Basics of Modeling in a Data Mining Context

Keith Cranford
Child Support Division, Office of the Attorney General of Texas

Abstract

Data mining has become an industry buzz word, but what is it and how does it differ from traditional statistical methods? This paper addresses these questions by discussing the modeling process involved in data mining. The emphasis is on the process and model evaluation with only brief mention of modeling techniques. SAS® Enterprise Miner is used to demonstrate certain aspects of the process.

Data Mining

Data mining has become a hot topic, but what is it? A couple of definitions found in the data mining literature give some insight into this question.

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*[1]

> *Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database.*[2]

These definitions indicate that first, data mining involves a process. This helps organize your approach to a question and gives you direction. Second, the data used in data mining is not initially collected for the purpose of the current question. This means that a large part of data mining is bringing together disparate data with the current question in mind. This is an important, and usually very time-consuming, part of the process. Also, this part of the process involves the preparation of the data.

Thirdly, data mining involves a search for relationships in the data. This is what most would consider the "mining" aspect of the process. Both definitions indicate a search for "unsuspected" or "unknown" relationships. Many times, though, the resulting relationships may be intuitive, but the data mining exercise confirms what experience has shown. Also, the techniques used for searching for relationships can include standard statistical methods; they are just applied in new ways. The important thing is that you can still utilize these familiar techniques.

Finally, the result of data mining should be useful and understandable. An important aspect of this part of the definition is that to be useful, you must start with a clear question. This helps direct your inquiry, so that the answer you get is to the correct question. As Dorian Pyle states, "It is no use looking for an answer unless there is a question."[3]

These aspects of data mining will be discussed by defining the process of data mining. This will give a basis for examining the finer points, although time and space will limit some of the discussion.

Data Mining Process

Since data mining is a process, it is important to clearly delineate this process. Although each miner will have his or her own approach, the general steps should be similar. The following is the process that will be discussed in this paper.

1. Define the relevant business question
2. Determine and prepare the data that meet requirements based on the business question
3. Model with current data
4. Evaluate competing models with current data
5. If possible, re-evaluate selected model with alternate data from different time frame
6. Apply model to answer question
7. Continuously evaluate selected model to ensure model is still valid

This paper will touch on each of these steps, but will concentrate on steps 3 and 4. These are the steps that involve SAS Enterprise Miner. However, you must not neglect the other steps.

What's the Question?

The first step in this process is to define the relevant business question. This may sound simple, but it can sometimes be the most difficult, and it is unquestionably the most important. The question drives the entire process. If the question is not clear or is ill-defined, the data collection step will be difficult and will most likely result in the wrong data being collected. The well-known caveat "Garbage in, garbage out" generally refers to bad data will result in a bad model. A primary reason for bad data in data mining is a poorly defined question.

Secondly, the question should be "business" related. The answer to the question should be actionable, that is, is there something you can do, once you have an answer to the question. Experience and subject matter knowledge plays a big role in the question definition. You may need to enlist others, especially those with a vested interest in the outcome, to help frame the question.

Where's the Data?

If defining the business question is the most important step, the second step of collecting and preparing appropriate data is the most time-consuming. This is also another step where you will probably have to rely on others who have the knowledge of various sources of data. Data collection is also where data mining tends to stray from traditional statistics, where data is commonly collected for the express purpose of answering a certain question. In data mining, the data has been collected for other purposes, for example, financial or customer demographic information, but you are bringing the data together to answer the new question.

In addition to collecting the data, there is generally a need to prepare the data. This may entail transforming character data into numeric values, aggregating categorical data, combining several variables into a single variable, or normalizing numeric variables in some way. Data preparation also helps you to gather an understanding of the data before the modeling begins, preparing both the data and the data miner. A better understanding of the data results in a better model.

How Do the Data Relate?

Once the data has been prepared, you are ready to start the first modeling step. Modeling is where SAS excels, so SAS, and especially Enterprise Miner, can play an important role. You might notice that the modeling was broken into several steps – model, evaluate and re-evaluate (and re-evaluate continually). In this first model step, building a model for the data, it is important to consider the other model steps as well. One way this is done is to divide your data into three partitions – training, validation, and testing. The first two of these groups are used in the modeling, whereas the testing group is used for evaluation.

1. Training: used in estimation of model parameters
2. Validation: used in "validating" model, including the stopping point for some techniques such as neural networks
3. Testing: used to test the model more purely (not involved in estimation)

The training data, since it is used for parameter estimation, should be as large as possible. This is of particular concern if there is limited data available. One conception of data mining is that you are working with large amounts of data. Many times this is true, but, if the question restricts the data in some way, then this may not be the case.

Validation data is not used for parameter estimation, but can influence the final model. Some iterative techniques, such as neural networks and decision trees, use the validation data to determine when to stop iterating. This data does not have to be very large, but should be sufficiently large to evaluate a model.

Finally, testing data is used to test competing models. This data is not used in parameter estimation or for individual model selection, but instead gives an indication of how well a model would work with new data. This helps show how consistent or robust a model may be when deployed. A model may perform very well on the training data (self-selective), but not so well on the testing data. This would indicate that the model would not work very well when it is deployed.

There are different schemes that can be used for dividing the data. A couple of suggestions are:

- 40-30-30: Balanced approach
- 60-30-10: Greater training allocation

The first of these is the default in SAS Enterprise Miner. It provides more balance among the partitions. This would be most appropriate for very large data sets. The second was suggested by Gordon Linoff and allocates more to the training set. This provides more data for estimating the model parameters. There is no magic in these numbers. Some other options would be a 70-20-10 split, if you have very limited data, or 60-20-20, which would give more data for the testing component.

After the data is partitioned, you are ready to apply various models. Here is where you can reach back and use the statistical methods with which you are familiar or try some new ones. The most common techniques used in data modeling are:

- Linear regression
- Logistic regression
- Decision, or classification, trees
- Neural network

This paper is not intended to give a full discourse on these methods, but a brief description and potential use for these will be addressed.

Linear regression models the relationship between a continuous response variable with a set of predictor, or explanatory, variables. In particular, the relationship is linear in the parameters. The simplest example of this is

$$y = a + bx + e,$$

where y is the response variable, x is the predictor variable, and e is the error term for a particular values of y and x.

Logistic regression is similar to linear regression, except the response variable is categorical and usually binary, and the linear relationship is relative to the log odds ratio, that is,

$$\log [B / (1 - B)] = a + bx + e,$$

where B = Prob(y is the event of interest).

Decision, or classification, trees create groups of observations according to division rules. These divisions are made to make the groups as homogeneous as possible in terms of the response variable. At each step, an explanatory variable is used to split the observations into groups. This is done until further splitting does not improve the classification according to certain rules. The final result is a set of groups with certain characteristics of the explanatory variables, which separate the events based on the response variable. This technique produces a very heuristic approach to the modeling, but less structure.

Neural networks are a more complex technique making them a little harder to explain. A neural network is composed of nodes, sometimes referred to as neurons, within layers, called perceptrons, with various weighted connections among the nodes. The explanatory variables are used as inputs and the response variable is the output signal to the network. The result is a highly non-linear function of the explanatory variables to predict the response variable. This very flexible model form is one of the advantages of neural networks. The model is not restricted to a linear form and can adapt to the nuances of the relationship between the response variable and the explanatory variables. The biggest disadvantage is interpretability. It is difficult to provide a simple description of the model's relationship.

A difference between data mining and traditional statistics is the use of these methods. Traditionalists would typically choose a single technique and attempt to find the best model using this technique. Data miners try various techniques and let the data determine which performs the best, with no prior bias to the technique to use.

Which Model is Best?

Evaluation of the potential models using the current data is the fourth step in the data mining process. Enterprise Miner provides both graphical methods and statistical measures to evaluate the competing models. Examples of both will be discussed. There are many more that are provided by Enterprise Miner

that will not be covered here, but those included give a good sampling and are those I have found most useful.

The graphical methods provide nice visual assessments of the models. These make it easy to see quickly which models perform better or worse. They can be used to narrow your focus to the best few models, before a more detailed evaluation is made. These methods include:

- Lift Chart
- Concentration Chart
- Receiver Operating Characteristic (ROC) Curve
- Score Distribution

The lift chart gives cumulative lift in response when working with a certain top percentage of the data. This lift is in comparison to a random, or mean, model. This chart can be used in a couple of ways. Comparing the lift chart of a single model for the training versus the validation sample, gives an indication of whether the model is fitting well and if the model is consistent, or robust. If the curve starts high and drops quickly, the model is showing strong predictability. Also, if the training and validation curves follow one another closely, it indicates some consistency (although this will be more relevant with the testing data). If they deviate, this is the first sign of trouble and that the model may not hold up well to a new set of data. Figure 1 illustrates using the lift chart in this manner. The training and validation curves lie close to one another in this example, which is what you would like to see.

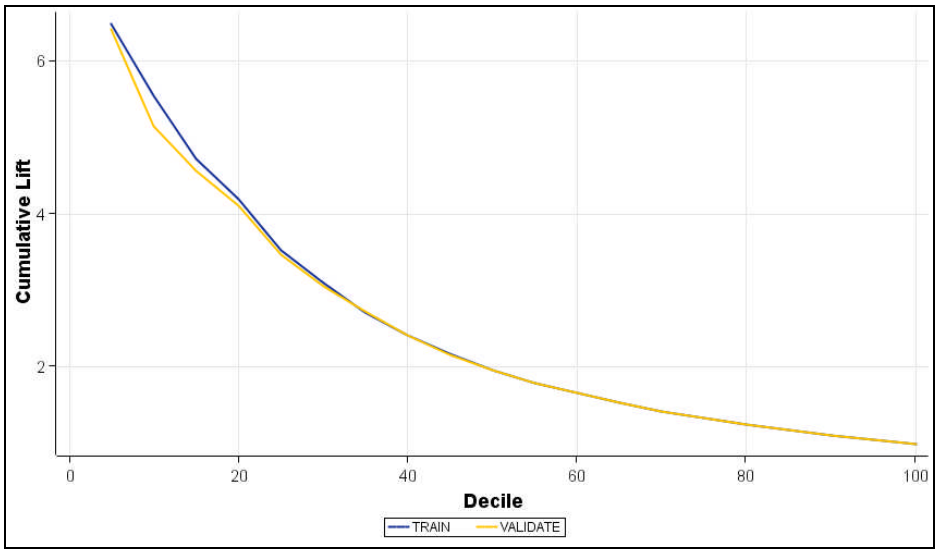

Figure 1: Lift Chart with Training vs. Validation

A lift chart is also used to compare competing models. This is best done with the test data, because this is a better judge of how the competing models will work with fresh data. In this comparison, you again would prefer models with the highest curves that remain higher for a longer period. Since these are cumulative charts, all of the curves eventually will converge, so it's the first segments that are of concern. In figure 2 it appears that Reg8 performs best for a longer period, followed by the Decision Tree. Neural3 does very well in the first decile, but falls off quickly. If you expected to only work or market a small percentage of the list, then the neural network model would be a good choice, but otherwise it would not be the best choice.

A concentration chart provides similar information to the lift chart. This chart graphs the cumulative captured response for a given percentage of the data. You would prefer models with a larger percentage of responses for the lowest percentage of data, which would result from more responses for higher probability scores. As with the lift chart, this gives a good visual comparison of competing models. Figure 3 compares the same set of models as figure 2, except with a concentration chart. A similar conclusion can be made that the Reg8 lies consistently above the rest, followed by the Decision Tree. Neural3 does particularly poorly after the second decile.

Receiver operating characteristic (ROC) curves give a measure of predictive accuracy for a logistic (binary) model. This curves graphs 1 minus specificity versus sensitivity for various cutoff values. Sensitivity is the

proportion of events predicted as such, or true positives, and specificity is the proportion of nonevents predicted as events, or false positives. The higher this curve deviates from a 45° line, the better the model (really the lower the error rate). Similar to the lift and concentration charts, it is best to use the test data to compare models, and the model with the highest curve is best. Figure 4 illustrates the ROC curve. This curve follows the concentration chart very closely. Note that Enterprise Miner 5.2 adds the baseline curve for comparison, but this has the effect of changing the colors of the lines. Reg8 is purple in figure 3, but cyan in figure 4.

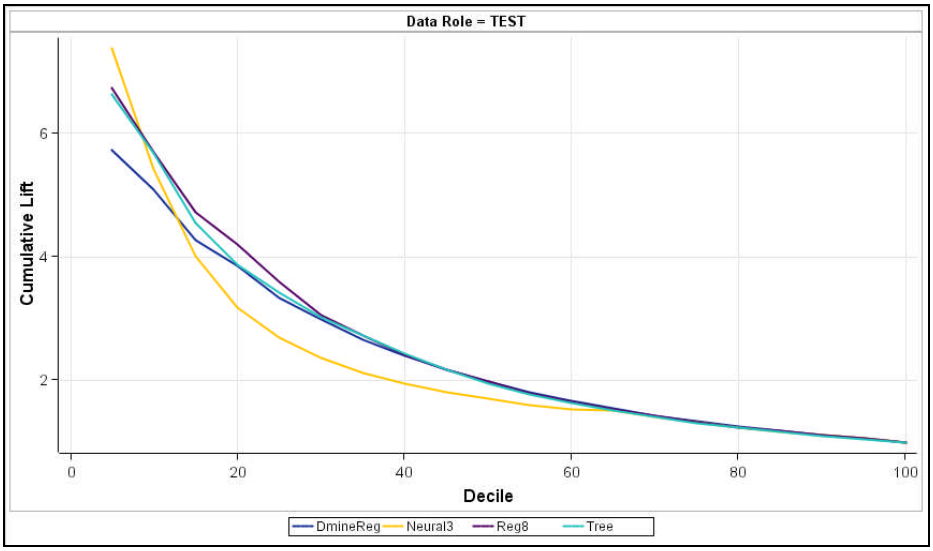

Figure 2: Lift Chart with Testing for Comparing Models

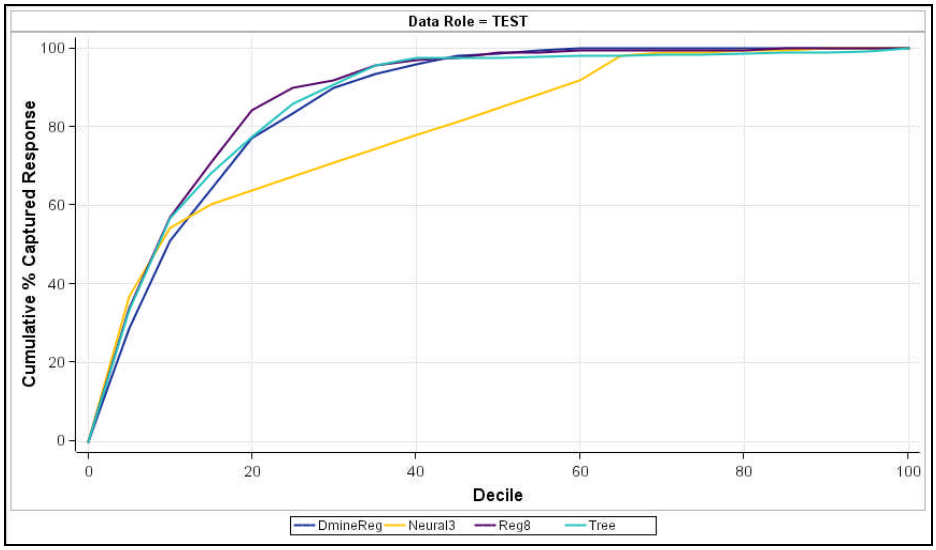

Figure 3: Concentration Chart

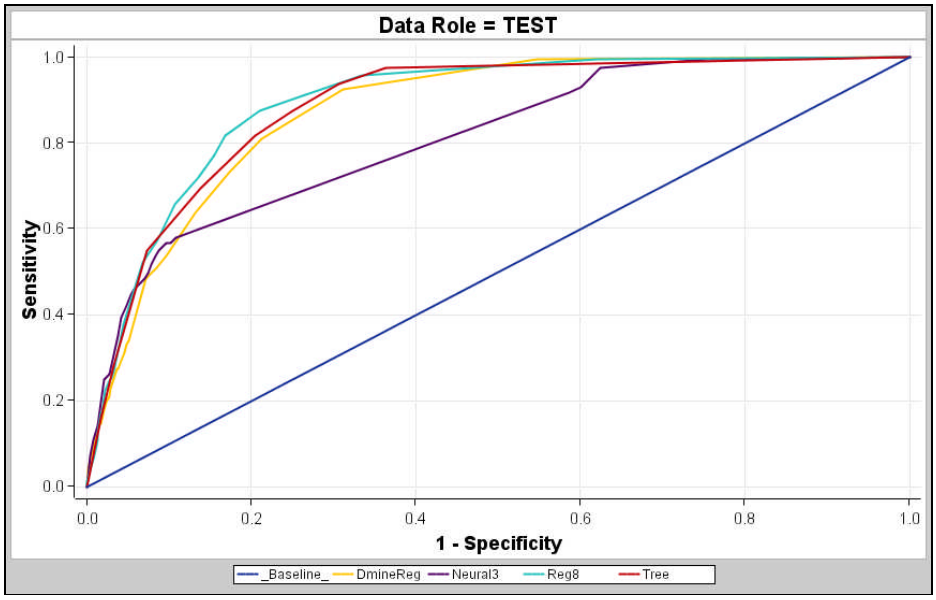

Figure 4: Receiver Operating Characteristic (ROC) Curve

The final graphical technique is a distribution of scores. This shows the distribution of model scores for events and non-events, graphing model scores by the percent of either events or non-events in a score range. You want non-events to be more heavily concentrated to the left (lower scores) and the events to the right (higher scores). This technique is a little more difficult to use for comparing models, but it is good for evaluating single models. In figure 5, over 80% of the non-events have a score less than .05, while a larger percentage of the events are associated with higher scores. This also shows the difficulty in interpreting the graph when you have rare events.

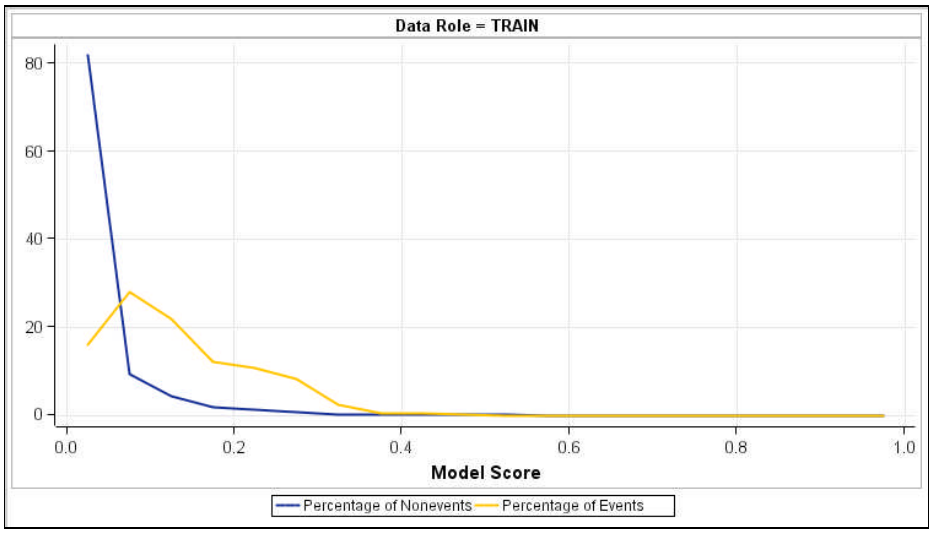

Figure 5: Distribution of Scores

Although graphs are good for visually comparing competing models, you usually need some hard numbers as well. There are several statistical measures that can be used to compare and evaluate models. The ones discussed here are:

- Average squared error (standard deviation)
- Misclassification error
- ROC Index
- Kolmogorov-Smirnov statistic

The average squared error is basically the standard deviation of the model. It compares actual values to predicted values, so the lower the average squared error the better. This measure also gives some degree of confidence in future predicted values.

Misclassification rates are available only for binary models and depend on the value of the cutoff value for classifying an observation as an event or non-event. This measures the proportion of false positives and false negatives, so you would like for these to be low. Unless you have a good idea of a reasonable cutoff value, this measure will not be very meaningful, so caution should be used with this measure.

The ROC Index relates to the ROC curve discussed earlier. The ROC Index is the area under the ROC curve and ranges from .5 (equating to the 45^{o} line) to 1 (no errors). The better the model, the closer the ROC Index will be to 1.

Based on a test for comparing an empirical distribution to a standard distribution, such as a test for normality, the Kolmogorov-Smirnov, or K-S, statistic is used in data mining to compare the model distribution of events to a random distribution. The statistic is simply the maximum difference between the two cumulative distributions, so the larger the value the greater the separation and the better the model.

Table 1 displays these statistics for the models shown earlier in the graphs, with highlights indicating the model with the "best" value for each statistic. Again, the misclassification rate does not have much meaning in this context, especially with rare events. The decision tree had the lowest average squared error. Since this measure is dependent on scale, you should only compare model values from the same data. The logistic regression model has the highest values for ROC Index and K-S statistic. The ROC Index of nearly .90 indicates that this is a very strong predictive model, and the K-S statistic of .67 indicates this as well. This confirms what was seen in the graphical comparisons, but with statistics to back it up.

Table 1: Comparison of Evaluation Statistics

Model	Misclassification Rate	Average Squared Error	ROC Index	K-S Statistic
DmineReg	0.031100	0.028193	0.87833	0.63344
Neural3	0.031300	0.027803	0.80913	0.47526
Reg8	0.031499	0.027390	0.89481	0.67272
Tree	0.031300	0.027312	0.88083	0.63173

Does the Model Work with New Data?

In addition to testing a model with a test, or hold-out, sample, evaluation of the model with data in a different time period provides information on how a model can be generalized. This test also assures that the model is not time-dependent. This aspect is important, because the model will be applied to new data in a different time frame as well. Data is gathered in a similar manner using the same criteria as the modeling data. The same statistics used in the model evaluation can be used in the out of period testing. This step can be though of as a confirmation step, giving you confidence that the model will work well with new data.

What's the Answer?

At this point you are ready to implement the model by applying the results to answer the question posed in the first step. Implementation typically entails applying the model to a new (and current) set of data and taking some action according to results. The importance of making the answer actionable becomes evident at this step. For example, for a binary model, the records are scored with the probability of the event. Some action, such as mailing a solicitation, is performed on the top scoring records.

Is the Model Still Good?

Finally, once a model is implemented, the same evaluation criteria can be used to ensure that a model is still valid. This constant evaluation helps determine when a new model may be necessary.

Conclusion

Data mining is a process, and, as with any process, you must take care to follow through with each step. Each step is important and has its place in producing the best possible model, and, ultimately, provides the best answer to the question at hand. This will also give you the greatest chance at success.

Contact Information

Keith Cranford
keith.cranford@cs.oag.state.tx.us

Footnotes

[1] From Principles of Data Mining by Hand, David, et. al.
[2] From Applied Data Mining by Paolo Giudici.
[3] Quote from Dorian Pyle in Data Preparation for Data Mining.

References

Hand, David, Heikiki Mannila, and Padhraic Smyth. Principles of Data Mining. Cambridge, Massachusetts: The MIT Press, 2001.

Giudici, Paolo. Applied Data Mining. West Sussex, England: John Wiley & Sons Ltd., 2003.

Pyle, Dorian. Data Preparation for Data Mining. San Francisco: Morgan Kaufmann Publishers, 1999.

Pyle, Dorian. Business Modeling and Data Mining. San Francisco: Morgan Kaufmann Publishers, 2003.

Rud, Olivia Parr. Data Mining Cookbook. New York: John Wiley & Sons, Inc., 2001.