Predictive Modeling in Enterprise Miner Versus Regression

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

We investigate the difference between regression models in SAS/Stat and compare them to the predictive models in Enterprise Miner. In large samples, the p-value becomes meaningless because the effect size is virtually zero. Therefore, there must be another way to determine the adequacy of the model. In addition, logistic regression cannot be used to predict rare occurrences. Such a model will be highly accurate, generally predicting all occurrences as non-occurrence. However, it will have no practical use whatsoever in identifying those at high risk. In contrast, predictive modeling in Enterprise Miner was designed to accommodate large samples and rare occurrences as well as providing many measures of model adequacy.

INTRODUCTION

Predictive modeling includes regression, both logistic and linear, depending upon the type of outcome variable. It can also include the generalized linear model. However, there are other types of models also available, including decision trees and artificial neural networks under the general term of predictive modeling. Predictive modeling includes nearest neighbor discriminant analysis, also known as memory based reasoning. These other models are nonparametric and do not require that you know the probability distribution of the underlying patient population. Therefore, they are much more flexible when used to examine patient outcomes. Because predictive modeling uses regression in addition to these other models, the end results will improve upon those found using just regression by itself.

Some, but not all, of the predictive models require that all of the x-variables are independent. However, predictive models must still also generally assume the uniformity of data entry. Because of the flexibility in the use of variables to define confounding factors, we can consider the presence or absence of uniformity in the model itself. We can define a variable to model outcome, and to see how the inputs impact the severity outcome. Since the datasets used in predictive modeling are generally too large for a p-value to have meaning, predictive modeling uses other measures of model fit. Generally, too, there are enough observations so that the data can be partitioned into two or more datasets. The first subset is used to define (or train) the model. The second subset can be used in an iterative process to improve the model. The third subset is used to test the model for accuracy. It is also known as a holdout sample.

The definition of “best” model needs to be considered in this context as well. Just what do we mean by “best”? In a regression model, the “best” model is one that satisfies the criterion of uniform minimum variance unbiased estimator. In other words, it is only “best” in the class of unbiased estimators. As soon as the class of estimators is expanded, “best” no longer exists, and we must define the criteria that we will use to determine a “best” fit. There are several criteria to consider. For a binary outcome variable, we can use the misclassification rate. However, especially in medicine, misclassification can have different costs. For example, a false positive error is not as costly as a false negative error if the outcome involves the diagnosis of a terminal disease.

Another difference when using predictive modeling is that many different models can be used, and compared to find the one that is the best. We can use the traditional regression, but also decision trees and neural network analysis. We can combine different models to define a new model. Generally, use of multiple models has been frowned upon because it is possible to “shop” for one that is effective. Indeed, the nearest neighbor discriminant analysis can always find a model that predicts correctly 100% of the time when defining the model, but predicts 0% of the time for any subsequent data. When using multiple models, it is essential to define a holdout sample that can be used to test the results.

BACKGROUND

Predictive modeling routinely makes use of a holdout sample to test the accuracy of the results. Figure 1 demonstrates predictive modeling. In SAS, there are two different regression models, three different neural network models, and two decision tree models. There is also a memory based reasoning model, otherwise known as nearest neighbor discriminant analysis. These models are discussed in detail in Cerrito (2007). It is not our intent here to provide an introductory text on neural networks; instead, we will demonstrate how they can be used effectively to investigate the outcome data.

Figure 1. Predictive Modeling of Patient Outcomes

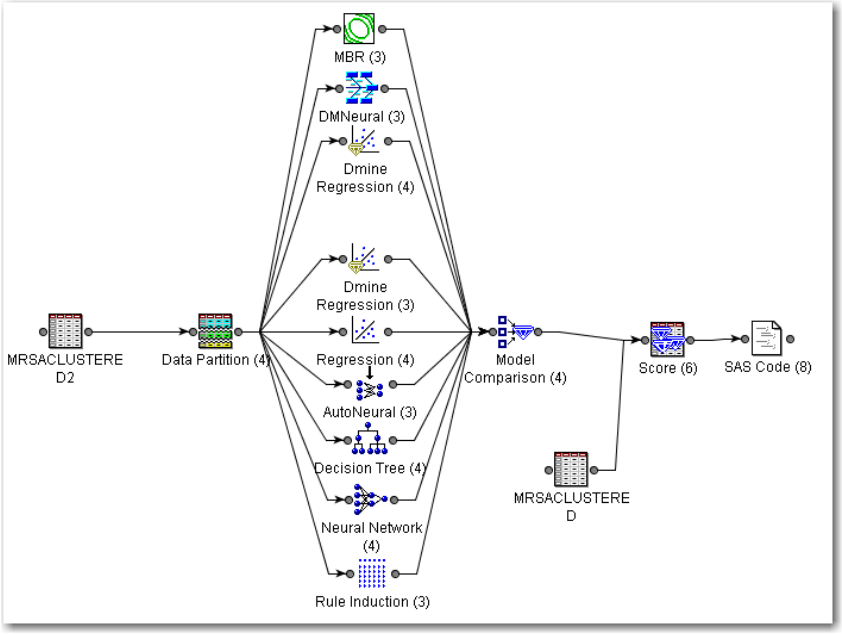


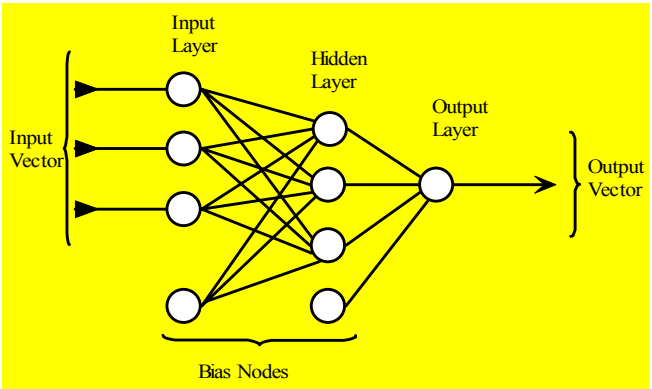
Figure 1 shows that many different models can be used. Once defined, the models are compared and the optimal model chosen based upon pre-selected criteria. The node labeled Model Comparison is used for this purpose. It compares all of the models and then chooses the optimal one based upon the pre-selected criterion. Model comparison can use several different statistics for comparison. The default here is the misclassification rate on the holdout sample.

Then, additional data can be scored (using the score node as shown in Figure 1) so that new patients who are admitted subsequently can have a severity level assigned to them. This figure also includes a data partition so that a holdout sample can be extracted in order to test the model results. It is important to be able to use the model to score subsequent data. When a patient severity model is defined, it should be tested on new data to demonstrate reliability.

There is still limited use of predictive modeling, with the exception of regression models, in medical studies. Most of the use of predictive modeling is fairly recent. (Sylvia, et al., 2006) While most predictive models are used for examining costs (Powers, Meyer, Roebuck, & Vaziri, 2005), they can be invaluable in improving the quality of care. (Hodgman, 2008; Tewari, et al., 2001; Weber & Neeser, 2006; Whitlock & Johnston, 2006) One recent study does indicate that predictive modeling can be used to target the most high risk patients for more intensive case management. (Weber & Neeser, 2006) It has also been used to examine workflow in the healthcare environment. (Tropsha & Golbraikh, 2007) Some studies focus on particular types of models such as neural networks. (Gamito & Crawford, 2004) In many cases, administrative (billing) data are used to identify patients who can benefit from interventions, and to identify patients who can benefit the most. Most of the use of predictive modeling is fairly recent.

Neural networks act like black boxes. There is no definite model or equation, and the model is not presented in the concise format available for regression. Its accuracy is examined similar to the diagnostics of the regression curve, including the misclassification rate, the AIC (Akaike's Information Criterion), and the average error. The simplest neural network contains a single input (an independent variable) and a single target (a dependent variable) with a single output. Its complexity increases with the addition of hidden layers and additional input variables (Figure 2).

Figure 2. Diagram of a Neural Network



With no hidden layers, the results of a neural network analysis resemble those of regression. Each input variable is connected to each variable in the hidden layer, and each hidden variable is connected to each outcome variable. The hidden layers combine inputs and apply a function to predict outputs. Hidden layers are often nonlinear.

The architecture of the neural network is used to define the model. There are two major types of neural network used, the MLP and the GLIM. MLP, the multi-layer perceptron, is the default model. A perceptron is a classifier that maps an input x to an output, f(x). The GLIM represents the more standard generalized linear model discussed in detail in Chapter 3. You should compare these two models to see the impact on the results. You can also define your own model, although this method is not recommended for beginners.

Decision trees provide a completely different approach to classification. A decision tree develops a series of if-then rules. Each rule assigns an observation to one segment of the tree, at which point another if-then rule is applied. The initial segment, containing the entire data set, is the root node for the decision tree. The final nodes are called leaves. Intermediate nodes (a node plus all its successors) form a branch of the tree. The final leaf containing an observation is its predictive value.

Unlike neural networks and regression, decision trees do not always work with interval data. Decision trees work better with nominal outcomes that have more than two possible results and with ordinal outcome variables. Missing values can be used in creating if-then rules. Therefore, imputation is not required for decision trees, although you can use it.

Predictive Modeling in SAS Enterprise Miner

For predicting a rare occurrence, one more node is added to the model in Figure 1, the sampling node (Figure 3). This node uses all of the observations with the rare occurrence, and then takes a random sample of the remaining data. While the sampling node can use any proportional split, we recommend a 50:50 split. Figure 4 shows how the defaults are modified in the sampling node of SAS Enterprise Miner to make predictions. Starting a project in SAS Enterprise Miner was discussed in Chapter 1.

Rule induction is a special case of a decision tree model. Figure 3 also shows three different neural network models and two regression models. The second regression model automatically categorizes all interval independent variables. There is one remaining model in Figure 3; the MBR or memory-based reasoning model. It represents nearest neighbor discriminant analysis. We first discuss the use of the sampling node in the process of predictive modeling. We start with the defaults for sampling node as modified in Figure 4.

The first arrow indicates that the sampling is stratified, and the criterion is level based. The rarest level (in this case, mortality) is sampled so that it will consist of half (50% sample proportion) of the sample to be used in the predictive model.

Figure 3. Addition of Sampling Node

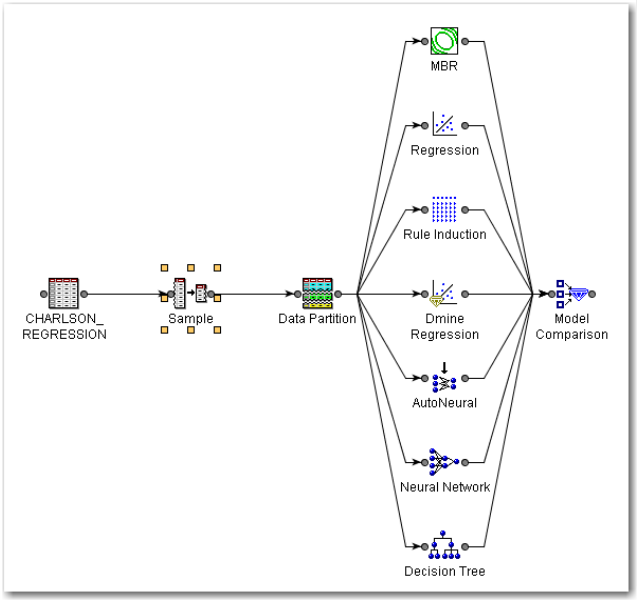


Figure 4. Change to Defaults in Sampling Node

Property	Value
Node ID	Smpl
Imported Data	
Exported Data	
Variables	
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Level Based
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Rarest Level
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No

In the following examples, we use a 50/50 split in the data. We use just three patient diagnoses of pneumonia, septicemia, and immune disorder to predict mortality. We use a 50/50 split in the data. We use all of the models depicted in Figure 1. According to the model comparison, the rule induction provides the best fit, using the

misclassification criterion as the measure of “best”. We first look at the regression model, comparing the results to those when a 50/50 split was not performed. The overall misclassification rate is 28%, with the divisions as shown in Table 1.

Table 1. Misclassification in Regression Model

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.1	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

The misclassification becomes more balanced between false positives and false negatives with a 50/50 split in the data. The model gives heavier weight to false positives than it does to false negatives. Table 2 shows the contrast without a 50/50 split. The false negative rate is extremely high even while the overall accuracy is 97%.

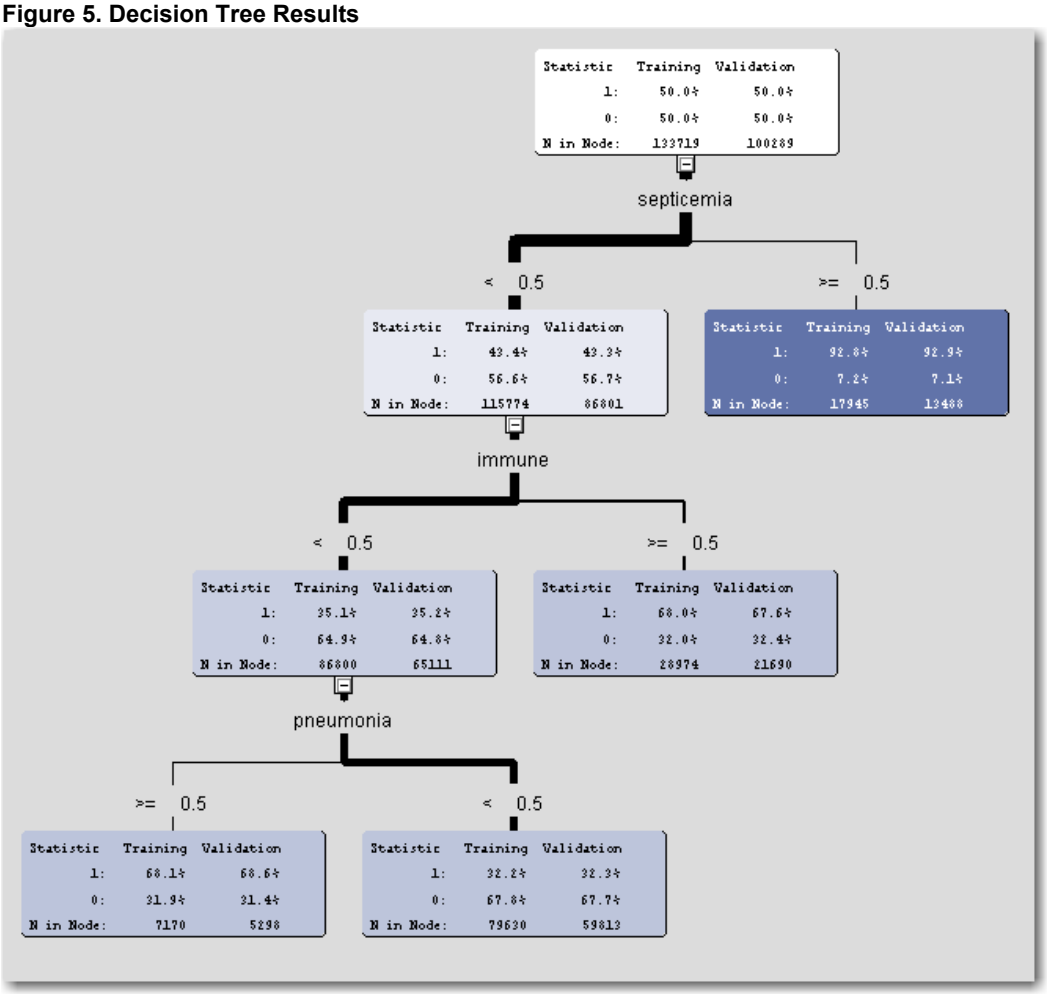
Table 2. Classification Table for Logistic Regression With Pneumonia and Septicemia

Classification Table									
Prob Level	**Correct**		**Incorrect**		**Percentages**				
	Event	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.580	782E4	0	167E3	0	97.9	100.0	0.0	2.1	.
0.600	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.620	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.640	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.660	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.680	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.700	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.720	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.740	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.760	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.780	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.800	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.820	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.840	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.860	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.880	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.900	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.920	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.940	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.960	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.980	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1
1.000	0	167E3	0	782E4	2.1	0.0	100.0	.	97.9

We first want to examine the decision tree model. While it is not the most accurate model, it is one that clearly describes the rationale behind the predictions. This tree is given in Figure 5. The tree shows that the first split occurs on the variable, Septicemia. Patients with Septicemia are more likely to suffer mortality compared to patients without Septicemia. The Immune Disorder has the next highest level of mortality, followed by Pneumonia.

Figure 5. Decision Tree Results



Since rule induction is identified as the best model, we examine that one next. The misclassification rate is only slightly smaller compared to the regression model. Table 3 gives the classification table.

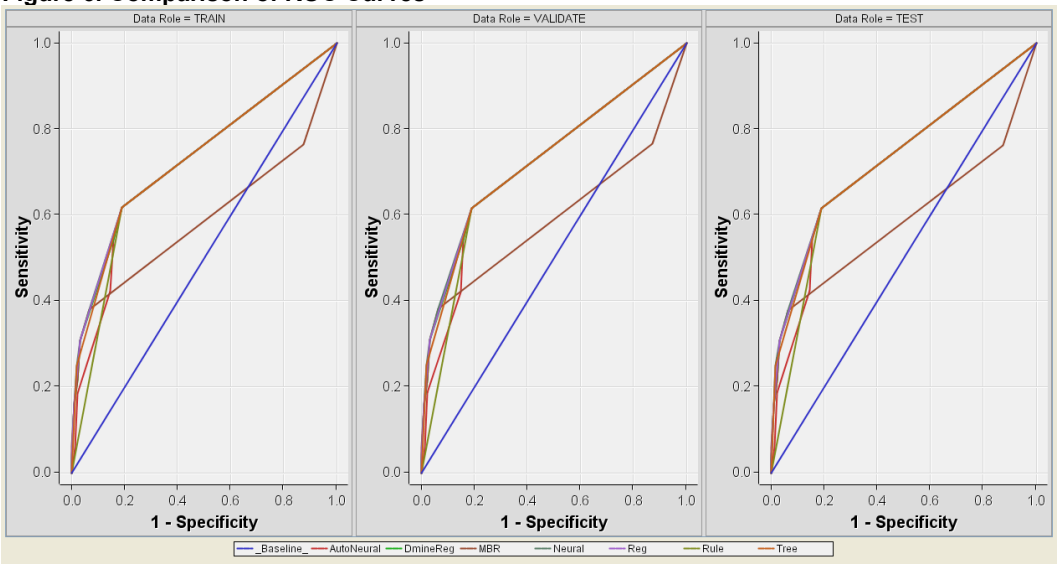
Table 3. Misclassification in Rule Induction Model

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.8	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

The results look virtually identical to those in Table 1. For this reason, the regression model, although not defined as the best, can be used to predict outcomes when only these three variables are used. The similarities in the models can also be visualized in the ROC (received-operating curve) that graphs the sensitivity versus one minus the specificity (Figure 6). The curves for rule induction and regression are virtually the same.

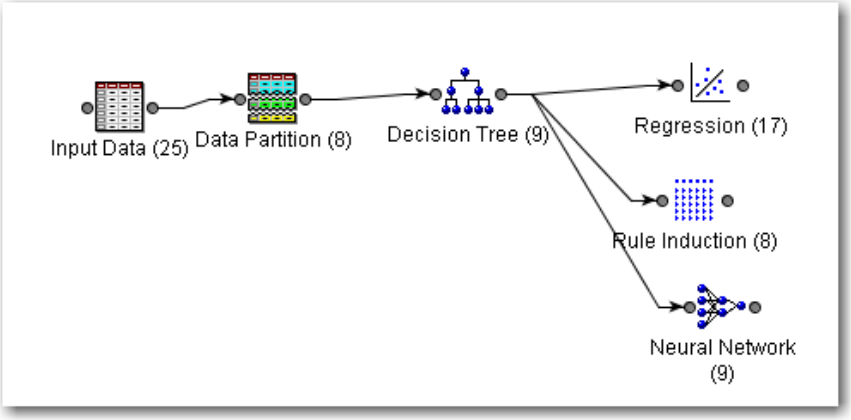
Figure 6. Comparison of ROC Curves



Many Variables in Large Samples

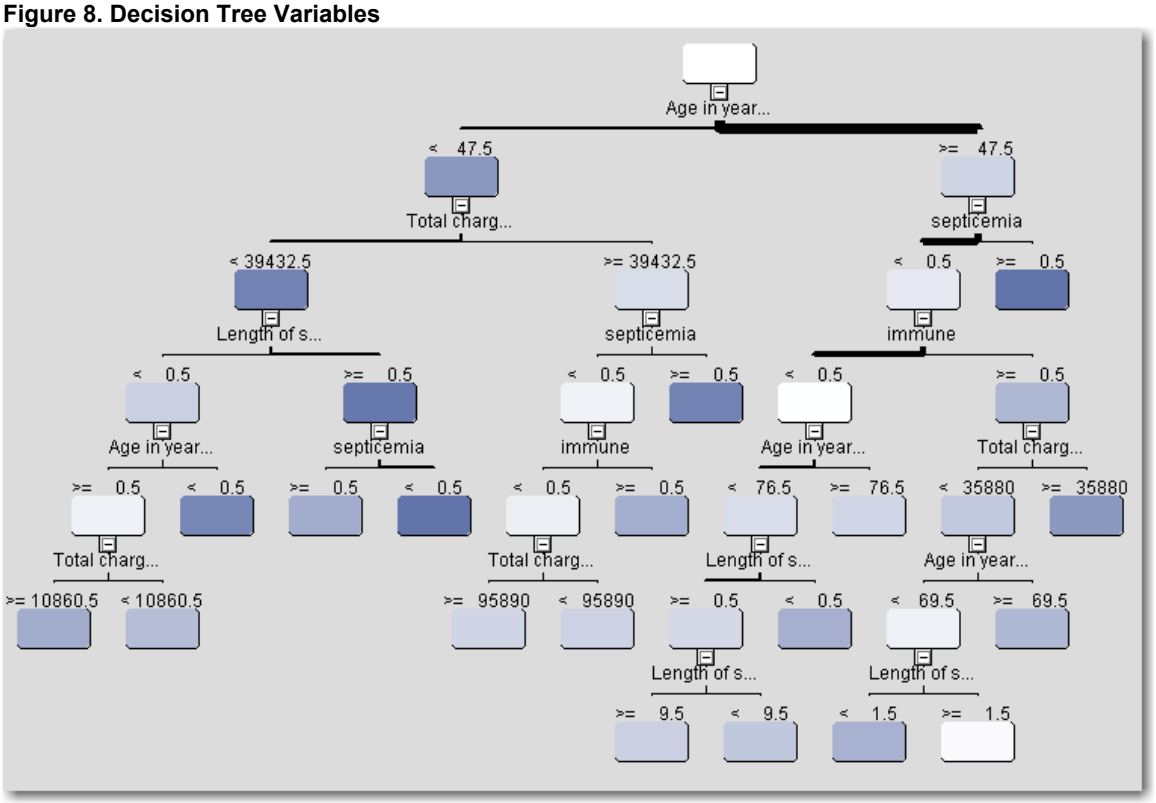
There can be hundreds if not thousands of variables collected for each patient. There can be far too many to include in any predictive model. We want to include all those variables that are crucial to the analysis, including potential confounders, but the use of too many variables can cause the model to over-fit the results, inflating the outcomes. Therefore, there needs to be some type of variable reduction method. In the past, factor analysis has been used to reduce the set of variables prior to modeling the data. However, there is now a more novel method available (Figure 7). In our example, there are many additional variables that can be considered in this analysis. Therefore, we use the variable selection technique to choose the most relevant. We first use the decision tree followed by regression, and then regression followed by the decision tree.

Figure 7. Variable Selection



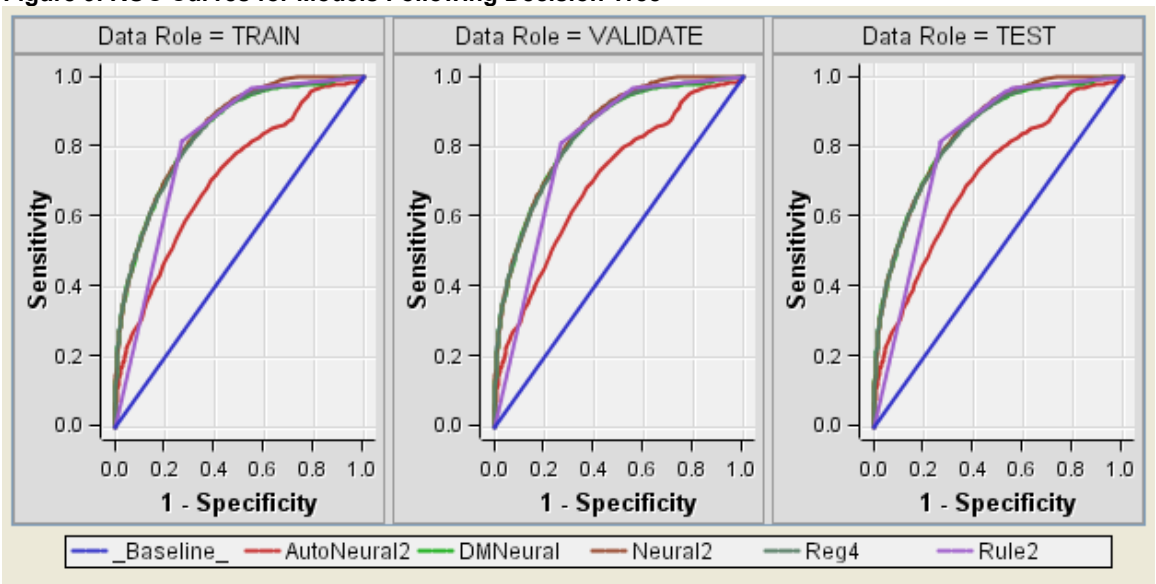
Using the decision tree to define the variables, Figure 8 shows the ones that remain for the modeling. Note that age, charges, and length of stay are at the beginning of the tree.

Figure 8. Decision Tree Variables



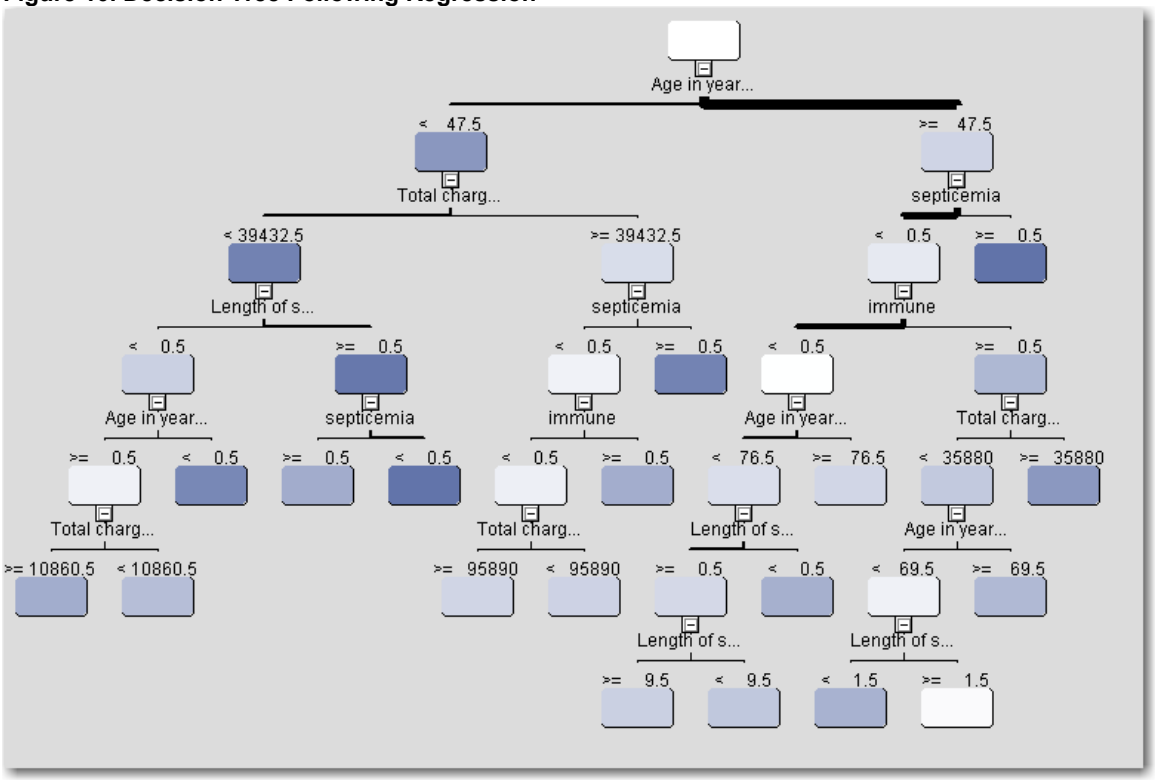
This tree shows that age, length of stay, having septicemia, immune disorder, length of stay and total charges are related to mortality. The remaining variables have been rejected from the model. The rule induction is the best model, and the misclassification rate decreases to 22% with the added variables. The ROC curve looks considerably improved (Figure 9).

Figure 9. ROC Curves for Models Following Decision Tree



The ROC curve is much higher compared to that in Figure 6. If we use regression to perform the variable selection, the results remain the same. In addition, a decision tree is virtually the same when it follows the regression compared to when it precedes regression (Figure 10).

Figure 10. Decision Tree Following Regression



The above example only used three possible diagnosis codes. We want to expand upon the number of diagnosis codes, and also to use a number of procedure codes. In this example, we restrict our attention to patients with a primary diagnosis of COPD (chronic obstructive pulmonary disease). This is approximately 245,000 patients in the NIS dataset. Table 4 gives the list of diagnosis codes used. Table 5 gives the list of procedure codes used to classify the patient's level severity. Here, we first examine a prediction of mortality.

Table 4. Diagnosis Codes Used to Predict Mortality

Condition	ICD9 Codes
Acute myocardial infarction	410, 412
Congestive heart failure	428
Peripheral vascular disease	441,4439,7854,V434
Cerebral vascular accident	430-438
Dementia	290
Pulmonary disease	490,491,492,493,494,495,496,500,501,502,503,504,505
Connective tissue disorder	7100,7101,7104,7140,7141,7142,7148,5171,725
Peptic ulcer	531,532,533,534
Liver disease	5712,5714,5715,5716
Diabetes	2500,2501,2502,2503,2507
Diabetes complications	2504,2505,2506

Condition	ICD9 Codes
Paraplegia	342,3441
Renal disease	582,5830,5831,5832,5833,5835,5836,5837,5834,585,586,588
Cancer	14,15,16,17,18,170,171,172,174,175,176,179,190,191,193, 194,1950,1951,1952,1953,1954,1955,1958,200,201,202,203, 204,205,206,207,208
Metastatic cancer	196,197,198,1990,1991
Severe liver disease	5722,5723,5724,5728
HIV	042,043,044

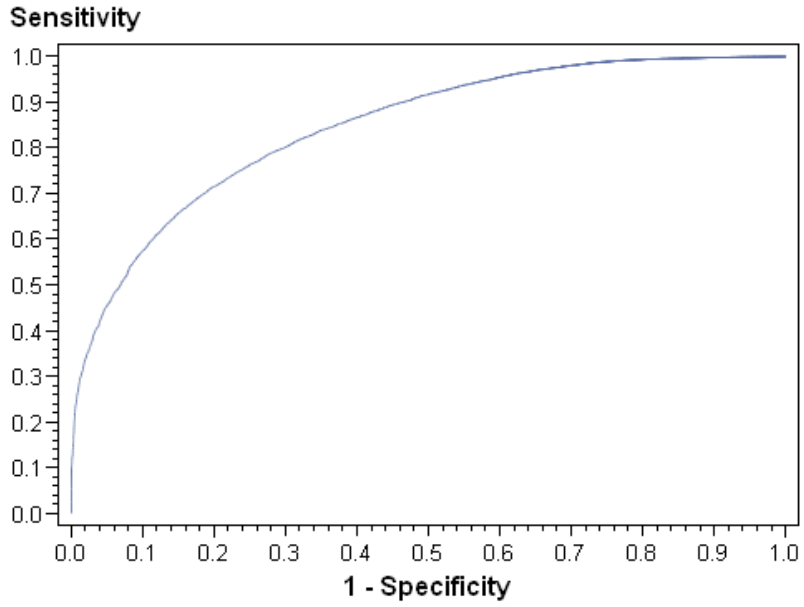
Table 5. Procedure Codes Used to Predict Mortality

pr	Procedure Translation	Frequency	Percent
9904	Transfusion of packed cells	17756	7.05
3893	Venous catheterization, not elsewhere classified	16142	6.41
9671	Continuous mechanical ventilation for less than 96 consecutive hours	10528	4.18
3324	Closed [endoscopic] biopsy of bronchus	8315	3.30
9672	Continuous mechanical ventilation for 96 consecutive hours or more	8243	3.27
3491	Thoracentesis	8118	3.22
3995	Hemodialysis	8083	3.21
9604	Insertion of endotracheal tube	7579	3.01
9921	Injection of antibiotic	6786	2.69
9394	Respiratory medication administered by nebulizer	6309	2.50
9390	Continuous positive airway pressure	7868	1.48
8856	Coronary arteriography using two catheters	7622	1.44
4516	Esophagogastroduodenoscopy [EGD] with closed biopsy	7516	1.42
966	Enteral infusion of concentrated nutritional substances	7203	1.36
3722	Left heart cardiac catheterization	6652	1.25
8853	Angiocardiography of left heart structures	6350	1.20
4513	Other endoscopy of small intestine	6343	1.19
3404	Insertion of intercostal catheter for drainage	5693	1.07
8741	Computerized axial tomography of thorax	5538	1.04
9915	Parenteral infusion of concentrated nutritional substances	5169	0.97
9907	Transfusion of other serum	4962	0.93
9396	Other oxygen enrichment	4937	0.93
4311	Percutaneous [endoscopic] gastrostomy	4831	0.91
3895	Venous catheterization for renal dialysis	4726	0.89
0331	Spinal tap	4362	0.82
3891	Arterial catheterization	3867	0.73
3327	Closed endoscopic biopsy of lung	3776	0.71

pr	Procedure Translation	Frequency	Percent
9339	Other physical therapy	3492	0.66
311	Temporary tracheostomy	3406	0.64
4523	Colonoscopy	3404	0.64

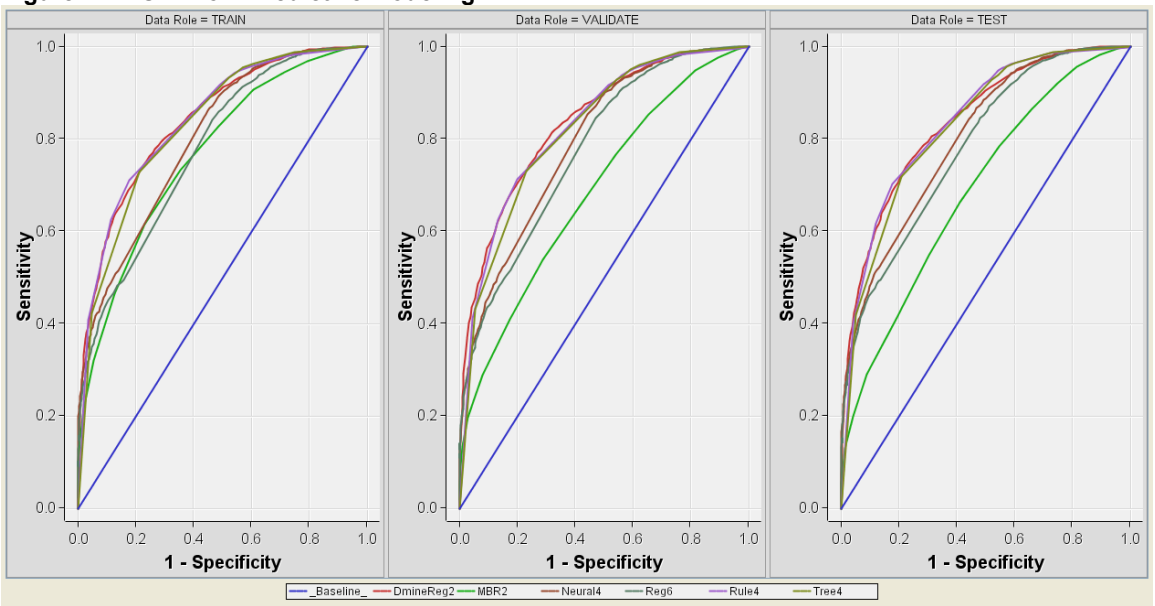
If we perform standard logistic regression without stratified sampling, the false positive rate remains small (approximately 3-4%), but with a high false negative rate (minimized at 38%). Given the large dataset, almost all of the input variables are statistically significant. The percent agreement is 84% and the ROC curve looks fairly good (Figure 11).

Figure 11. ROC Curve for Traditional Logistic Regression



If we perform predictive modeling and stratify the sample to the rarest level, the accuracy rate drops to 75%, but the false negative rate is considerably improved. Figure 12 gives the ROC curve from predictive modeling. It shows that the model predicts considerably better than chance in the testing set. We will examine the stratified sampling in more detail in the next section.

Figure 12. ROC From Predictive Modeling



Change in Split in the Data

The analyses in the previous section assumed a 50/50 split between mortality and non-mortality. We want to look at the results if mortality composes only 25% of the data, and 10% of the data. Table 6 gives the regression classification breakdown for a 25% sample; Table 7 gives the breakdown for a 10% sample.

Table 6. Misclassification Rate for a 25% Sample

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	80.4	96.6	10070	72.5
1	0	19.6	70.9	2462	17.7
0	1	25.6	3.3	348	2.5
1	1	74.4	29.1	1010	7.3
Validation Data					
0	0	80.2	97.1	7584	72.8
1	0	19.8	71.7	1870	17.9
0	1	23.7	2.9	229	2.2
1	1	76.2	28.2	735	7.0

Note that the ability to classify mortality accurately is decreasing with the decrease of the split; almost all of the observations are classified as non-mortality, but also at a cost of a high level of false positives. The decision tree for a 25% sample (Figure 13) is considerably different from that in Figure 10 with a 50/50 split. Now, the procedure of Esophagogastroduodenoscopy gives the first leaf of the tree; in Figure 10, the first split was on age followed by charges and length of stay.Thus, a change in the sampling can in and of itself be responsible for the outcomes predicted by the model.

Figure 13. Decision Tree for 25/75 Split in the Data

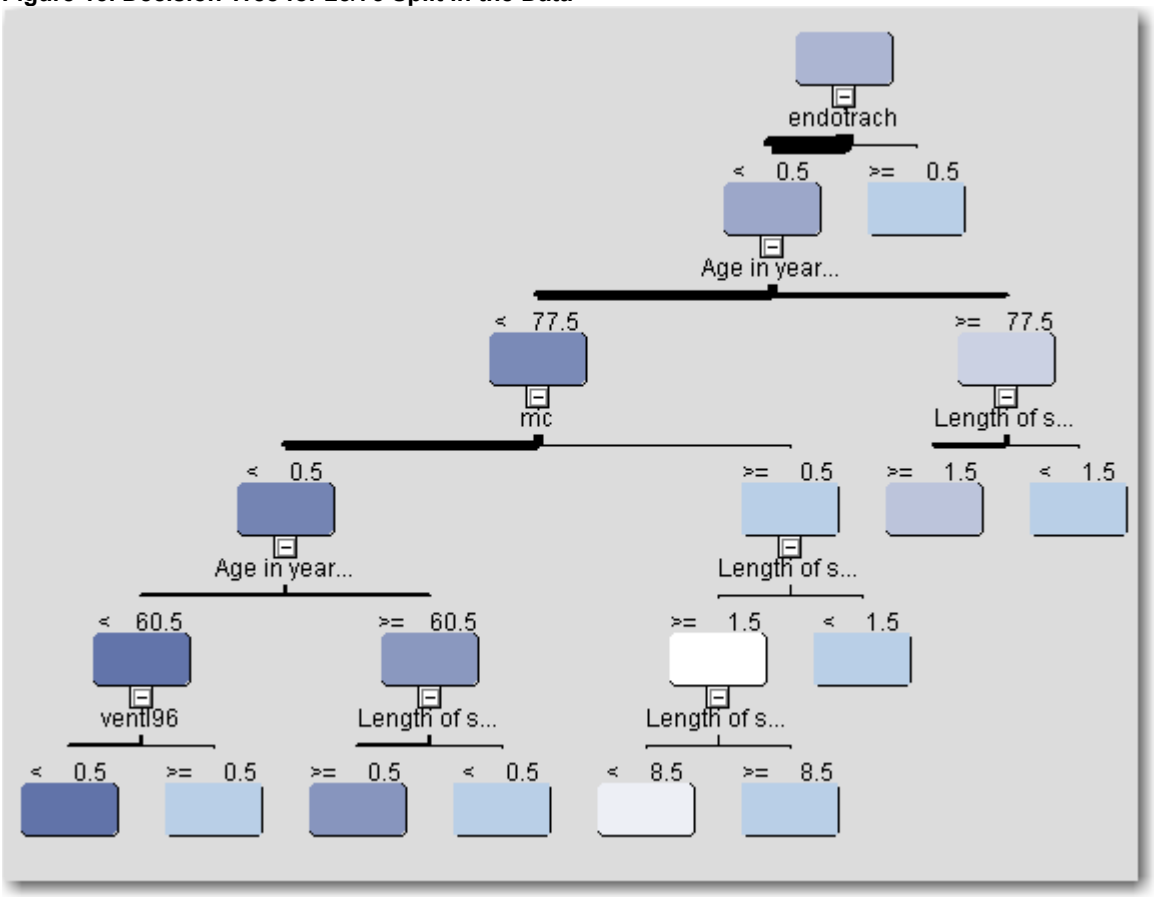
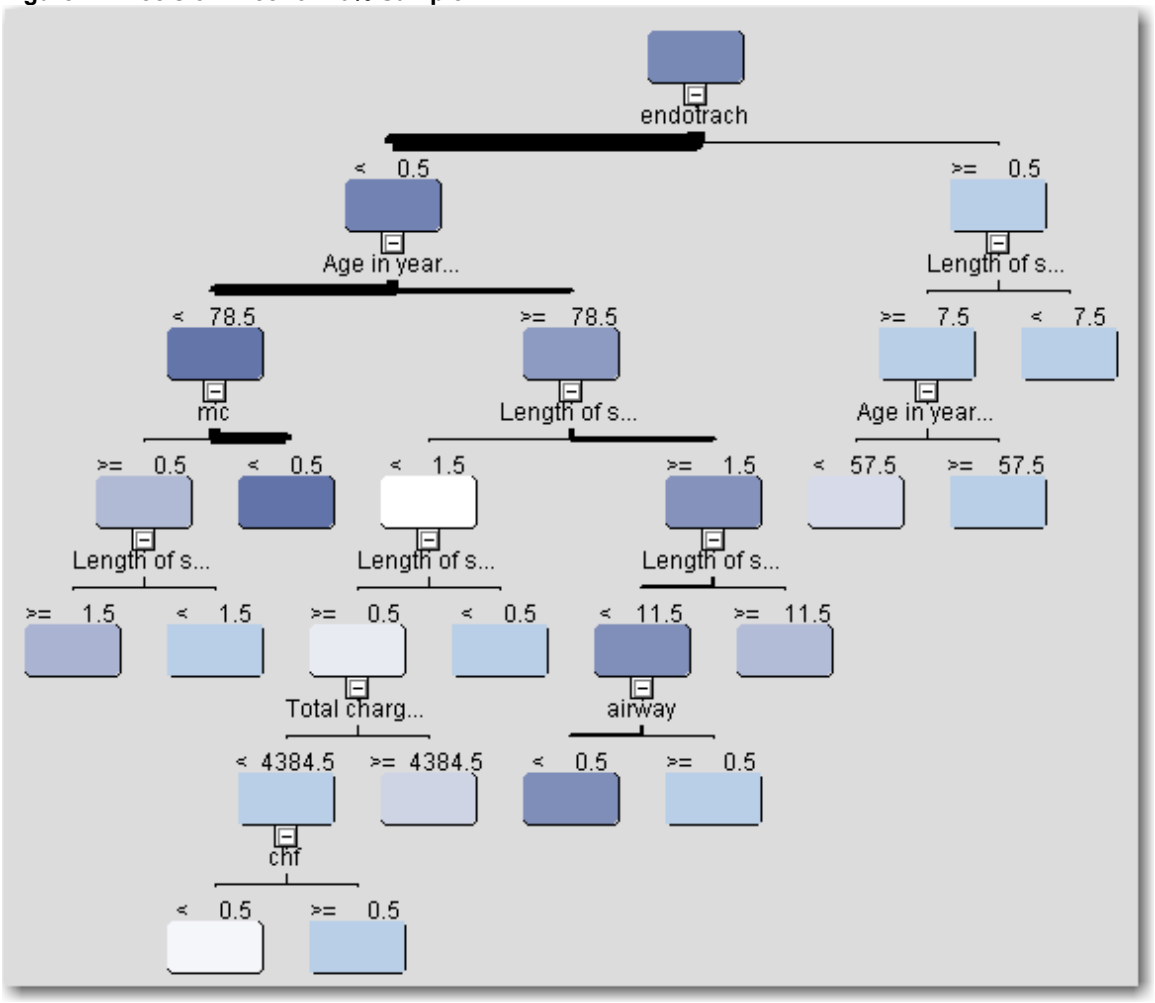


Table 7. Misclassification Rate for a 10% Sample

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	91.5	99.3	31030	89.4
1	0	8.5	83.5	2899	8.3
0	1	27.3	0.7	216	0.6
1	1	72.6	16.5	574	1.6
Validation Data					
0	0	91.5	99.2	23265	89.3
1	0	8.4	82.4	2148	8.2
0	1	27.8	0.7	176	0.7
1	1	72.2	17.5	457	1.7

Note that the trend shown in the 25% sample is even more exaggerated in the 10% sample. Figure 14 shows that the decision tree has changed yet again. It now includes the procedure of continuous positive airway pressure and the diagnosis of congestive heart failure.

Figure 14. Decision Tree for 10% Sample



Addition of Weights for Decision Making

In most medical studies, a false negative is more costly to the patient compared to a false positive. This occurs because a false positive generally leads to more invasive tests; however, a false negative means that a potentially life-threatening illness will go undiagnosed, and hence, untreated. Therefore, we can weight a false negative at higher cost, and then change the definition of a "best" model to one that minimizes costs. The problem is to determine which costs to use.

The best thing to do is to experiment with magnitudes of difference in cost between the false positive and false negative to see what happens. At a 1:1 ratio, the best model is still based upon the misclassification rate. Change to a 5:1 ratio indicates that a false negative is five times as costly compared to a false positive. A 10:1 ratio makes it ten times as costly. We need to determine if changes to this ratio result in changes to the optimal model.

Introduction to Lift

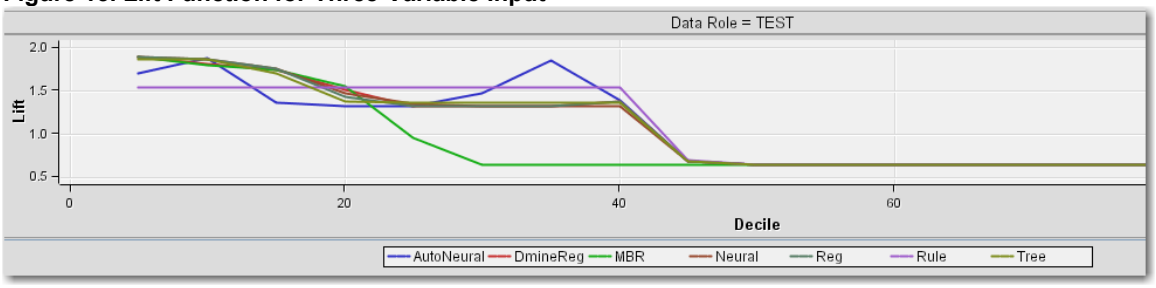
Lift allows us to find the patients at highest risk for occurrence, and with the greatest probability of accurate prediction. This is especially important since these are the patients we would want to take the greatest care for, and who will incur the highest costs and longest length of stay.

Using lift, true positive patients with highest confidence come first, followed by positive patients with lower confidence. True negative cases with lowest confidence come next, followed by negative cases with highest confidence. Based on that ordering, the observations are partitioned into deciles, and the following statistics are calculated:

- The *Target density* of a decile is the number of actually positive instances in that decile divided by the total number of instances in the decile.
- The *Cumulative target density* is the target density computed over the first *n* deciles.
- The *lift* for a given decile is the ratio of the target density for the decile to the target density over all the test data.
- The *Cumulative lift* for a given decile is the ratio of the cumulative target density to the target density over all the test data.

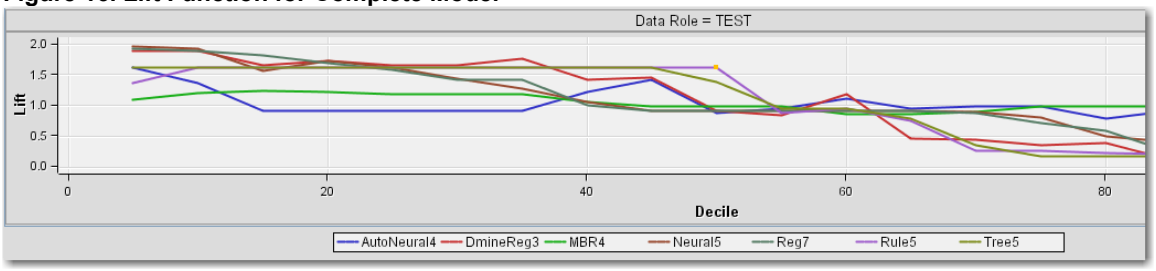
Given a lift function, we can decide on a decile cutpoint so that we can predict the high risk patients above the cutpoint, and predict the low risk patients below a second cutpoint, while failing to make a definite prediction for those in the center. In that way, we can dismiss those who have no risk, and aggressively treat those at highest risk. Lift allows us to distinguish between patients without assuming a uniformity of risk. Figure 15 shows the lift for the testing set when we use just the three input variables of pneumonia, septicemia, and immune disorder.

Figure 15. Lift Function for Three-Variable Input



Random chance is indicated by the lift value of 1.0; values that are higher than 1.0 indicate that the observations are more predictable compared to random chance. In this example, 40% of the patient records have a higher level of prediction than just chance. Therefore, we can concentrate on these 4 deciles of patients. If we use the expanded model that includes patient demographic information plus additional diagnosis and procedure codes for COPD, we get the lift shown in Figure 16. The model can now predict the first 5 deciles of patient outcomes.

Figure 16. Lift Function for Complete Model



Therefore, we can predict accurately those patients most at risk for death; we can determine which patients can benefit from more aggressive treatment to reduce the likelihood that this outcome will occur.

Predictive Modeling to Rank the Quality of Providers

Ultimately, the definition of a patient severity index is used in a model to rank the quality of healthcare providers. Unlike the standard logistic regression investigation of mortality, what we want to do to predict the quality of providers is to look not at the similarity between actual and predicted values, but to look at the difference between them. Quality rankings assume that if a provider does better than predicted, then it must be because the provider is delivering better care compared to a provider who does worse than predicted. This approach assumes that the predicted value is the established norm for a patient with a certain level of severity and demographics, and any deviation from that norm is a result of the quality of care. This assumption has not yet been validated.

We first look at the logistic regression model defined considering just the three patient conditions of pneumonia, septicemia, and immune disorder. Any choice of a threshold value will have a high false negative rate. It we use a threshold value of 0.720 or less, then the predicted value of mortality is equal to $4907/(782 \times 10^4)$. This is approximately 0.06% of the time overall. If we choose a threshold value above 0.760, the predicted mortality level becomes 0.034%. The only change in determining quality rankings when changing the threshold value will be to change the predicted value but not the order of the ranking of the providers. This is because the predicted mortality level is not really determined by the patient's actual severity; rather, it is defined uniformly for all patients.

We examine Table 2 together with a defined threshold value will determine the rankings of providers. Then, the worse the model is in predicting a provider's true mortality, the better that provider will appear in terms of quality. A model that can define a ranking will be "good" regardless of its ability to actually predict mortality.

Given that the three conditions of pneumonia, septicemia, and immune disorder all have higher mortality rates compared to patients generally, and patients with two of the three conditions can have a higer rate still, it is clear that hospitals with higher proportions of such patients will have higher mortality rates. We will examine a random selection of ten hospitals in detail. We will compare their rates of the three diseases, their overall actual mortality rate in comparison to the predicted value, and how the ten hospitals would be ranked by this model. Table 8 gives the proportion of death by hospital. The mortality rate ranges from a low of 0 to a high of 3.16. We want to know if the hospital with zero deaths has patients that are as severe as the hospital with 3.16% deaths.

Table 8. Mortality (All Causes) by Hospital

Table of DSHOSPID by DIED				**Table of DSHOSPID by DIED**			
Hospital Code	**DIED**		**Total**	**Hospital Code**	**DIED**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**			0	1	
1	2795 97.12 9.21	83 2.88 12.56	2878	**6**	5237 96.84 17.25	171 3.16 25.87	5408
2	1460 96.95 4.81	46 3.05 6.96	1506	**7**	1476 98.07 4.86	29 1.93 4.39	1505

Table of DSHOSPID by DIED				Table of DSHOSPID by DIED			
Hospital Code	**DIED**		**Total**	**Hospital Code**	**DIED**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**			0	1	
3	884 97.46 2.91	23 2.54 3.48	907	**8**	938 100.00 3.09	0 0.00 0.00	938
4	7652 97.76 25.21	175 2.24 26.48	7827	**9**	5370 98.62 17.69	75 1.38 11.35	5445
5	2369 97.89 7.80	51 2.11 7.72	2420	10	**10**	2172 99.63 7.16	8 0.37 1.21

Table 9 gives the proportion of patients with septicemia by hospital.

Table 9. Patients with Septicemia by Hospital

Table of DSHOSPID by septicemia				Table of DSHOSPID by septicemia			
Hospital Code	**septicemia**		**Total**	**Hospital Code**	**Septicemia**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**			0	1	
1	2782 96.66 9.24	96 3.34 10.49	2878	**6**	5035 93.10 16.73	373 6.90 40.77	5408
2	1444 95.88 4.80	62 4.12 6.78	1506	**7**	1498 99.53 4.98	7 0.47 0.77	1505
3	892 98.35 2.96	15 1.65 1.64	907	**8**	938 100.00 3.12	0 0.00 0.00	938
4	7628 97.46 25.34	199 2.54 21.75	7827	**9**	5360 98.44 17.81	85 1.56 9.29	5445
5	2347 96.98 7.80	73 3.02 7.98	2420	**10**	2175 99.77 7.23	5 0.23 0.55	2180

Note that hospital #8 with zero deaths also has zero patients with septicemia. Hospital #6 with the highest death rate has almost 7% patients with septicemia, which is the highest of the ten hospitals. This hospital should probably be investigated to determine whether this high rate of septicemia is a result of nosocomial infection, or whether patients enter the hospital with it. Table 10 gives the rate of pneumonia. Does this hospital take in sicker patients compared to the other nine hospitals?

Table 10. Patients with Pneumonia by Hospital

Table of DSHOSPID by septicemia				Table of DSHOSPID by septicemia			
Hospital Code	**septicemia**		**Total**	**Hospital Code**	**Septicemia**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**			0	1	
1	2638 91.66 9.14	240 8.34 11.23	2878	**6**	4802 88.79 16.63	606 11.21 28.34	5408
2	1382 91.77 4.79	124 8.23 5.80	1506	**7**	1416 94.09 4.90	89 5.91 4.16	1505
3	830 91.51 2.87	77 8.49 3.60	907	**8**	932 99.36 3.23	6 0.64 0.28	938
4	7416 94.75 25.68	411 5.25 19.22	7827	**9**	5149 94.56 17.83	296 5.44 13.84	5445
5	2273 93.93 7.87	147 6.07 6.88	2420	**10**	2038 93.49 7.06	142 6.51 6.64	2180

Hospital #6 again has the highest rate of pneumonia to go with the highest death rate; hospital #8 has the lowest rate of pneumonia; in fact, it is the only hospital with a rate of less than 1%. Table 11 gives the rate for immune disorder. The trend is similar; hospital #8 has the lowest rate, hospital #6 has the highest.

Table 11. Patients with Immune Disorder by Hospital

Table of DSHOSPID by septicemia				Table of DSHOSPID by septicemia			
Hospital Code	**septicemia**		**Total**	**Hospital Code**	**Septicemia**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**			0	1	
1	2198 76.37 8.84	680 23.63 11.05	2878	**6**	3599 66.55 14.48	1809 33.45 29.40	5408
2	1164 77.29 4.68	342 22.71 5.56	1506	**7**	1362 90.50 5.48	143 9.50 2.32	1505
3	638 70.34 2.57	269 29.66 4.37	907	**8**	878 93.60 3.53	60 6.40 0.97	938
4	6324 80.80 25.44	1503 19.20 24.42	7827	**9**	4868 89.40 19.58	577 10.60 9.38	5445
5	1928 79.67 7.76	492 20.33 7.99	2420	**10**	1901 87.20 7.65	279 12.80 4.53	2180

These three tables suggest that hospital #6 has a very good reason to have a higher mortality rate. For this reason, we compare the expected mortality to the actual mortality. We use a predictive model with hospital, septicemia, immune disorder, and pneumonia as the input variables and mortality as the output variable. Figure 17 gives the results, indicating that Dmine regression gives the best fit.

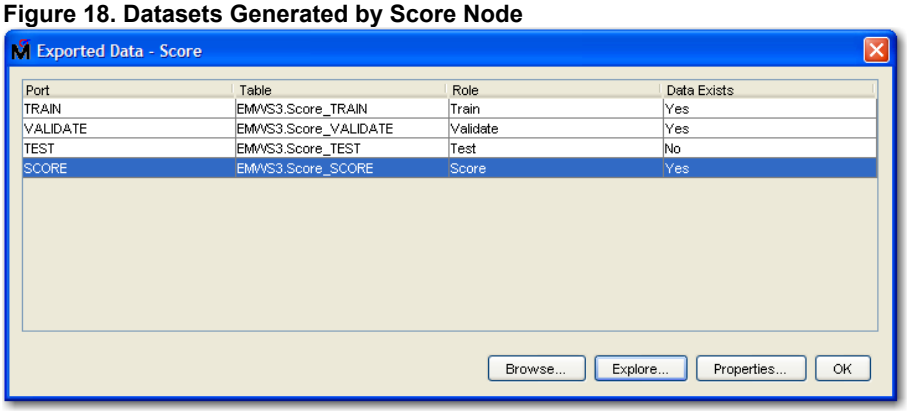
Figure 17. Model Comparison to Predict Mortality

Fit Statistics
Model selection based on _TMISC_

Selected Model	Model Node	Test: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Test: Average Squared Error
	AutoNeural	0.41206	0.24282	0.26105	0.25049
	DMNeural	0.29899	0.20388	0.21929	0.18838
Y	DmineReg	0.29397	0.20973	0.22101	0.19603
	MBR	0.41206	0.33253	0.27731	0.27399
	Neural	0.30653	0.20294	0.21838	0.19115
	Reg	0.30402	0.20854	0.22038	0.19452
	Rule	0.29899	.	.	.
	Tree	0.29899	0.21273	0.22793	0.20109

The best misclassification rate is still almost 30%. We partition the data to define the model; we then score the entire dataset so that we can examine the difference between the predicted and actual values. Figure 18 shows the datasets generated by the score node in Enterprise Miner.

Figure 18. Datasets Generated by Score Node



The dataset EMWS3.Score_Score contains the predicted values as well as the actual values. We can use PROC FREQ in SAS to examine the relationship to hospital. Table 12 gives the actual and predicted values by hospital.

Table 12. Actual Versus Predicted Mortality Values by Hospital

Hospital	Actual Mortality	Predicted Mortality	Difference
1	2.88	30.06	27.18
2	3.05	29.68	26.63
3	2.54	35.06	32.52
4	2.24	23.89	21.65
5	2.11	25.95	23.84
6	3.16	39.87	36.71
7	1.93	14.49	12.56
8	0	0	0
9	1.38	15.76	14.38
10	0.37	0.23	-0.14

By this process, hospitals #8 and #10 have the smallest differential between the actual and predicted values. Therefore, they would be ranked the lowest even though they both have very low mortality values. In contrast, hospital #6 with the highest actual mortality would rank the highest because the difference between the actual and predicted mortality rates is the greatest. However, a hospital with 0 actual mortality has very little room for an increase in the predicted mortality; a hospital with a higher actual mortality is more likely to "game" the system by increasing the predicted mortality.

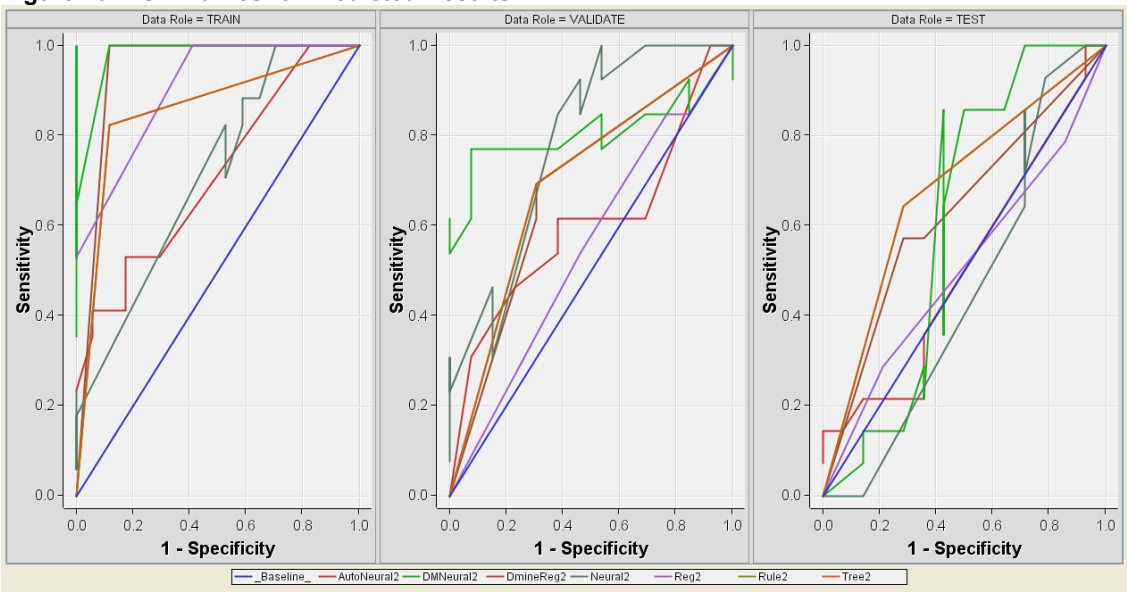
Our second example is restricted to the treatment of patients with a primary diagnosis of COPD. We use a predictive model that is similar to that in the previous section, but now we add a hospital identifier. The results are given in Figure 19 with the ROC curve in Figure 20. Note that the minimum error rate is still 32%.

Figure 19. Mortality Prediction for Patients with COPD

Fit Statistics
Model selection based on _TMISC_

Selected Model	Model Node	Test: Misclassification Rate
	AutoNeural2	0.50000
	DMNeural2	0.42857
	DmineReg2	0.39286
	Neural2	0.53571
	Reg2	0.42857
Y	Rule2	0.32143
	Tree2	0.32143

Figure 20. ROC Curves for Predicted Results



The ROC curves indicate that accuracy decreases considerably on the test data compared to the training data. Table 13 gives the actual and predicted mortality levels by hospital.

Table 13. Actual Versus Predicted Mortality Values by Hospital for Patients with COPD

Hospital	Actual Mortality	Predicted Mortality	Difference
1	3.94	23.62	19.68
2	7.94	33.33	25.39
3	2.13	44.68	42.55
4	4.78	29.57	24.79
5	4.65	29.07	24.42
6	3.21	27.98	24.77
7	3.85	50.00	46.15
8	No COPD Patients		
9	4.11	43.84	39.73
10	0	0	0

The provider that has the largest difference between actual and predicted mortality is #7. The overall ranking is 1>3>9>2>4>6>5>1>10; again, a hospital with zero mortality is penalized using this system. Usually, zero mortality would be considered good. In fact, regardless of the actual mortality, a hospital with zero predicted morality will rank low in comparison to other providers.

In a third example, we will restrict attention to ten hospitals, and examine patients undergoing just one procedure, that of cardiovascular bypass surgery. We will compare actual mortality rates across these hospitals, and look at the relationship of patient diagnosis to prediction of mortality.

We will use a different set of hospitals from the ones in the COPD example since not all of those hospitals perform bypass surgery. Then we will examine the ranking that the model gives to the hospitals. Cardiovascular bypass (or CABG) is assigned an ICD9 procedure code of 36.1. We will restrict attention to patients for whom 36.1 is the primary procedure. In this example, we will use the list of patient conditions as given in Table 4 to define a patient severity level. We will use a stratified sample to define the predicted value of mortality. Then, we will compare the predicted results to the actual results by patient and by hospital. Note that the list in Table 4 contains a condition for congestive heart failure and for myocardial infarction. However, it does not include a code for congested arteries.

This example differs from the previous example because the patient condition will be considered in defining the predicted value. Figure 21 gives the results of the predictive model, with hospital included as one of the input variables. The best misclassification rate is 26%.

Figure 21. Results of Predictive Model

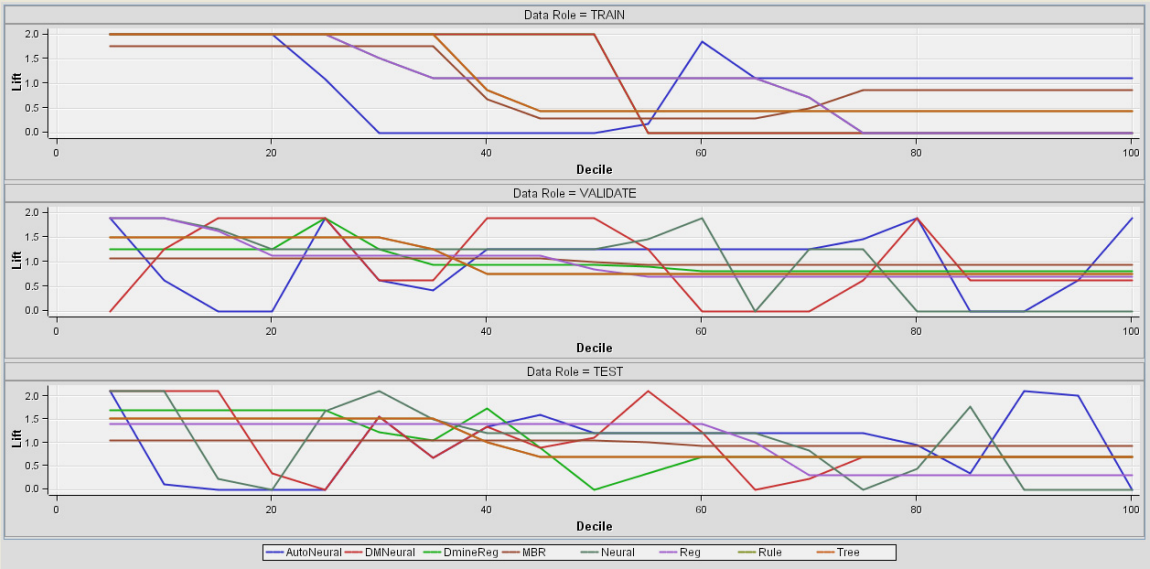
Fit Statistics

Model selection based on _TMISC_

Selected Model	Model Node	Test: Misclassification Rate
	AutoNeural	0.63158
	DMNeural	0.36842
Y	DmineReg	0.26316
	MBR	0.47368
	Neural	0.36842
	Reg	0.31579
	Rule	0.31579
	Tree	0.31579

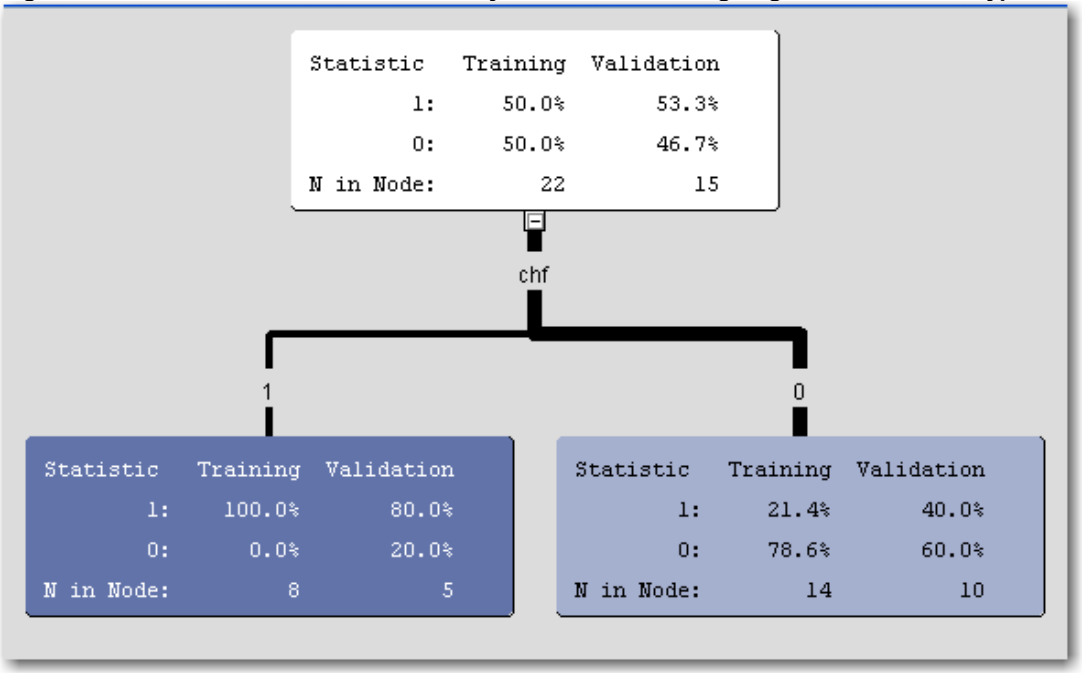
The lift function indicates that the top half of the data can be predicted easily (Figure 22). However, the test data is much less predictable compared to the training set.

Figure 22. Lift Function for Predictive Model



The decision tree indicates that prediction is based largely upon the occurrence of congestive heart failure in the model (Figure 23). For this reason, a provider that can increase the proportion of patients identified as having congestive heart failure will rank higher compared to those who do not inflate the proportion. This condition is loosely defined as a disease that weakens the heart muscle, or weakens the ability of the heart to pump. The definition tends to be vague, and the condition can be assigned differently by different providers.

Figure 23. Decision Tree to Predict Mortality for Patients Undergoing Cardiovascular Bypass Surgery



We compare the difference between the actual and predicted mortality by hospital (Table 16). Table 17 gives the actual and predicted values by procedure. We translate these procedures in Table 14; Table 15 gives the number of procedures by hospital. In both cases, the total sample size is restricted to the ten hospitals with the given procedures.

Table 14. Procedures Related to Cardiovascular Surgery

Procedure	Translation	Frequency	Percent
3610	Aortocoronary bypass for heart revascularization, not otherwise specified	8	0.02
3611	(Aorto)coronary bypass of one coronary artery	5112	11.14
3612	(Aorto)coronary bypass of two coronary arteries	13449	29.30
3613	(Aorto)coronary bypass of three coronary arteries	13176	28.70
3614	(Aorto)coronary bypass of four or more coronary arteries	7208	15.70
3615	Single internal mammary-coronary artery bypass	6419	13.98
3616	Double internal mammary-coronary artery bypass	508	1.11
3617	Abdominal - coronary artery bypass	3	0.01
3619	Other bypass anastomosis for heart revascularization	24	0.05

Table 15 shows the relationship of hospital to procedure. It shows that there is a considerable difference in the procedures performed across the hospitals. For example, #1 has over 50% in 3615, Single internal mammary-coronary artery bypass. The r5emaining hospitals are more divided in their procedures. Hospital #4 has almost 30% in 3614, (Aorto)coronary bypass of four or more coronary arteries, suggesting that it treats patients with very severe blockage in the coronary vessels. The same hospital has approximately 20% of its procedures in 3611, 3612, and 3613.

Table 15. Procedures by Hospital

Table of DSHOSPID by PR1								
HOSPID	**PR1(Principal procedure)**							**Total**
Frequency Row Pct Col Pct	**3611**	**3612**	**3613**	**3614**	**3615**	**3616**	**3619**	
1	16 4.92 14.81	37 11.38 12.85	47 14.46 11.69	24 7.38 9.09	193 59.38 51.47	8 2.46 28.57	0 0.00 0.00	325
2	3 27.27 2.78	4 36.36 1.39	1 9.09 0.25	0 0.00 0.00	2 18.18 0.53	1 9.09 3.57	0 0.00 0.00	11
3	5 4.46 4.63	10 8.93 3.47	19 16.96 4.73	7 6.25 2.65	69 61.61 18.40	2 1.79 7.14	0 0.00 0.00	112
4	22 8.70 20.37	59 23.32 20.49	82 32.41 20.40	75 29.64 28.41	15 5.93 4.00	0 0.00 0.00	0 0.00 0.00	253
5	2 2.27 1.85	14 15.91 4.86	34 38.64 8.46	29 32.95 10.98	9 10.23 2.40	0 0.00 0.00	0 0.00 0.00	88
6	5 2.86 4.63	23 13.14 7.99	51 29.14 12.69	36 20.57 13.64	58 33.14 15.47	2 1.14 7.14	0 0.00 0.00	175
7	21 7.42 19.44	79 27.92 27.43	95 33.57 23.63	60 21.20 22.73	15 5.30 4.00	12 4.24 42.86	1 0.35 100.00	283

Table of DSHOSPID by PR1								
HOSPID	**PR1(Principal procedure)**							**Total**
Frequency Row Pct Col Pct	**3611**	**3612**	**3613**	**3614**	**3615**	**3616**	**3619**	
8	7 15.56 6.48	11 24.44 3.82	17 37.78 4.23	9 20.00 3.41	1 2.22 0.27	0 0.00 0.00	0 0.00 0.00	45
9	15 19.23 13.89	23 29.49 7.99	23 29.49 5.72	7 8.97 2.65	9 11.54 2.40	1 1.28 3.57	0 0.00 0.00	78
10	12 12.50 11.11	28 29.17 9.72	33 34.38 8.21	17 17.71 6.44	4 4.17 1.07	2 2.08 7.14	0 0.00 0.00	96
Total	108	288	402	264	375	28	1	1466

Table 15 shows that there is a considerable difference in the procedures performed across the hospitals. For example, #1 has over 50% in 3615, Single internal mammary-coronary artery bypass. The remaining hospitals are more divided in their procedures. Hospital #4 has almost 30% in 3614, (Aorto)coronary bypass of four or more coronary arteries, suggesting that it treats patients with very severe blockage in the coronary vessels. The same hospital has approximately 20% of the procedures in 3611, 3612, and 3613.

Table 16. Percent of Predicted Versus Actual Mortality by Procedure

Procedure	Actual Mortality	Predicted Mortality
3610	2.78	42.59
3611	2.78	42.36
3612	2.49	39.55
3613	1.14	39.02
3614	1.07	40.53
3615	0	32.14
3616	0	32.14
3617	0	0
3619	0	100

Table 16. Predicted Versus Actual Mortality by Hospital (Given as Percent)

Hospital	Actual Mortality	Predicted Mortality
1	1.23	15.08
2	0	0
3	4.46	25.89
4	0.79	11.86
5	1.14	15.91
6	0.57	16.00
7	3.53	26.15

Hospital	Actual Mortality	Predicted Mortality
8	2.22	20.00
9	1.28	15.38
10	3.13	16,67

Note that the difference between the predicted mortality and the actual value is considerable, both by procedure and by hospital. Therefore, the ability of this model to rank hospitals is highly questionable. There is a large difference between the average and the maximum values of outcomes (Table 18). In particular, at least one patient stayed 150 days or more for 3611, 3612, 3613, and 3614. How should these outliers be considered when ranking quality? We use the following code to find the kernel density estimation functions (Figures 24 and 25).

```
proc sort data=nis.cardiovascular out=work.cardiovascular2;
by pr1;
proc kde data=work.cardiovascular2;
univar los/gridl=0 gridu=15 out=nis.kdecardlos;
univar totchg/gridl=20000 gridu=100000 out=nis.kdecardchg bwm=.9;
by pr1;
run;
```

Table 18. Length of Stay and Total Charges by Procedure

Principal procedure	N Obs	Variable	Mean	Std Dev	Minimum	Maximum
3610	8	TOTCHG LOS	65496.75 5.8750000	20519.46 1.3562027	39277.00 3.0000000	99584.00 7.0000000
3611	5112	TOTCHG LOS	90656.52 8.9047340	68257.93 7.5870299	84.0000000 0	829195.00 161.0000000
3612	13449	TOTCHG LOS	96585.12 9.3835973	73855.08 7.4754109	534.0000000 0	997836.00 153.0000000
3613	13176	TOTCHG LOS	101269.45 9.5980571	75537.11 7.3233339	2029.00 0	998991.00 188.0000000
3614	7208	TOTCHG LOS	103371.69 9.6594062	74343.53 7.4100765	839.0000000 0	918286.00 155.0000000
3615	6419	TOTCHG LOS	92813.63 8.7963857	66991.61 6.6919820	484.0000000 0	898653.00 114.0000000
3616	508	TOTCHG LOS	89716.19 8.1909449	56349.01 6.1948121	20786.00 1.0000000	461205.00 78.0000000
3617	3	TOTCHG LOS	78057.33 6.3333333	56807.84 2.3094011	43820.00 5.0000000	143632.00 9.0000000
3619	24	TOTCHG LOS	88172.17 8.0833333	56691.47 7.6437907	20741.00 1.0000000	282273.00 32.0000000

Figure 24 shows the length of stay by hospital. As shown in Table 18, stay differed considerably by procedure. It also differs considerably by hospital. Hospital #6 has the greatest probability of a shorter length of stay compared to the other hospitals. Hospital #2 has the highest probability of a longer length of stay. Hospital #7 tends to be in the middle in probability for both a high and low length of stay, as does hospital #1.

Figure 24. Length of Stay by Hospital for Cardiovascular Surgery

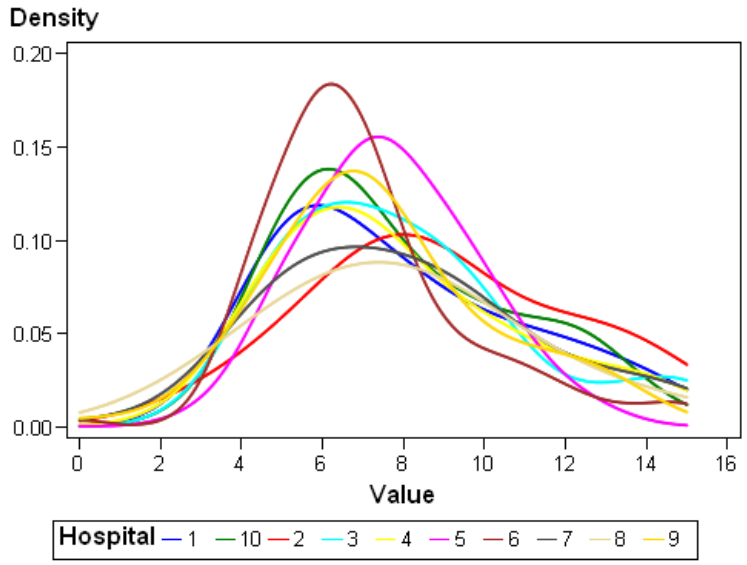
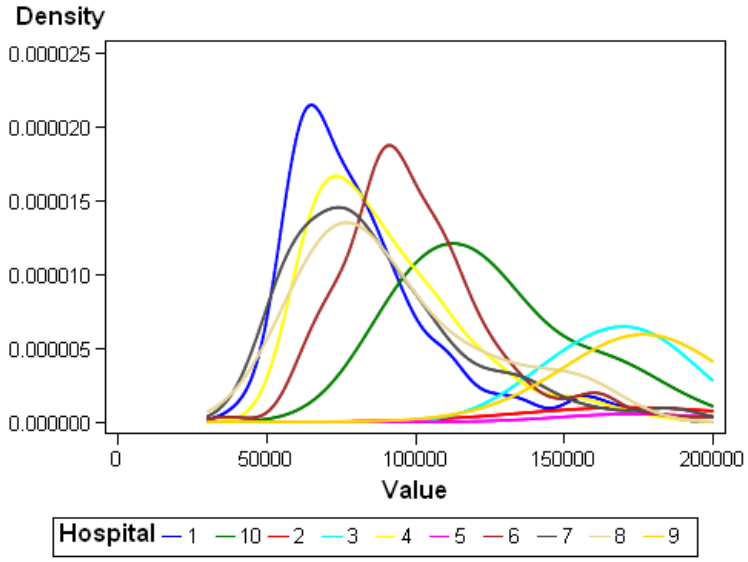


Figure 25 shows the total charges compared to hospital. There is a definite shift in the curves, indicating that some hospitals charge far more compared to other hospitals, especially hospitals #3 and #9. Hospital #1 has the least charges, reinforcing the fact that it more generally performs a procedure that is less risky compared to the other procedures. Interestingly, hospital #7, while performing higher risk procedures, also tends to charge a lower amount.

Figure 25. Total Charges by Hospital for Cardiovascular Surgery



Next, we examine the length of stay by procedure for a specific hospital. We contrast hospital #7 to hospital #6. For hospital #6, there is a natural hierarchy in the kernel density estimators, demonstrating the severity of each of the procedures. Procedure 3613 has the highest probability of a long length of stay; Procedures 3611 and 3616 have the highest probability of a short length of stay. Figure 31 shows the length of stay for hospital #7.

Figure 26. Length of Stay for Hospital #6 by Procedure

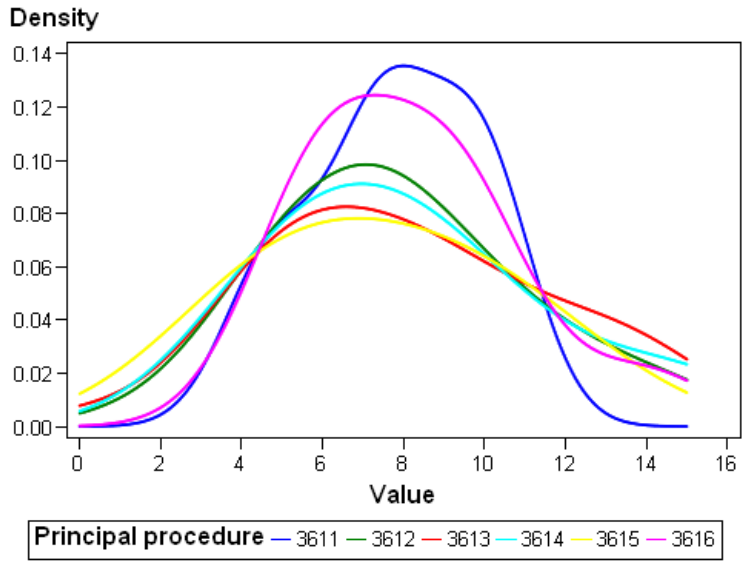
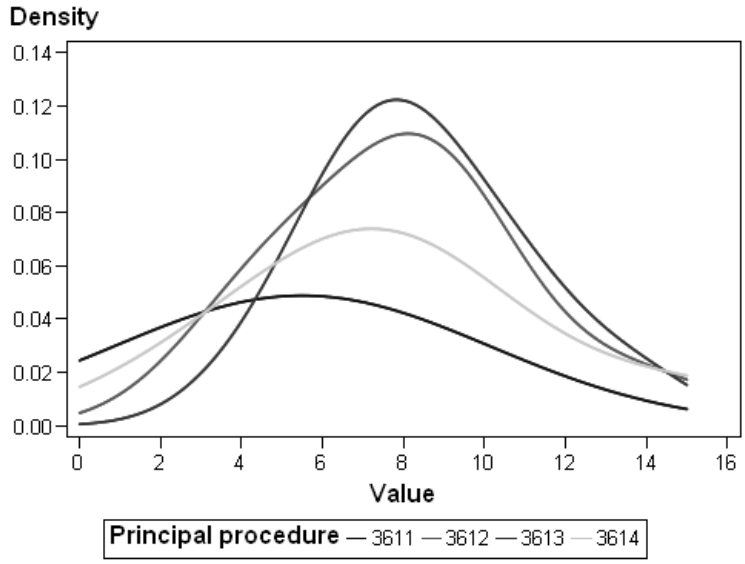
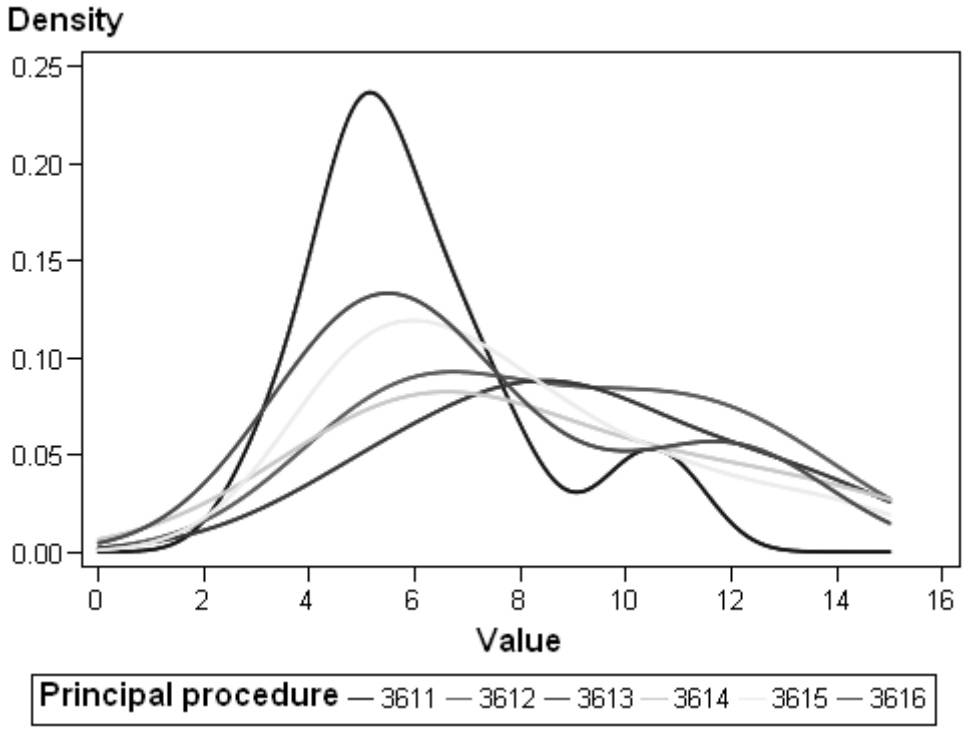


Figure 27. Length of Stay for Hospital #7 by Procedure



First, hospital #6 does only four of the procedures. In contrast to hospital #7, procedure 3611 has the lowest probability of a short length of stay for hospital #6; procedure 3613 has the highest probability of a short length of stay. The ordering of the procedures is completely different for the two hospitals. We do a final examination in Figure 28 for hospital #1.

Figure 28. Length of Stay for Hospital #1 by Procedure



In this figure, it is clear that procedure 3612 has the highest probability of a long length of stay while 3611 has a high probability of a short length of stay. In other words, we have three different hospitals and they have three very different graphs, indicating that there is almost no relationship between procedure and length of stay when comparing the different hospitals.

DISCUSSION

Predictive modeling already includes all regression models. Therefore, it will be used much more often when analyzing health outcomes than it has been used in the past. It needs to be brought more commonly into the curriculum for students specializing in health outcomes research before predictive modeling will become more common. Departments of Biostatistics and Informatics need to recognize the availability of data mining tools and their use in health outcomes research. The process of predictive modeling should be substituted for the now common use of regression models. Misclassification and cost should be used instead of p-values and odds ratios to have more accurate results generally in health outcomes research.

The process of data mining automatically incorporates important components that are not generally a part of more traditional statistical methods. These components include sampling, partitioning, and model comparison. In addition, they include a component for scoring new data and defining a more meaningful definition of a “best” model. Best does not always mean the most accurate. Often in healthcare, we can sacrifice some accuracy in the false positive prediction to greatly reduce the false negative rate. However, as predictive modeling automatically incorporates regression models, the process of predictive modeling is essential to health outcomes research.

In addition, the common practice of using the difference between observed and predicted outcomes to rank the quality of providers should be reconsidered. As it is now, providers with low mortality can be penalized compared to providers with high mortality since the differential between actual and predicted can be much larger for providers with higher adverse outcomes.

REFERENCES

Gamito, E. J., & Crawford, D. E. (2004). Artificial neural networks for predictive modeling in prostate cancer. *Current Oncology Reports, 6*(3), 216-221.

Hodgman, S. B. (2008). Predictive modeilng & outcomes. *Professional Case Management, 13*(1), 19-23.

Powers, C. A., Meyer, C. M., Roebuck, M. C., & Vaziri, B. (2005). Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Medical Care, 43*(11), 1065-1072.
Sylvia, M. L., Shadmi, E., Hsiao, C.-J., Boyd, C. M., Schuster, A. B., & Boult, C. (2006). Clinical features of high risk older person identified by predictive modeling. *Disease Management, 9*(1), 56-62.
Tewari, A., Porter, C., Peabody, J., Crawford, E., Demers, R., Johnson, C., et al. (2001). Predictive modeling techniques in prostate cancer. *Molecular Urology, 5*(4), 147-152.
Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Current Pharmaceutical Design, 13*(34), 3494-3504.
Weber, C., & Neeser, K. (2006). Using individualized predictive disease modeling to identify patients with the potential to benefit from a disease management program for diabetes mellitus. *Disease Management, 9*(4), 242-256.
Whitlock, T., & Johnston, K. (2006). Using predictive modeling to evaluate the financial effect of disease management. *Managed Care Interface, 19*(9), 29-34.

AUTHOR CONTACT

Patricia Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6010
pcerrito@gmail.com