Paper 212-2008

Data Mining to Predict the Occurrence of Resistant Infection

Patricia Cerrito, University of Louisville, Louisville, KY

ABSTRACT

In order to demonstrate the use of predictive modeling in SAS Enterprise Miner, We will examine the problem of resistant infection in the hospital using data from the National Inpatient Sample. The National Inpatient Sample contains approximately 8 million records and represents 37 states. A sample of 800,000 inpatient events or about 10% of the records will be used to investigate the problem. Of 800,000 events, 6000 or 0.75% of the records had a diagnosis of resistant infection. Using data visualization techniques and predictive modeling as well as SAS Text Miner, we found the infection treated with an infusion of antibiotics in only 197 patients, with an infusion of linezolid in 95 patients, indicating the use of antibiotics is under-reported, or under-utilized.

Since there are thousands of codes used to define patient condition and treatment procedures, in order to use predictive modeling, these codes must be reduced to a manageable number. SAS Text Miner can be used for this purpose. Predictive modeling demonstrated that, diagnoses are not highly correlated with the occurrence of resistant infection, but treatment procedures are and can be used to predict patient risk. Using these predictions, techniques can be used to reduce the incidence of infection.

INTRODUCTION

Resistant infection is a serious nosocomial problem (meaning that the infection occurred because of the hospital). There are few antibiotics that are effective in treating these infections, and the patient consequences can be significant, including amputation and death. Moreover, these infections add significantly to the cost of patient care and add to the patient length of stay. Medicare has just recently announced that they will no longer reimburse healthcare providers for the added cost of nosocomial infection, meaning that the provider will have to shoulder the complete cost of treating the infection until patient discharge.[1] The most difficult of the resistant infections is MRSA (Methicillin Resistant Staph), and we will use MRSA to represent the general category of resistant infections to see if we can predict its occurrence.

In the past, focus has been on general infection control procedures[2, 3], assuming a uniform patient risk, or on targeting patients using specific demographic and health parameters.[4, 5] In these studies, one patient location (for example, ICU)[6, 7] or one patient treatment was investigated.[8] To date, the infections have not been controlled or prevented very successfully.[9, 10] Attempts to predict the infection have also not been terribly successful.[11] If predictive modeling on a more general basis can be used to identify those patients at high risk for acquiring resistant infection, preventive measures can be used on a targeted group of patients to reduce the likelihood of infection.[12]

METHOD

We use data from the National Inpatient Sample that contain information from all inpatient events for 37 participating states, with 8 million records for the year 2004. More information about the dataset is available at http://www.hcup-us.ahrq.gov/nisoverview.jsp#databases. The data contain some basic patient demographics, fifteen columns concerning patient condition, and another fifteen columns identifying patient treatments. These columns use ICD9 codes detailing health information.[13] There are thousands of possible codes that can be listed. We want to see if we can use treatment choices to find those with higher risk of infection compared to other treatments.

Before any predictive modeling is possible, we must compress these diagnosis and procedure clusters in some way that is meaningful. It is not enough to examine the codes; we must also use all combinations of codes, the number of which increases factorially. To do this compression, we use SAS Text Miner. In this way, combinations are linked, and can be compressed into a small number of clusters. We create a series of clusters for patient diagnoses and another series of clusters for patient procedures. Both sets of clusters are then used in the predictive model. We then use the clusters and patient demographics to predict the occurrence of resistant infection. We want to identify those patients at highest risk so that aggressive, preventive treatment can be initiated.

In addition, we want to identify the added cost of MRSA to the hospital to get an estimate of the added costs that the healthcare provider must absorb. The costs are considerable, but also vary across patient treatment and diagnosis clusters.

RESULTS

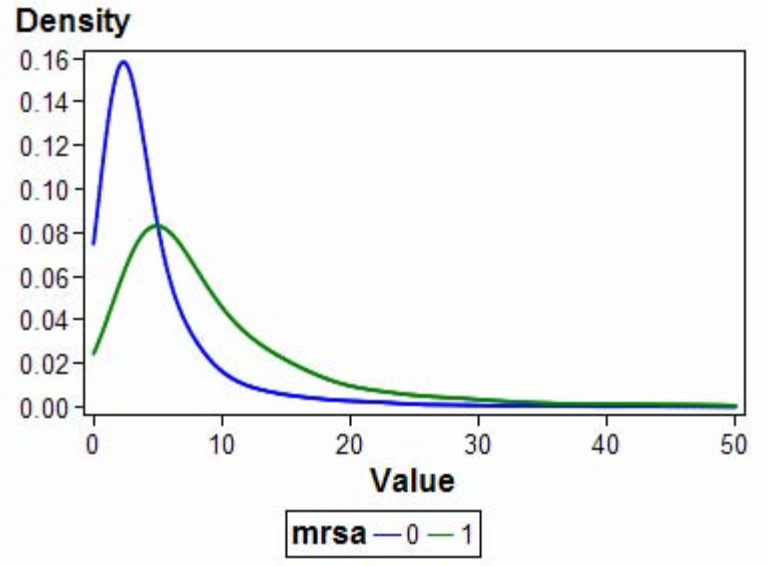
We use a 10% sample of 800,000 patients. Approximately 6000 have a diagnosis of resistant infection. The average length of stay for the patients without resistant infection is 4.59 days ($20,270 average total charges) compared to 9.85 days ($36,979 average total charges) for patients with resistant infection. Approximately 2% of patients without resistant infection died compared to 4% of resistant patients who died.

To investigate the problem of resistant infection across the entire distribution of patients, we use PROC KDE to estimate the population distribution using the code:

```
proc sort data=mrsa.nis out=mrsa.sortednis;
by mrsa;
proc kde data=mrsa.sortednis;
univar los/gridl=0 gridu=50 out=mrsa.kdelos method=os;
univar totchg/gridl=0 gridu=50000 out=mrsa.kdetotchg method=os bwm=.2;
by mrsa ;
run;
```

Figure 1 shows the difference in hospital length of stay and Figure 2 shows the difference in total charges. The two graphs show that patients with MRSA have a higher probability of longer stay and higher charges compared to patients without MRSA. However, this representation is incomplete since it could also be that patients with more severe problems are the ones who are more likely to acquire MRSA. We need to consider patient risk.

Figure 1. Process Flow for Length of Stay Between Patients With and Without MRSA.

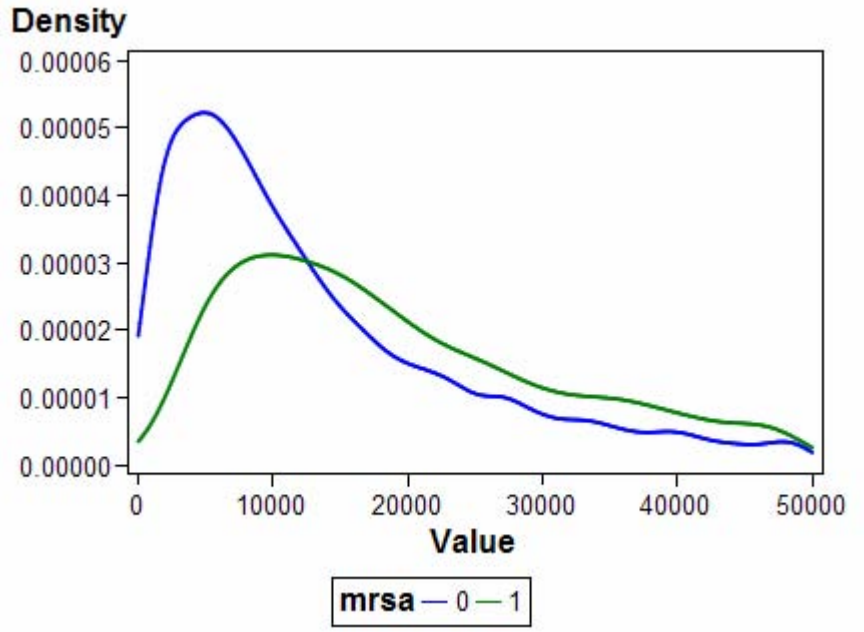


To consider the patient diagnoses and treatment procedures, we first use the CATX procedure in a data step to combine the 15 columns into one text string,

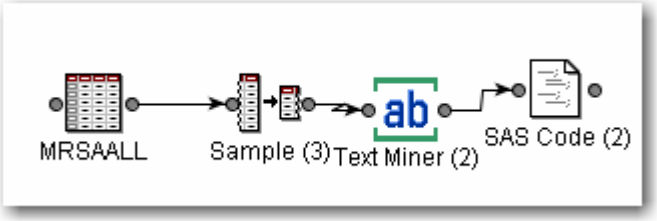
```
CATX(' ',dx1,dx2,dx3,dx4,dx5,dx6,dx7,dx8,dx9,dx10,dx11,dx12,dx13,dx14,dx15);
CATX(' ',pr1,pr2,pr3,pr4,pr5,pr6,pr7,pr8,pr9,pr10,pr11,pr12,pr13,pr14,pr15);
```

Next, we use SAS Text Miner to cluster the text strings so that we can use patient condition in the predictive model (Display 1). Display 2 shows the changes to the defaults to stratify the sample, needed because MRSA remains a rare occurrence. Display 3 shows the changes to the defaults for Text Miner so that the icd9 number codes can be used in the analysis.

Figure 2. Probability Density of Total Charges Between Patients With and Without MRSA



Display 1. Process Flow for SAS Text Miner

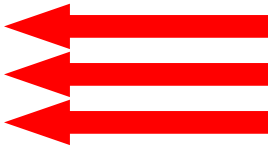


Display 2. Stratified Sampling for Rare Occurrence

Property	Value
Node ID	Smpl3
Imported Data	
Exported Data	
Variables	
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	
Percentage	35.0
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

Display 3. Changes to the Defaults for SAS Text Miner

Property	Value
Node ID	TEXT2
Imported Data	
Exported Data	
Variables	
Interactive	
Rerun	No
Parse	
Parse Variable	Diagnoses
Language	ENGLISH
Stop List	
Start List	EMWS.startmrsa
Stem Terms	Yes
Terms in Single Document	No
Punctuation	No
Numbers	Yes
Different Parts of Speech	No
Ignore Parts of Speech	
Noun Groups	No
Synonyms	SASHELP.ENGSYNMS
Find Entities	No
Types of Entities	



Display 4 shows the 31 clusters defined for the procedure codes; similarly, a total of 30 clusters were defined for the diagnoses codes. The codes in the clusters in Display 4 can be identified through the website, http://icd9cm.chrisendres.com/index.php. We consider some of the clusters below:

For cluster 1:
39.95 Hemodialysis
40.0 Incision of lymphatic structures
39.43 Removal of arteriovenous shunt for renal dialysis
38.95 Venous catheterization for renal dialysis
39.42 Revision of arteriovenous shunt for renal dialysis

For cluster 22:
39.93 Insertion of vessel-to-vessel cannula
80.82 Other local excision or destruction of lesion of joint (elbow)
54.3 Excision or destruction of lesion or tissue of abdominal wall or umbilicus
85.0 Mastotomy
82.01 Exploration of tendon sheath of hand

Both of these clusters (as shown in Table 1) have a high incidence rate of MRSA. Contrast these with cluster 2 with a low incidence rate of MRSA:

81.02 Other cervical fusion, anterior technique
84.51 Insertion of interbody spinal fusion device
81.62 Fusion or refusion of 2-3 vertebrae
81.08 Lumbar and lumbosacral fusion, posterior technique
81.51 Total hip replacement

Display 4. Text Clusters of ICD9 Procedure Codes

#	Descriptive Terms	Freq
1	3995, 400, 3943, 3895, 3942, 9749, 3927, 8609, 3993, 9782	203
2	8102, 8451, 8162, 8108, 8051, 7779, 8103, 8163, 8452, 8101	45
3	9389, 9959, 9324, 8721, 8603, 8591, 6952, 6497, 5899, 5305	4290
4	8191, 8026, 806, 8183, 7781, 8016, 8086, 835, 8303, 8076	144
5	8872, 4223, 8944, 3799, 8111, 3995, 3895, 9962, 3943, 9214	135
6	741, 7534, 0395, 9948, 6632, 6639, 7537, 7532, 7279, 7535	171
7	7936, 3327, 2219, 3324, 8611, 9604, 9605, 9656, 9672, 3322	150
8	6561, 5459, 6859, 6829, 6839, 7092, 6549, 7050, 5979, 684	170
9	8774, 6851, 598, 6029, 560, 5631, 9762, 613, 5304, 610	144
10	3722, 8856, 8853, 3794, 8872, 3612, 3613, 3614, 3615, 3723	192
11	8876, 5122, 5491, 8838, 5123, 5185, 4701, 8753, 8801, 5361	261
12	9468, 863, 8674, 8628, 8659, 8622, 7788, 8604, 7789, 8669	715
13	0309, 8161, 9354, 7779, 9902, 8163, 9339, 8105, 8108, 8051	101
14	9920, 8853, 3722, 8857, 3607, 3601, 8856, 3605, 3606, 3723	192
15	966, 9702, 9999, 9703, 4311, 9672, 4513, 311, 0125, 9390	157
16	4523, 4513, 4292, 4542, 4525, 4341, 4524, 4443, 4573, 9702	117
17	9919, 9918, 9917, 9929, 9923, 8952, 9396, 8954, 9394, 9922	163
18	7562, 7301, 7271, 734, 7359, 7309, 6902, 7569, 736, 7534	437
19	9907, 8605, 9390, invl, 9465, 8417, 8609, 843, 3995, 3895	600
20	734, 741, 0391, 7309, 9649, 7534, 0390, 7532, 7535, 7561	91
21	9547, 5421, 9983, 3892, 4576, 4709, 5459, 9915, 3891, 4562	357
22	3893, 8082, 543, 850, 8201	475
23	9671, 9672, 9604, 3891, 311, 3893, 9960, 3322, 3323, 4311	602
24	8627, 9427, 3404, 7935, 9462, 3783, 3772, 3491, 3421, 270	323
25	8604, 8319, 8622, 3893, 7863, 8666, 8401, 7867, 8689, 8201	352
26	4542, 4422, 4523, 4516, 4525, 4443, 4836, 4341, 4292, 4824	168
27	3950, 3929, 3990, 8849, 8848, 3812, 8842, 3818, 3949, 8847	128
28	9546, 2591, 9955, 640, 9547, 9541	251
29	9905, 9925, 8605, 8891, 8607, 0331, 8703, 8841, 4131, 8543	325
30	9375, 9444, 9381, 9439, 9425, 9438, 9423, 9408, 9401, 9339	102
31	8604, 211, 7109	387

We next examine the relationship of procedure cluster to the occurrence of MRSA using the stratified sample with 50% MRSA. We first examine the relationship of procedure cluster to the occurrence of resistant infection (Table 1). While the stratified sampling introduces a bias into the table, the table does indicate which clusters are more likely to have the MRSA.

Table 1. Procedure Cluster by Percentage of MRSA

Procedure Cluster	MRSA		Total	Procedure Cluster	MRSA		Total
	0	1			0	1	
1	56 27.59 0.94	147 72.41 2.46	203	17	100 61.35 1.67	63 38.65 1.05	163
2	43 95.56 0.72	2 4.44 0.03	45	18	433 99.08 7.25	4 0.92 0.07	437
3	2471 57.60 41.36	1819 42.40 30.45	4290	19	200 33.33 3.35	400 66.67 6.70	600

Procedure Cluster	MRSA		Total	Procedure Cluster	MRSA		Total
	0	1			0	1	
4	30 20.83 0.50	114 79.17 1.91	144	20	87 95.60 1.46	4 4.40 0.07	91
5	42 31.11 0.70	93 68.89 1.56	135	21	231 64.71 3.87	126 35.29 2.11	357
6	166 97.08 2.78	5 2.92 0.08	171	22	37 7.79 0.62	438 92.21 7.33	475
7	40 26.67 0.67	110 73.33 1.84	150	23	119 19.77 1.99	483 80.23 8.09	602
8	164 96.47 2.75	6 3.53 0.10	170	24	178 55.11 2.98	145 44.89 2.43	323
9	91 63.19 1.52	53 36.81 0.89	144	25	17 4.83 0.28	335 95.17 5.61	352
10	140 72.92 2.34	52 27.08 0.87	192	26	113 67.26 1.89	55 32.74 0.92	168
11	177 67.82 2.96	84 32.18 1.41	261	27	75 58.59 1.26	53 41.41 0.89	128
12	115 16.08 1.93	600 83.92 10.04	715	28	248 98.80 4.15	3 1.20 0.05	251
13	55 54.46 0.92	46 45.54 0.77	101	29	183 56.31 3.06	142 43.69 2.38	325
14	185 96.35 3.10	7 3.65 0.12	192	30	55 53.92 0.92	47 46.08 0.79	102
15	35 22.29 0.59	122 77.71 2.04	157	31	5 1.29 0.08	382 98.71 6.39	387
16	83 70.94 1.39	34 29.06 0.57	117	Total	5974	5974	11948

Another cluster with a high proportion of MRSA is 31:

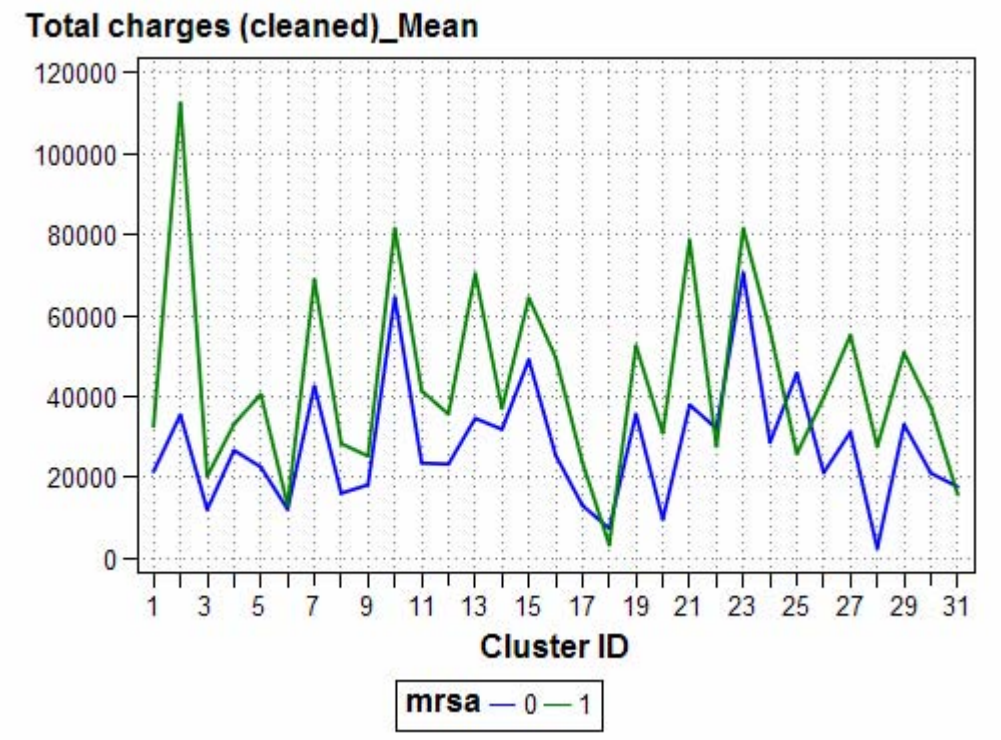
86.04 Other incision with drainage of skin and subcutaneous tissue
21.1 Incision of nose
71.09 Other incision of vulva and perineum

Similarly, cluster 25 starts in the same way, with procedures related to wounds:

86.04 Other incision with drainage of skin and subcutaneous tissue
83.19 Other division of soft tissue
86.22 Excisional debridement of wound, infection, or burn
38.93 Venous catheterization, not elsewhere classified
78.6 Removal of implanted devices from bone
86.66 Homograft to skin

These two clusters clearly show that incisions increase the risk of MRSA. Figure 3 shows the total charges for the procedure clusters; Figure 4 shows the length of stay.

Figure 3. Total Charges by Procedure Cluster



Although cluster 2 has a low rate of MRSA, the cost differential is considerable, adding more than 15 days to the length of stay and almost $100,000 additional cost per patient. In contrast, procedure cluster 6 adds almost no additional cost or stay for MRSA. Procedures in cluster 6 are all related to childbirth. Similarly, cluster 22 with a very high rate of MRSA adds very little additional length of stay but no additional cost. Therefore, added cost of prevention provides few financial benefits.

Figure 5 gives the average added cost for MRSA*frequency of occurrence for each cluster. It shows that cluster 3 with a moderate rate of occurrence adds substantially to the cost. Therefore, patients in this cluster should probably be treated as if they are high risk. It is primarily related to prosthesis.

Figure 4. Length of Stay by Procedure Clusters

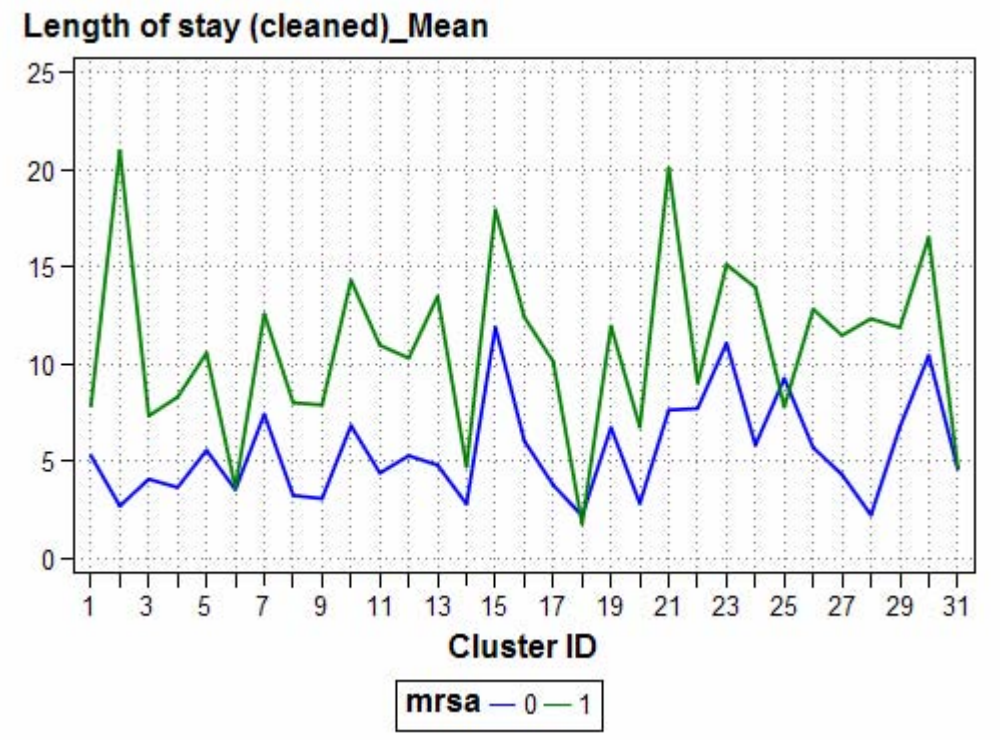
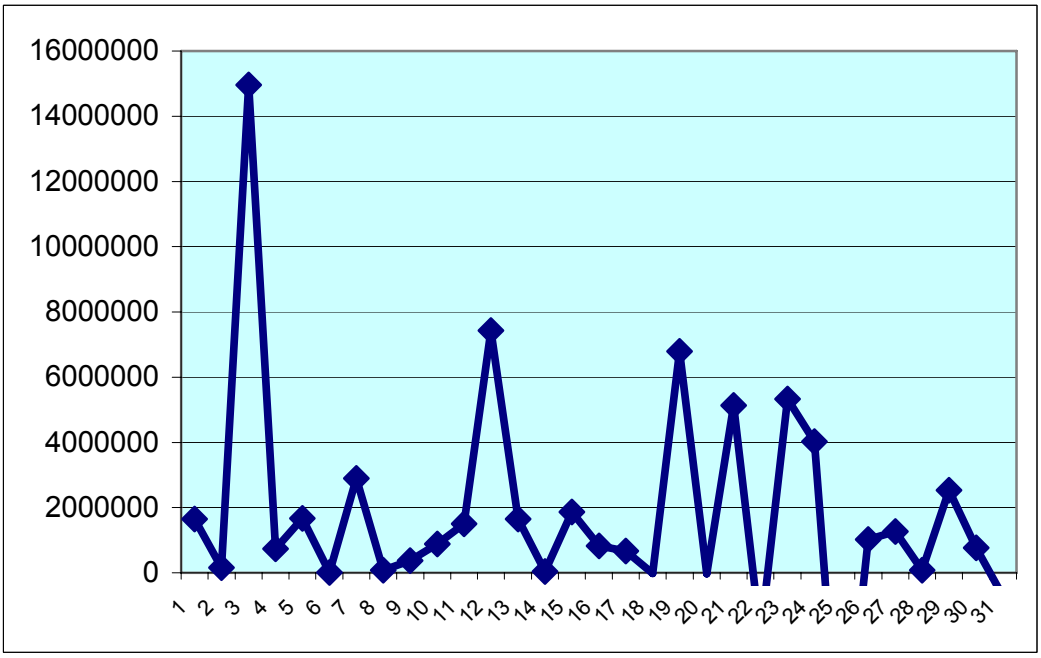


Figure 5. Total Cost of MRSA by Procedure



Similarly, we examine patient diagnoses. Display 5 shows the 32 diagnosis clusters.

Display 5. Diagnosis Clusters.

#	Descriptive Terms	Freq
2	2720, 2724, 4019, 78659, 60000, 25000, 42789, 73300, 4240, 53081	370
3	64821, 27801, 2859, 65421, 66331, 64841, 65971, 64891, 66411, 66401	307
4	2768, 2752, 78791, 2765, 3051, 2859, 5849, 4019, 51881, 5990	310
5	49390, 30183, 5409, 53081, 27801, 6820, 6822, 311, 30000, 04111	293
6	44023, 70715, 25070, 25082, 25080, 7318, 04111, 25081, 6827, 73007	214
7	04111, 683, 6825, 6164, 6822, 6084, 68100, 6820, 042, 7048	416
8	v451, 29410, 42731, 99662, 6826, 40391, 5939, 3310, 7907, 4280	965
9	04111, 5672, 99859, 99832, 6822, 99812, 04119, 5601, 0417, 27801	270
10	27800, 4019, 78057, 2724, 25000, 6822, 53081, 3051, 49390, 6826	260
11	59654, 7070, 9072, 5990, 3441, 0417, 2639, 0389, 03811, 5070	338
12	319, 3439, 78039, v440, v441, 5990, 51881, 5070, 2761, 311	315
13	2724, v4501, 41400, v4581, 412, 4280, v4582, 42731, 4439, 25000	439
14	v3001, 77089, 7706, 77081, 76518, 7793, 7755, 76528, 76519, v3101	323
15	04104, 64403, 5990, 64663, 0414, 0413, 59080, 04111, 0417, 7907	274
16	2639, 71590, 2948, 7872, 5990, 5070, 43820, v1259, v441, 73300	505
17	2761, 2768, 2767, 5990, 49121, 2449, 5849, 51881, 2765, 2859	358
18	3829, 46611, 5589, 2765, 4589, 7906, 2639, 78791, 2875, 42789	218
19	6826, 9165, 8911, 9161, 6806, 04111, 72665, 3051, 9895, 72660	191
20	78552, 99662, 40391, 99673, v451, 99591, 03811, 99592, 0389, 5849	399
21	53081, 311, 5533, 49392, 78650, 3051, 60000, 56400, 4019, 2724	284
22	2449, 4019, 73300, v103, 2724, v1582, 2859, 311, 42731, 53081	328
23	9135, 72633, 6824, 04111, 6823, 3051, 30560, 9895, 30570, 30590	180
24	v053, 7661, 7678, 7626, 7746, v290, 76719, v3000, 76529, v298	398
25	58381, 3371, 25040, 36201, 3572, 25060, 25050, 7318, 70715, 25080	218
26	66401, 65961, 64881, 65811, 65221, 64231, 66551, 66331, 64891, 65421	508
27	42731, 4280, 5070, 51881, 496, 99592, 5849, 5119, 0389, 2767	593
28	4660, 49121, 30000, v1582, 42789, 4280, 3004, 56400, 51881, 42731	399
29	6262, v5789, 2851, v4365, 2800, 56210, 5119, 56400, 5180, 07054	952
30	4439, 4019, 25000, 99762, 6826, 6827, 04111, 68100, 45981, 71536	181
31	v0259, 311, 71590, 30000, 73300, 6822, 27800, 49121, 53081, 3051	292
32	412, 4280, 4111, v4582, 2724, v4501, 4019, 25000, 4139, 41401	515

Diagnosis 6 is related to patients with diabetes who have problems with diabetic foot ulcers:

440.23 atherosclerosis of the extremities with ulceration
707.15 ulcer of toes
250.70 diabetes with peripheral circulatory disorder
250.80 diabetes with other specified manifestations
682.7 cellulitis of foot
730.07 acute osteomyelitis of ankle and foot

Cluster 19 is related to bursitis and wounds:

916.5 insect bite, nonvenomous, infected
891.1 complicated open wounded of knee
916.1 abrasion or friction burn, infected
680.6 carbuncle and furuncle
726.65 prepatellar bursitis
989.5 venom

Cluster 22 is related to severe infection:

785.52 septic shock
996.62 complications of reattached hand
403.91 hypertensive kidney disease
996.73 complications of renal dialysis device
584.9 acute renal failure
036.9 meningococcal infection

Cluster 23 is related to substance abuse:
913.5 insect bite, infected
726.33 olecranon bursitis
682.4 cellulitis of hand
682.3 cellulitis, upper arm
305.1 tobacco use disorder
305.60 cocaine abuse
305.70 amphetamine abuse
305.90 unspecified drug abuse

Table 2 has the percentage of MRSA in the stratified sample. Figures 6 and 7 examine the total cost of MRSA. They show that clusters 8 and 27 add significantly to cost.

Table 2. Percentage of MRSA for Diagnosis Clusters

Diagnosis Clusters	MRSA		Total	Diagnosis Clusters	MRSA		Total
	0	1			0	1	
1	210 62.69 3.52	125 37.31 2.09	335	**17**	141 39.39 2.36	217 60.61 3.63	358
2	314 84.86 5.26	56 15.14 0.94	370	**18**	160 73.39 2.68	58 26.61 0.97	218
3	243 79.15 4.07	64 20.85 1.07	307	**19**	7 3.66 0.12	184 96.34 3.08	191
4	151 48.71 2.53	159 51.29 2.66	310	**20**	20 5.01 0.33	379 94.99 6.34	399
5	173 59.04 2.90	120 40.96 2.01	293	**21**	180 63.38 3.01	104 36.62 1.74	284
6	17 7.94 0.28	197 92.06 3.30	214	**22**	188 57.32 3.15	140 42.68 2.34	328
7	10 2.40 0.17	406 97.60 6.80	416	**23**	8 4.44 0.13	172 95.56 2.88	180
8	276 28.60 4.62	689 71.40 11.53	965	**24**	397 99.75 6.65	1 0.25 0.02	398
9	23 8.52 0.39	247 91.48 4.13	270	**25**	69 31.65 1.16	149 68.35 2.49	218
10	138 53.08 2.31	122 46.92 2.04	260	**26**	506 99.61 8.47	2 0.39 0.03	508

Diagnosis Clusters	MRSA		Total	Diagnosis Clusters	MRSA		Total
Frequency Row Pct Col Pct	0	1			0	1	
11	46 13.61 0.77	292 86.39 4.89	338	**27**	246 41.48 4.12	347 58.52 5.81	593
12	145 46.03 2.43	170 53.97 2.85	315	**28**	195 48.87 3.26	204 51.13 3.41	399
13	219 49.89 3.67	220 50.11 3.68	439	**29**	724 76.05 12.12	228 23.95 3.82	952
14	315 97.52 5.27	8 2.48 0.13	323	**30**	86 47.51 1.44	95 52.49 1.59	181
15	95 34.67 1.59	179 65.33 3.00	274	**31**	90 30.82 1.51	202 69.18 3.38	292
16	196 38.81 3.28	309 61.19 5.17	505	**32**	386 74.95 6.46	129 25.05 2.16	515
				Total	5974	5974	11948

Cluster 14, while having a low rate of MRSA, has the highest cost differential. The diagnoses in this cluster are related to a newborn:

V3001 single liveborn before current hospital admission
V3101 twin and mate liveborn before current hospital admission
770.89 other respiratory problems after birth
770.81 primary apnea of newborn
765.18 other preterm infant, 2000-2499 grams
779.3 feeding problem in newborn

Since a stratified sample was already used to increase the sampling of MRSA, there is no need to re-sample. The prior probabilities were defined to reflect the true probability of occurrence (Display 6). MRSA is identified as a binomial target.

Figure 6. Charges by Diagnosis

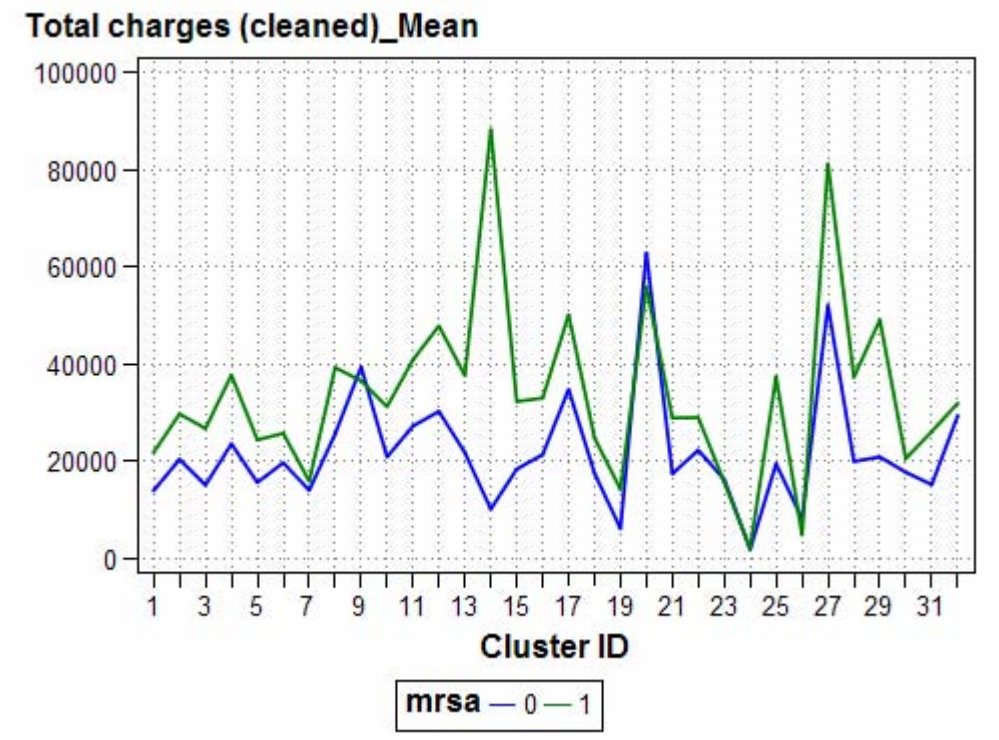


Figure 7. Length of Stay by Diagnosis

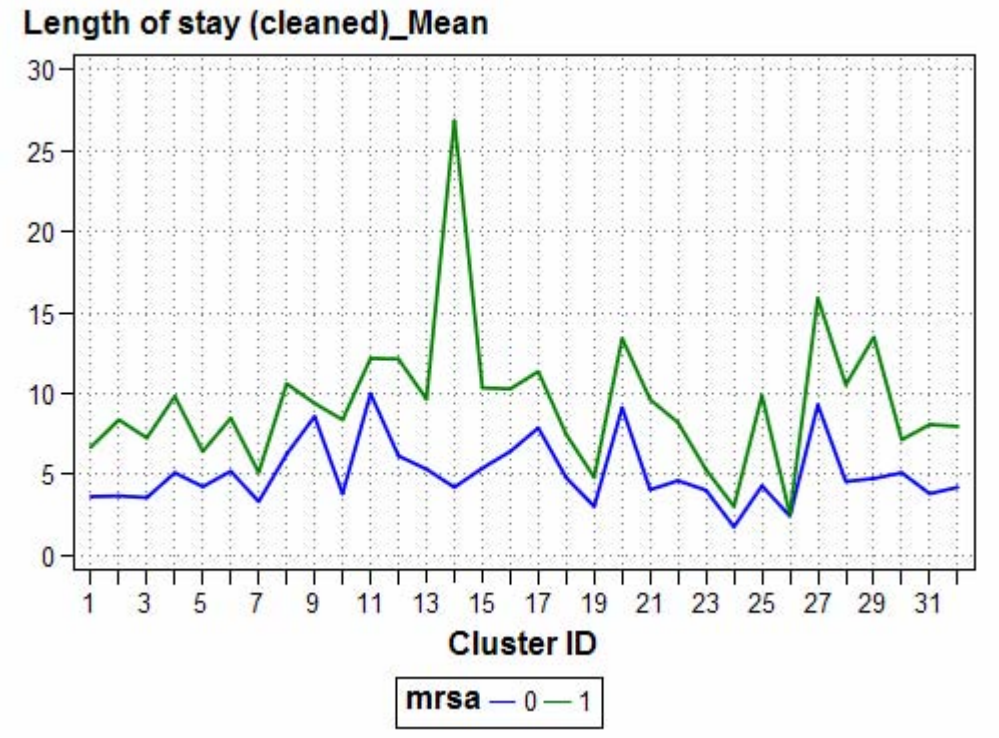
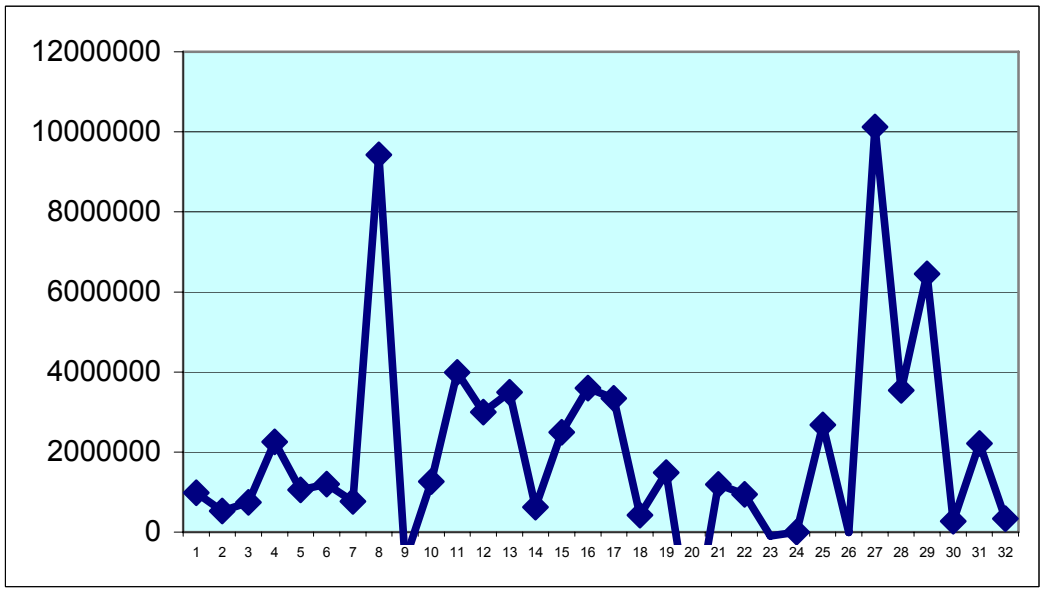
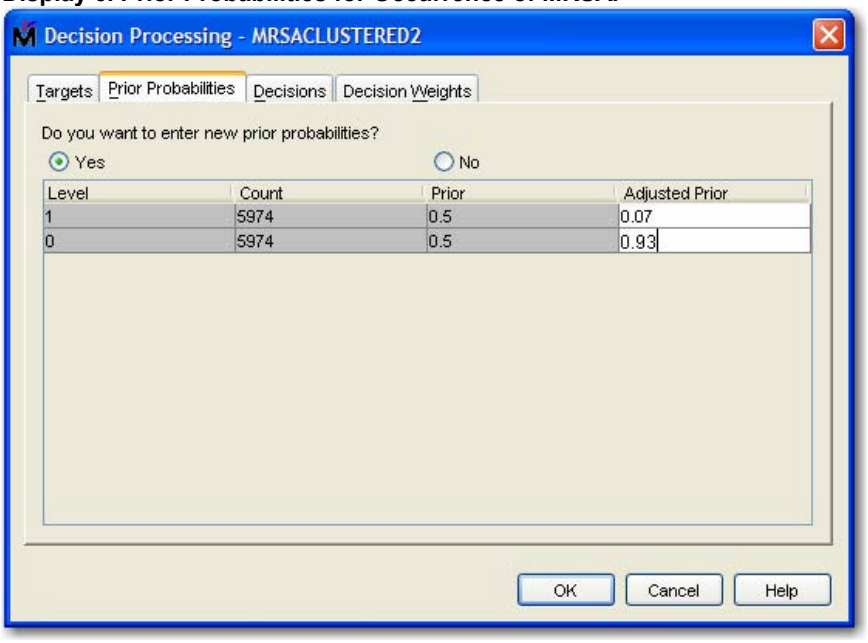


Figure 8. Total Cost of MRSA by Diagnosis

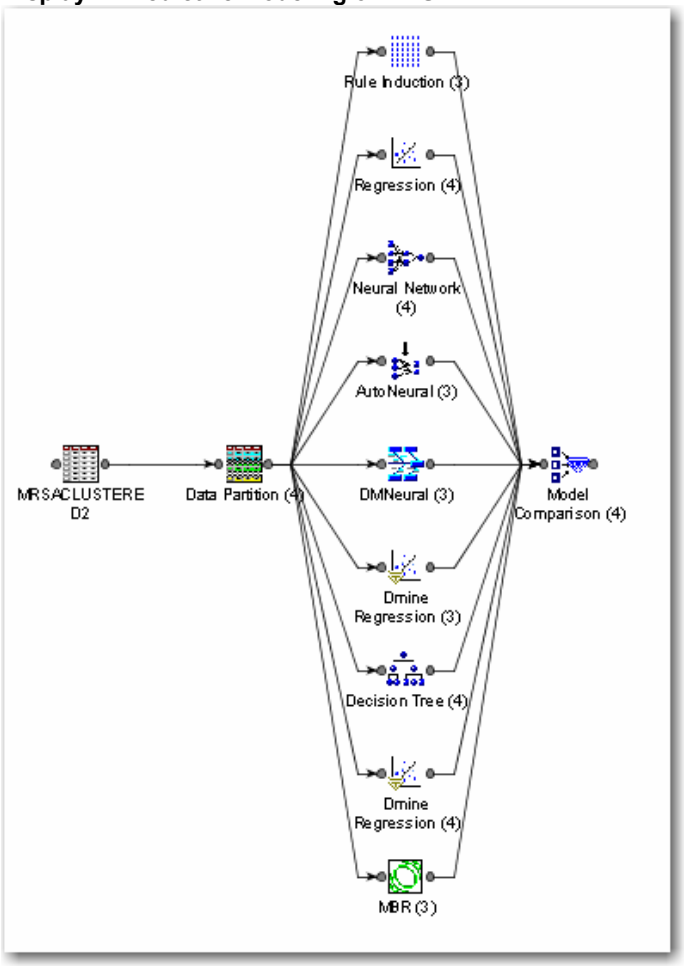


With the stratified sample, we retain the probability of actual occurrence (Display 6). Display 7 gives the predictive model.

Display 6. Prior Probabilities for Occurrence of MRSA.



Display 7. Predictive Modeling of MRSA



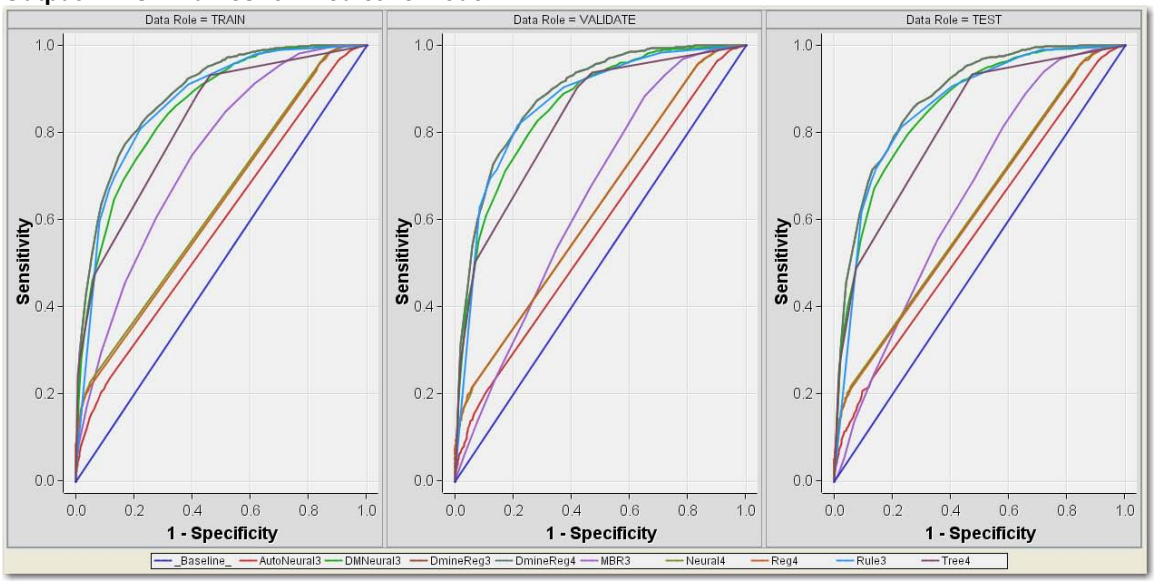
The model comparison node identifies the decision tree as the optimal model (Output 1).

Output 1. Model Comparison Output

Selected Model	Model Node	Test: Average Profit for mrsa	Train: Average Squared Error	Valid: Average Squared Error	Test: Average Squared Error	Train: Akaike's Information Criterion	Train: Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
	AutoNeural3	0.89905	0.24597	0.24895	0.24989	7539.77	0.49226	0.49498	0.49359
	DMNeural3	0.67724	0.15692	0.15616	0.15762	-5494.98	0.36773	0.36914	0.38120
	DmineReg3	0.59838	0.13848	0.13870	0.14154	.	0.32482	0.32785	0.33714
	DmineReg4	0.59838	0.13848	0.13870	0.14154	.	0.32482	0.32785	0.33714
	MBR3	.	0.20374	0.23458	0.23252	-4251.48	0.33654	0.40848	0.40212
	Neural4	0.90117	0.22676	0.23109	0.23225	6588.70	0.48891	0.49163	0.48968
	Reg4	0.90145	0.22696	0.23003	0.23252	6230.67	0.46170	0.46010	0.45929
	Rule3	0.57462	.	.	.	.	0.31833	0.32338	0.33212
Y	Tree4	0.90931	0.16359	0.16211	0.16638	.	0.37338	0.37388	0.37535

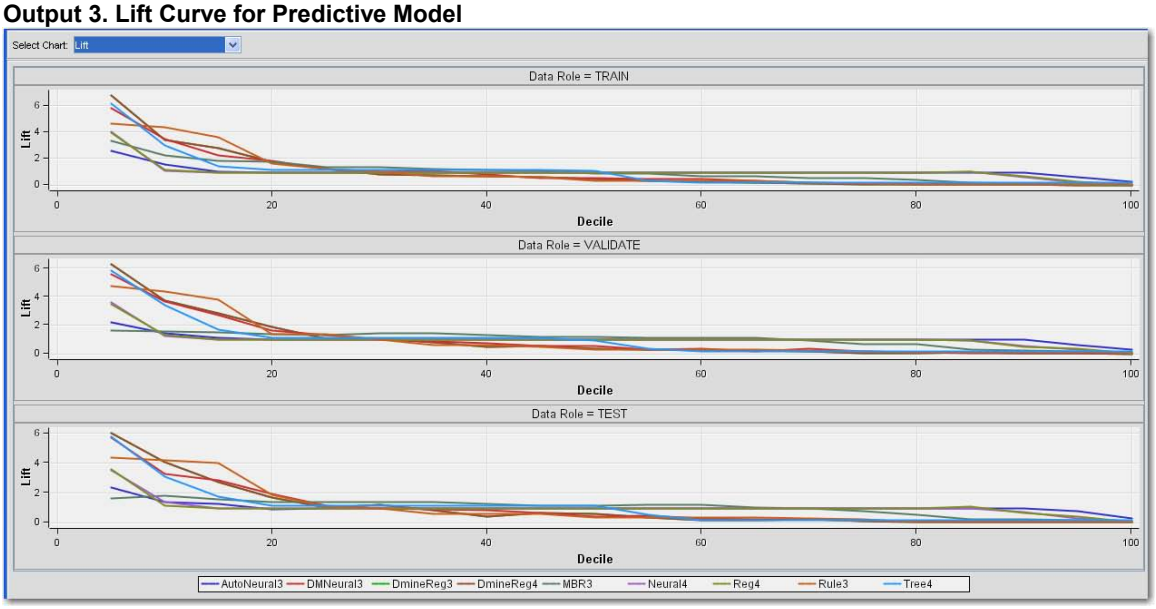
The ROC curves in Output 2 show that the Tree results are similar to those for DMNeural and DmineReg.

Output 2. ROC Curves for Predictive Model



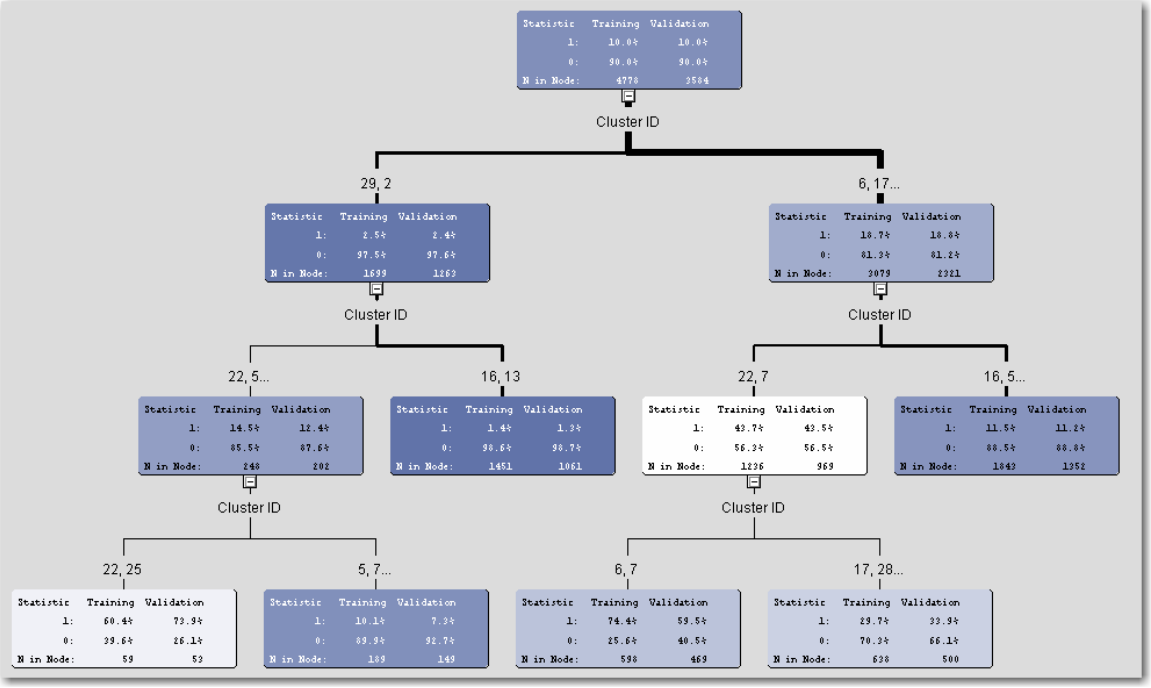
While the misclassification rates seem high, the lift curves show that the patients with the greatest risk can be predicted (Output 3).

Output 3. Lift Curve for Predictive Model



As the Decision Tree was the optimal model, we examine it in more detail (Output 4). Note that the tree decisions are based upon the procedure clusters rather than the diagnosis clusters. This implies that it is the treatments in the hospital that increase the risk of MRSA more so than the severity of the patient's condition, especially for those patients at highest risk.

Output 4. Decision Tree for MRSA



We expand upon Output 4 by providing the English Rules for the Decision Tree. They clearly show that the primary input variable is the procedure cluster.

IF Cluster ID IS ONE OF: 16 13 3 21 18 29 27 6 9 10 26 11 14 28 24 20 17 8 2 30
AND Cluster ID IS ONE OF: 29 2 18 26 32 1 3 14 24 21
THEN
NODE : 5
N : 2348
1 : 1.4%
0 : 98.6%

IF Cluster ID IS ONE OF: 16 5 13 3 21 29 27 9 10 26 11 14 24 17 8 2 30
AND Cluster ID IS ONE OF: 6 17 28 7 23 20 11 22 4 8 13 12 10 5 15 19 27 25 30 16 31 9
THEN
NODE : 6
N : 1729
1 : 11.5%
0 : 88.5%

IF Cluster ID IS ONE OF: 22 25 31
AND Cluster ID IS ONE OF: 29 2 18 26 32 1 3 14 24 21
THEN
NODE : 8
N : 18
1 : 60.4%
0 : 39.6%

IF Cluster ID IS ONE OF: 5 7 12 19 23 4
AND Cluster ID IS ONE OF: 29 2 18 26 32 1 3 14 24 21
THEN
NODE : 9
N : 188
1 : 10.1%
0 : 89.9%

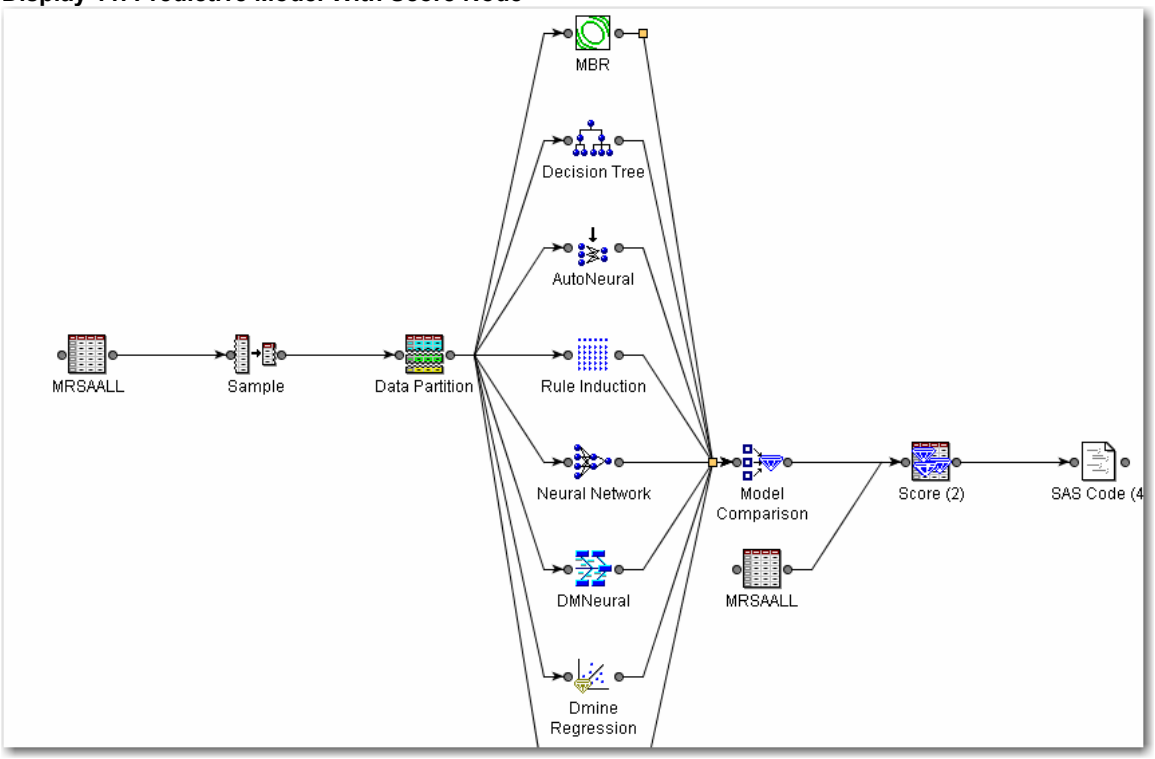
IF Cluster ID IS ONE OF: 17 28 22 8 13 12 10 5 27 25 31
AND Cluster ID IS ONE OF: 22 7 25 15 12 1 19 23 31 4
THEN
NODE : 14
N : 340
1 : 29.7%
0 : 70.3%

IF Cluster ID IS ONE OF: 6 7 23 20 11 4 15 19 30 16 9
AND Cluster ID IS ONE OF: 22 7 25 15 12 1 19 23 31 4
THEN
NODE : 15
N : 155
1 : 74.4%
0 : 25.6%

The decision rule clearly shows that patient treatment rather than patient condition is the leading predictor of MRSA. In particular, the top 2 deciles can be predicted much better than chance. Knowing that these patients have a high likelihood of contracting MRSA, steps can be taken to treat them prophylactically to prevent infection.

Once we find the optimal predictive model, we can score the original dataset to find the individual patients most at risk for MRSA (Display 11).

Display 11. Predictive Model With Score Node



The score node adds the following variables to the dataset (Output 5).

Output 5. Variables Added by Score Node

Score Output Variables

NAME	FUNCTION	CREATOR	LABEL
D_MRSA	DECISION	Rule	Decision: MRSA
EL_MRSA	ASSESS	Rule	Expected Loss: MRSA
EM_CLASSIFICATION	CLASSIFICATION	Score2	Prediction for mrsa
EM_DECISION	DECISION	Score2	Recommended Decision for mrsa
EM_EVENTPROBABILITY	PREDICT	Score2	Probability for level 1 of mrsa
EM_LOSS	ASSESS	Score2	Expect Loss for mrsa
EM_PROBABILITY	PREDICT	Score2	Probability of Classification
EM_SEGMENT	TRANSFORM	Score2	Segment
I_mrsa	CLASSIFICATION	Rule	Into: mrsa
P_MRSA0	PREDICT	Rule	Predicted: mrsa=0
P_MRSA1	PREDICT	Rule	Predicted: mrsa=1
U__bin	CLASSIFICATION		Unnormalized Into: _bin
U_mrsa	CLASSIFICATION	Rule	Unnormalized Into: mrsa
LEAF	TRANSFORM	Rule	Leaf
NODE	TRANSFORM	Rule	Node
WARN	ASSESS	Rule	Warnings
_bin	TRANSFORM		
b_mrsa	TRANSFORM	MdlComp	

To find the decile for an individual patient, we can use the following code in the code node (where the names of the datasets are extracted from Display 12 found by highlighting the link between the score node and the SAS code node. Note that the name of the scored dataset in Display 12 is equal to EMWS.score2_SCORE; it should be substituted for scored in the macro code.

```
/* extract number of observations from scored data */
%let dsid=%sysfunc(open(scored));
%let nobs=%sysfunc(attrn(&dsid,nobs));
%let rc=%sysfunc(close(&dsid));

/* add variable to randomize within score */
data a;
  set scored;
  ranvar=rand('uniform');
run;

/* sort by score and randomly within score */
proc sort data=a out=b;
  by descending score ranvar;
run;

/* assign deciles based on row */
data c;
  set b;
  decile=ceil(10*_n_/&nobs.);
run;

proc freq data=c;
  tables decile;
run;

proc print data=c;
  var id score decile ranvar;
run;
```

Display 12. Datasets Created by Score Node

From	Score2
To	EMCODE4
TRAIN	
Table	EMWS.Score2_TRAIN
Variables	
Role	Train
VALIDATE	
Table	EMWS.Score2_VALIDATE
Variables	
Role	Validate
TEST	
Table	EMWS.Score2_TEST
Variables	
Role	Test
SCORE	
Table	EMWS.Score2_SCORE
Variables	
Role	Score

CONCLUSION

One of the confounding factors with the occurrence of resistant infection is the practice of infection control in hospitals-and the variability in the adherence to infection control procedures. Even so, once different groups of procedures can be used to predict the occurrence of resistant infection, steps can be taken to reduce that occurrence through the use of prevention in the form of prophylactic antibiotics, or in increased adherence to infection control. Treatment procedures related to dialysis, for example, have a much higher risk of infection, and this knowledge can be used to reduce the problem.

REFERENCES

1. Pear R. Medicare Says It Won't Cover Hospital Errors. *New York Times.* August 19, 2007, 2007.
2. Henderson DK. Managing methicillin-resistant staphylococci: a paradigm for preventing nosocomial transmission of resistant organisms. *American Journal of Medicine.* 2006;119(6 Suyppl 1):S45-52.
3. Cola J, Duckworth G, Edwards D, et al. Guidelines for the control and prevention of meticillin-resistant Staphylococcus aureus in healthcare facilities. *Journal of Hospital Infection.* 2006;63(Suppl 1):S1-44.
4. Jr. TPL, McKinnon PS, Rybak M. Prediction model to identify patients with Staphylococcus aureus bacteremia at risk for methicillin resistance. *Infection Control & Epidemiology.* 2003;24(9):655-661.
5. Roghmann M. Predicting methicillin resistance and the effect of inadequate empiric therapy on survival in patients with Staphylococcus aureus bacteremia. *Archives of Internal Medicine.* 2001;160(7):1001-1004.
6. McBryde E, Pettit A, McElwain D. A stochastic mathematical model of methicillin resistant Staphylococcus aureus transmission in an intensive care unit: predicting the impact of interventions. *Journal of Theoretical Biology.* 2007;245(3):470-481.
7. Gerber SI, Jones RC, Scott MV, et al. Management of outbreaks of methicillin-resistant Staphylococcus aureus infection in the neonatal intensive care unit: a consensus statement. *Infection Control & Hospital Epidemiology.* 2006;27(2):139-145.
8. Masaki H, Nagatake T, Asoh N, et al. Significant reduction of nosocomial pneumonia after introduction of disinfection of upper airways using providone-iodone in geriatric wards. *Dermatology.* 2006;212(Suppl 1):98-102.
9. Bootsma M, Diekmann O, Bonten M. Controlling methicillin-resistant Staphylococcus aureus: quantifying the effects of interventions and rapid diagnostic testing. *Proceedings of the National Academy of Sciences in the United States of America.* 2006;103(14):5620-5625.

10. Loveday H, Pellowe C, Jones S, Pratt R. A systematic review of the evidence for interventions for the prevention and control of methicillin-resistant Staphylococcus aureus (1996-2004): report to the Joint MRSA Working Party. *The Journal of Hospital Infection.* 2006;63(Suppl 1):S45-70.
11. Malde D, Hardern L, Welch M. Is it possible to predict outcome in MRSA positive patients undergoing arterial reconstruction. *International Angiology.* 2006;25(1):78-83.
12. Thomas SC, Waghorn D, McIntyre A. The role of screening and antibiotic prophylaxis in the prevention of percutaneous gastrostomy site infection caused by methicillin-resistant Staphylococcus aureus. *Alimentary Pharmacology & Therapeutics.* 2007;25(5):593-597.
13. Anonymous. ICD9; 2006.

ACKNOWLEDGMENTS
The author wishes to thank John Cerrito, PharmD, for his invaluable expertise in medication and procedures in support of this project.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:

Name	Patricia B. Cerrito
Enterprise	University of Louisville
Address	Department of Mathematics
City, State ZIP	Louisville, KY 40292
Work Phone:	502-852-6010
Fax:	502-852-7132
E-mail:	pcerrito@louisville.edu
Web:	www.louisville.edu