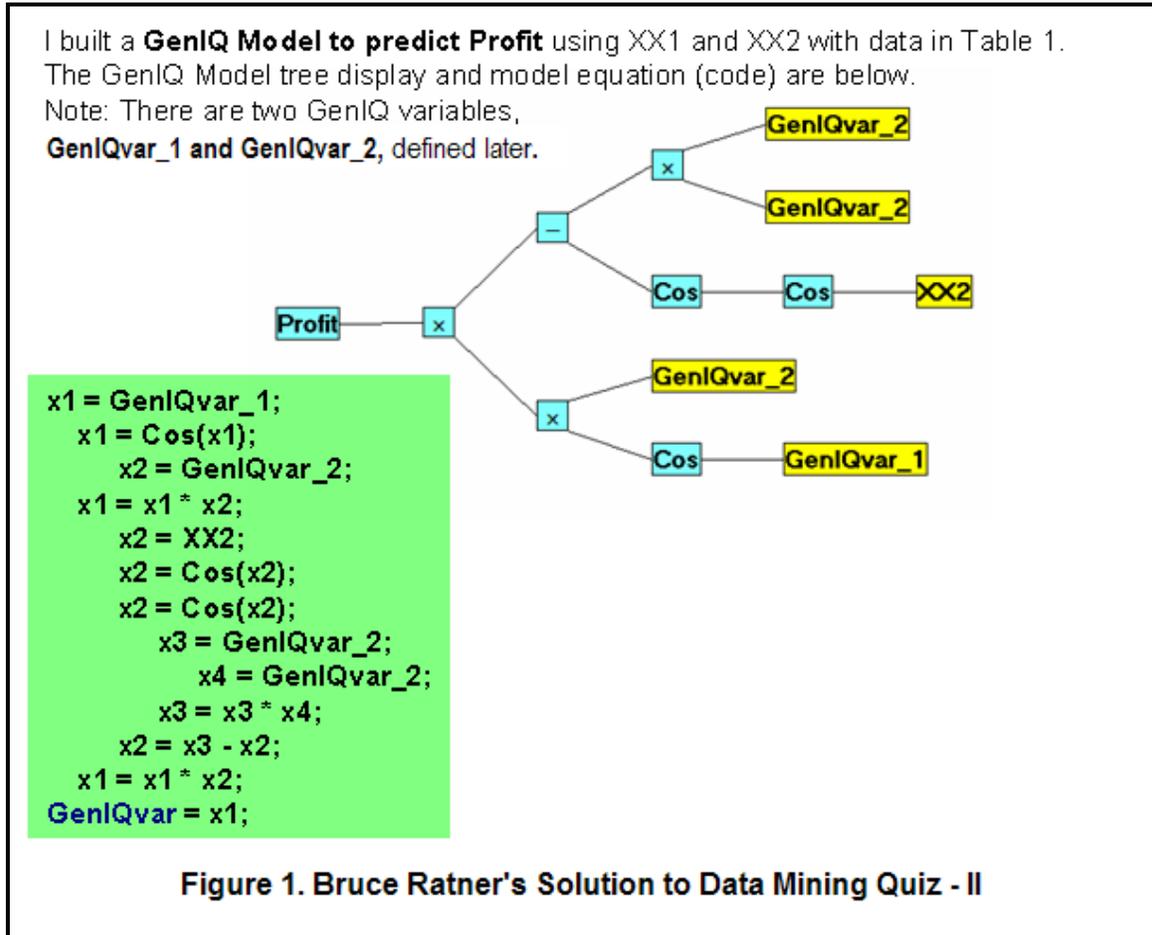
The objective of the data mining quiz is to build a model with XX1 and XX2 such that its solution descendingly ranks Profit values from 10 to 1, irrespectively of IDs, as close to, or exactly as in Table 2, below, indicating a perfect solution. My solution yields Table 2, and all its features are in Figures 1 – 3 on pages 3 – 5. *GenIQvar* is my solution's predicted *y-hat*. I do not present any detail of 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. (However, *the answer my friend is blowing in* [www.GenIQModel.com](http://www.GenIQModel.com).) 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. An almost never uttered distinguishing point regarding modeling is now getting its just-notice, discussed below: My solution is a *representative* model.

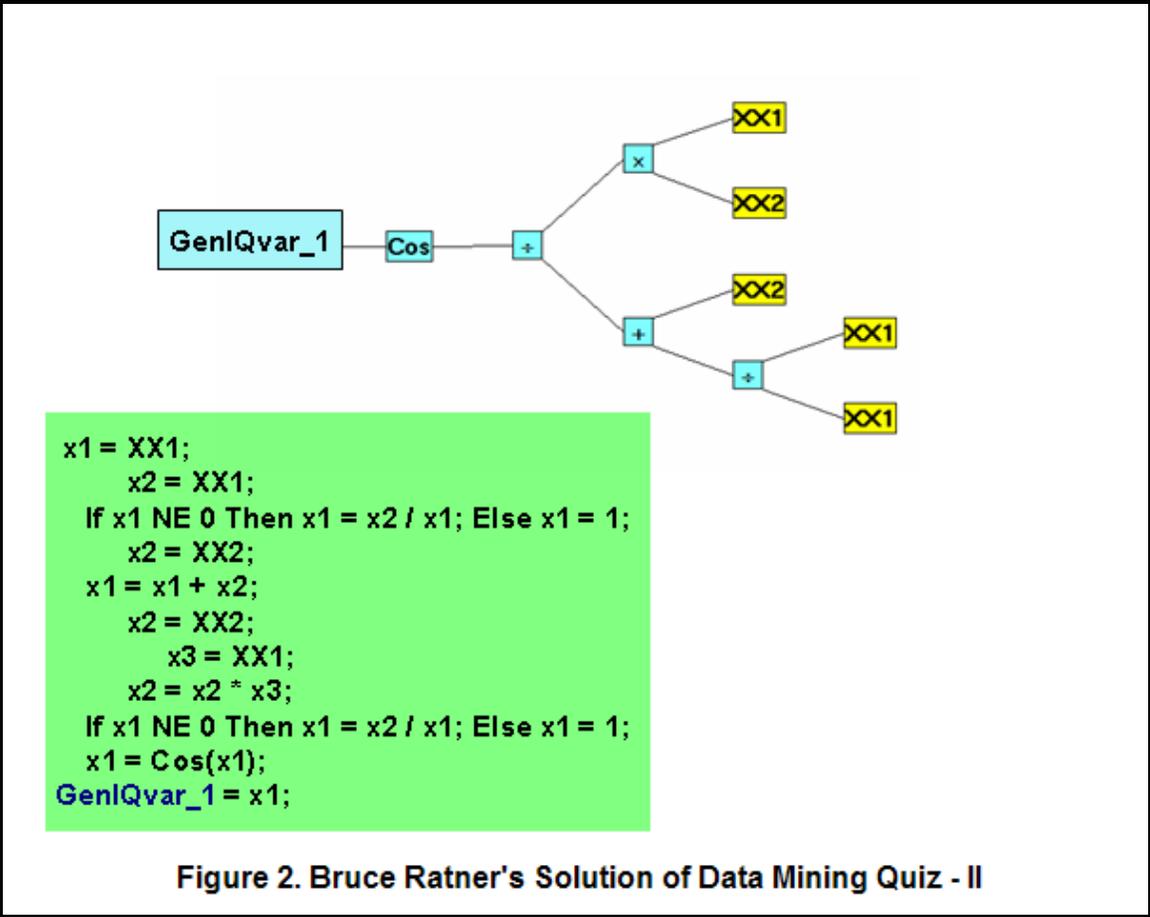
A model is built to *represent* a training data such that it will welcome a data-point visitor from a validation data. The model states to the visitor the refrain *Make yourself at home*. The visitor nicely nestles in the model, a snug fit that is free from significant error. The visitor receives an accurate prediction.

In contrast, a model is *not* built to *reproduce* a training data. If so, then a visitor from the validation data will not *feel at home*. The visitor encounters an uncomfortable fit in the model because s/he probabilistically *does not* look like a typical data-point from the training data. Thus, the misfit visitor takes a poor prediction. Statistical factoid: Any model can reproduce a dataset by adding predictor variables until there are perfect fits and predictions for each training data-point. Such models are useless models.

ID	XX1	XX2	Profit	GenIQvar
1	45	5	10	9.27644
2	32	33	9	6.22359
3	33	38	8	0.60790
4	32	23	7	0.29207
5	10	6	6	0.09151
6	46	38	5	0.06930
7	25	12	4	-0.12350
8	23	30	3	-0.19258
9	5	5	2	-0.26727
10	12	30	1	-0.31133

My solution with all the particulars is in Figures 1 – 3, below.

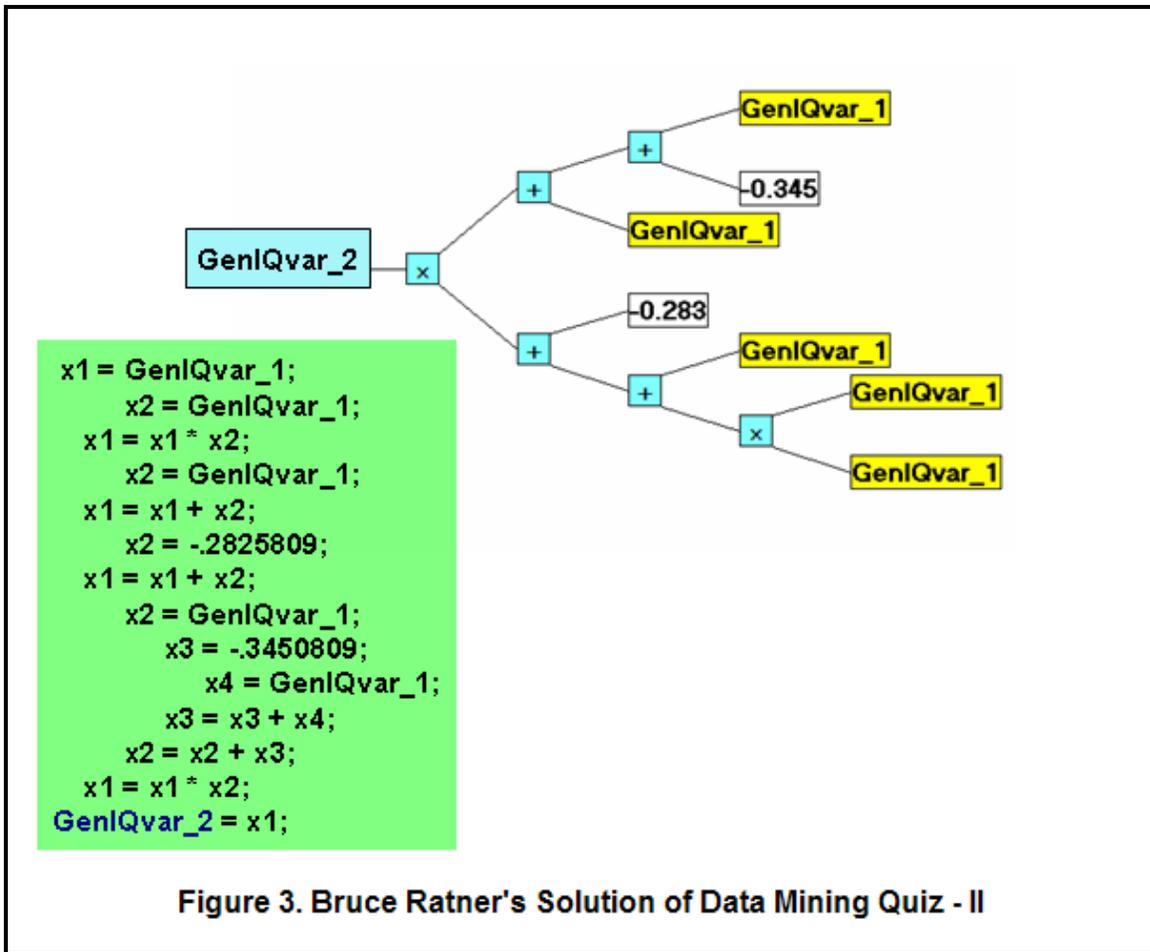




```

x1 = XX1;
x2 = XX1;
If x1 NE 0 Then x1 = x2 / x1; Else x1 = 1;
x2 = XX2;
x1 = x1 + x2;
x2 = XX2;
x3 = XX1;
x2 = x2 * x3;
If x1 NE 0 Then x1 = x2 / x1; Else x1 = 1;
x1 = Cos(x1);
GenlQvar_1 = x1;
  
```

Figure 2. Bruce Ratner's Solution of Data Mining Quiz - II



Well, now it's time for you to share your data-mining solution (using data in Table 1) to that self-seeking group of data miners. Please forward it to [me](#).

I *really* hope you care to share.

Regards,