BUSINESS UNDERSTANDING

Define the problem: The study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (Support) acquired vast amounts of data from seriously ill patients admitted into hospitals. The scope of data includes diagnosis, medical measurements, socioeconomic status, hospital charges, status at 2 months and 18 months, etc. Can the data be used to predict if a seriously ill hospital patient will die after 2 months? This would help doctors in clinical decision making and promote better communication among doctors, patients and patients' families.
Define the target variable: Death at two months.

DATA UNDERSTANDING

DATA DESCRIPTION

The SUPPORT2 dataset from the SAS Studio library contains data of all 9105 patients: 40 columns and 9105 rows. Out of the 40s variables, 11 are categorical, 5 of which are binary, and 29 are continuous.

Categorical Variables

- Dzgroup (specific disease) has 8 categories:
  - ARF/MOSF w/Sepsis (n=3515, 39%).
  - CHF (n=1387, 15%).
  - COPD (n=967, 11%).
  - Cirrhosis (n=508, 5%).
  - Colon Cancer (n=512, 6%).
  - Coma (n=596, 6%).
  - Lung Cancer (n=908, 10%).
  - MOSF w/Malig (n=712, 8%).
- Dzclass (category of disease) has 4 categories:
  - ARF/MOSF (n=4227, 46%).
  - COPD/CHF/Cirrhosis (n=2862, 31%).
  - Cancer (n=1420, 16%).
  - Coma (n=596, 7%).
- Income (income bracket):
  - 47% are under $11k (n=2855).
  - 25% are $11-$25k (n=1527).
  - 17% are $25-$50k (n=1057).
  - 11% are >$50k (n=684).
- Race:
  - 79.3% are white (n=7191).
  - 15.4% are black (n=1391).
  - 3.2% are Hispanic (n=290).
  - 1.2% are other (n=112).
  - 0.9% are Asian (n=79).
- Ca (cancer) has 3 categories:
  - Yes (n=1252, 14%).
  - No (n=5995, 66%).
  - Metastatic (n=1858, 20%).
- Sfdm2 (disability functionality at 2 months) in the order of most to least severe:
  - 40.5% are <2 mo. Follow up (died within 2 months, n=3123).
  - 0.5% are coma or intub (n=41).
  - 7.3% are SIP>=30 (Sickness Impact Profile score at 2 months >=30, n=564).
  - 11.9% are adl>=4 (>=5 if sur) (unable to do >=4 activities at month 2 or if surrogate was interviewed, ADL >=5, n=916).
  - 39.7% are no(M2 and SIP pres) (lived 2 months to get interviewed and showed no signs of moderate/severe functional disability, n=3061).

Binary Variables

- Sex is male (n=5125, 56%) and female (n=3980, 44%).
- Death is 0 (status not dead, n=2904, 32%) and 1 (status dead, n=6201, 68%).
- Hospdead is 0 (patient did not die in hospital, n=6745, 74%) and 1 (patient died in hospital, n=2360, 26%).
- Diabetes is 0 (no diabetes, n=7327, 80%) and 1 (with diabetes, n=1778, 20%).
- Dementia is 0 (no dementia, n=8809, 97%) and 1 (with dementia, n=296, 3%).

Continuous Variables

- An A number is assigned to each participant from 1 to 9105. It has no analytical meaning.
- Age: mean=62.7, median=64.9, std deviation= 15.6, min=18.0, max=101.8.
- Slos (days of study entry to discharge): mean=17.9, median=11, std deviation=22, min=3, max=343.
- D.time (days of follow-up): mean=478.4, median=233, std deviation=560, min=3, max=2029. T
- Num.co (number of comorbidities): mode=1, mean=1.8, median=2, std deviation=1.3, min=0, max=9.
- Edu (years of education): mean=11.7, median=12, std deviation=3.4, min=0, max=31.
- Scoma (coma score) (Figure 4): The original ordinal data, ranging from 3 to 15 on the Glasgow Coma Scale (GCS), has been previously rescaled to 0 to 100 (Knaus). In this rescaling, a score of 0 corresponds to a GCS of 15, while a score of 100 corresponds to a GCS of 3, indicating deep coma. Its basic statistics are mean =12, median =0, std deviation=24.6, min=0, max=100.
- Charges: mean=59995, median=25024, std deviation=102648, min=1169, max=1435423.
- Totcst (total RCC cost): mean=30825.9, median=14452.7, std deviation=45780.8, min=0, max=633212.
- Totmcst (total microcost): mean=28828.9, median=13223.5, std deviation=43604.3, min=-102.7, max=710682.
- Avtisst (average Therapeutic Intervention Scoring System score): mean=22.6, median=19.5, std deviation=13.2, min=1, max=83.
- Hday (days in hospital when patient entered study): mean=4.4, median=1, std deviation=9.1, min=1, max=148.
- Meanbp (mean blood pressure): mean=84.5, median=77, std deviation=27.7, min=0, max=195.
- Wblc (white blood cell count) (Figure 6): mean=12.3, median=10.6, std deviation=9.3, min=0, max=200.
- Hrt (heart rate): mean=97.2, median=100, std deviation=31.6, min=0, max=300.
- Resp (respiratory rate): mean=23.3, median=24, std deviation=9.6, min=0, max=90.
- Temp (temperature): mean=37.1, median=36.7, std deviation=1.3, min=31.7, max=41.7.
- Pafi (PaO2/FiO2 ratio) (Figure 3): mean=239.5, median=224, std deviation=109.7, min=12, max=890.4.
- Alb (albumin level) (Figure 10): mean=3.0, median=2.9, std deviation=0.9, min=0.4, max=29.
- Bili (bilirubin level) (Figure 2): mean=2.6, median=0.9, std deviation=5.3, min=0.1, max=63.
- Crea (creatinine level): mean=1.8, median=1.2, std deviation=1.7, min=0.1, max=21.5.
- Sod (sodium level): mean=137.6, median=137, std deviation=6.0, min=110, max=181.
- PH: mean=7.4, median=7.4, std deviation=0.1, min=6.8, max=7.8.
- Glucose: mean=159.9, median=135, std deviation=88.4, min=0, max=1092.
- BUN: mean=32.3, median=23, std deviation=26.8, min=1, max=300.
- Urine (output) (Figure 8): mean=2191.5, median=1968, std deviation=1455.2, min=0, max=9000.
- Adlp (activities of daily living): mean=1.2, median=0, std deviation=1.7, min=0, max=7.
- Adls (activities of daily living for the surrogate): mean=1.6, median=1, std deviation=2.2, min=0, max=7.
- Adlsc (imputed activities of daily living calibrated for the patient): mean=1.9, median=1, std deviation=2.0, min=0, max=7.

Data Exploration

To avoid bias when building certain types of predictive models, it is recommended to log10-transform distributions that have strong positive skewness, thereby promoting a more normal distribution. These variables include slos, d.time, BUN (Figure 1), charges, totmcst, hday, and bili. Totmcst includes 2 negative values whose validity is to be confirmed.

Avtisst and pafi (Figure 3) display moderate positive skewness, making it worth considering log10 transformation for these variables as well.

Some of the positively skewed variables have a minimum value of 0, requiring the addition of 1 to each value before log transformation. These variables include scoma (Figure 4), totcst, wblc (Figure 6), glucose (Figure 7), urine (Figure 8), adlp, adls, and adlsc.
PH (skewness=-1.0) (Figure 9) is the one distribution that exhibits negative skewness, requiring power transformation or flip transformation followed by log10 to make it suitable for analysis.
For the variable alb (Figure 10), the distribution may appear more normal if we remove the few outliers on the far right. To correct the skewness in pafi (Figure 3) and num.co (Figure 5), two options can be tried: remove the outliers on the right or log-10 transformation.

Data Quality

Among the categorical variables, three have missing values, all of which are missing completely at random (MCAR): Income (missing 2982), race (missing 42), and sfdm2 (missing 1400). These data can all be placed into their own group, "Missing."

Missing values are prevalent among the continuous variables.

- Edu is missing 1634 values, which may be imputed by several different methods: assign median number, assign a random number from the distribution, or predict it based on one's age and income, etc.
- Scoma has 1 missing value. It can be predicted based on a model that factors in activities of daily living and/or disability functionality.
- Charges, totcst, and totmcst all have missing values, which may be imputed with predictive models.
- Avtisst's 82 missing fields can be imputed by modeling.
- Medical measurements of Wblc, pafi, alb, bili, crea, pH, glucose, BUN, and urine will all be crucial towards the construction of the final models, so it's better to impute their missing values based on modeling rather than assigning mean value or random value from a distribution.
- Finally, adlp and adls have 5641 and 2867 missing values, respectively. Adlsc, however, has been calculated previously and has 0 missing number. It should function as a replacement for the above two variables.

DATA PREPARATION

Variable selection: variables dzclass and dzgroup are redundant, so I only selected dzgroup, for it's more detailed. Since we want to predict hospital death as soon as possible, almost right after admission, I cannot include d.time, death, adlp, adls, adlsc, charges, totcst, totmcst, slos, avtisst and smdf2 as input variables. These variables would cause future knowledge to leak into the model. All the other variables, except A, can be included as input variables.

Missing values: I imputed dummy variable "missing" for the missing income and race inputs. Upon preliminary modeling, I discovered that the medical measurements such as pafi and hrt were not typically important in the predictions, so instead of modeling them individually, I used normal values from https://biostat.app.vumc.org/wiki/Main/SupportDesc to impute some of the missing physiologic data. For missing values of edu, scoma, glucose, pH, sod, temp, resp, hrt, and meanbp, I imputed either mean or median. I built a decision tree model to impute the missing value of scoma.

Feature creation: I created the target variable death2m (death within 2 months) by calculation using patients' time-related variables of death, sfdm2, hospdead, and d.time.

MODELING

Considering the large number of input variables and its inclusion of both categorical and numeric data, I decided to build decision tree models to take advantage of their flexibility. Target variable was death2m. Input variables were dzgroup, income, race, ca, sex, diabetes, dementia, age, num.co, edu, scoma, meanbp, wblc, hrt, resp,temp pafi, alb bili, crea, sod, ph, glucose, bun, and urine. Using SAS, I tried different tree depths and all different combinations of pruning and growing methods.

MODEL ASSESSMENT

I was able to build a simple tree with tree depth of 5 and 7 leaves (Figure 11). The model did not overfit, as the error rate was (0.14, 0.13) for not dead and (0.55, 0.54) for dead (Figure 14). Throughout my report, the first number in parentheses will pertain to training data, the second number testing data.

Confusion matrices (Figure 14):
True positive (1120, 297); true negative (4117, 1015); false positive (1381, 350); false negative (667, 158); sensitivity (86.1%, 86.5%); specificity (44.8%, 45.9%).

AUC was 0.72 for both training and validation datasets (Figure 14).

In order to improve these numbers, I decided to incorporate principal component analysis into my modeling.

MODEL REVISION AND RE-ASSESSMENT

To prepare the numeric data for principal component analysis, I log10 transformed the following variables: BUN, Bili, Pafi, Scoma, WBLC, Glucose, and Urine. I deleted some of the outliers in alb. I incorporated all the numeric variables into my PCA. The resulting scree plot indicated that the first 4 principal components were the most important, so I built decision tree models with the four PCs, and with and without the original variables.

The best result I received was a tree with tree depth of 5 and 10 leaves (Figure 16), containing the PCs and all the original variables. The model did not overfit, as the error rate was (0.08, 0.09) for not dead and (0.61, 0.61) for dead (Figure 20).

Confusion matrices (Figure 20):
True positive (955, 237); true negative (4319, 1097); false positive (413, 114); false negative (1582, 370); sensitivity (91.2.1%, 90.6%); specificity (37.4%, 39.0%).

AUC was 0.70 for both training and validation datasets (Figure 20).

Compared to the first decision tree model, this model containing the PCs is harder to interpret. But its higher sensitivity allows it to correctly identify more patients who may die within 2 months. On the other hand, its low specificity indicates its weakness in finding patients who are at low risk of dying at two months.

APPENDIX

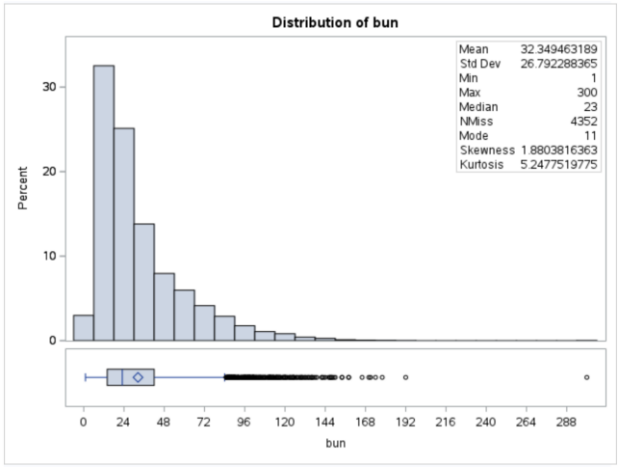

Figure 1. Distribution plot and basic statistics of BUN

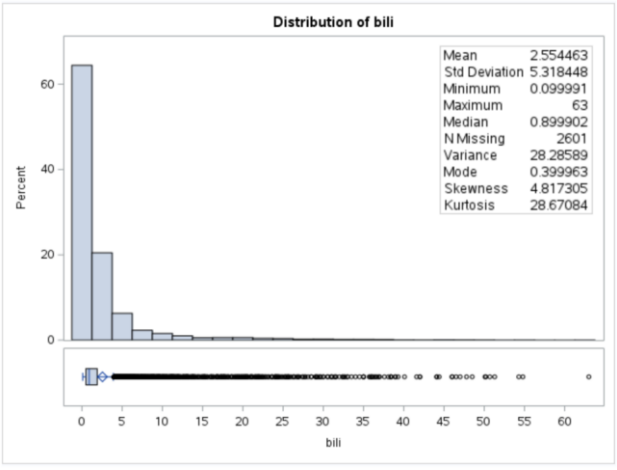

Figure 2. Distribution plot and basic statistics of Bilirubin

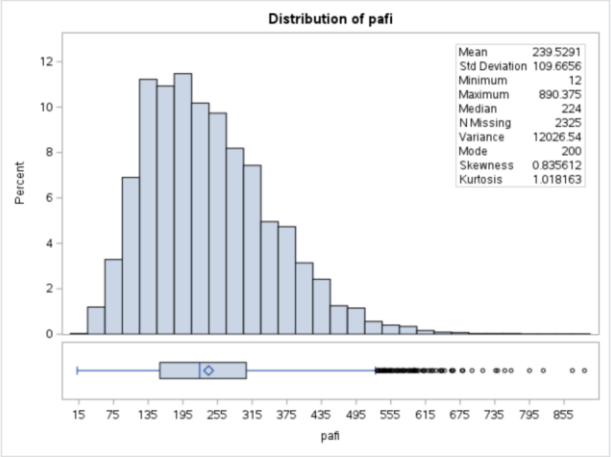

Figure 3. Distribution plot and basic statistics of PaO2/FiO2 ratio

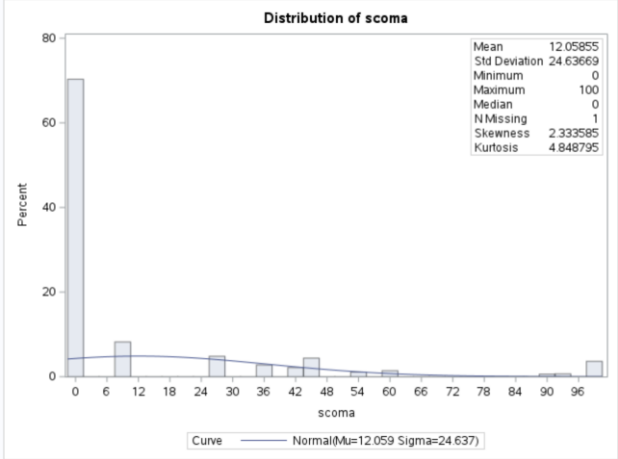

Figure 4. Distribution plot and basic statistics of coma score

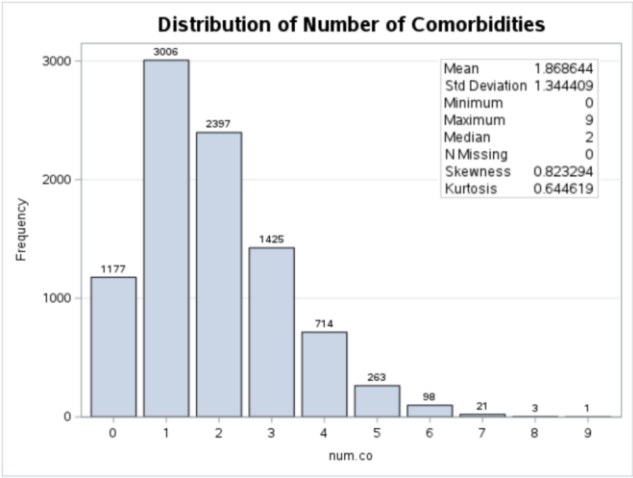

Figure 5. Distribution plot and basic statistics of number of comorbidities

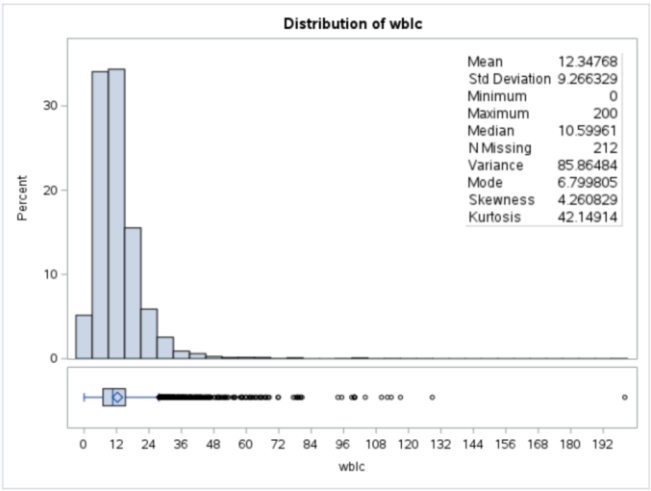

Figure 6. Distribution plot and basic statistics of white blood cell

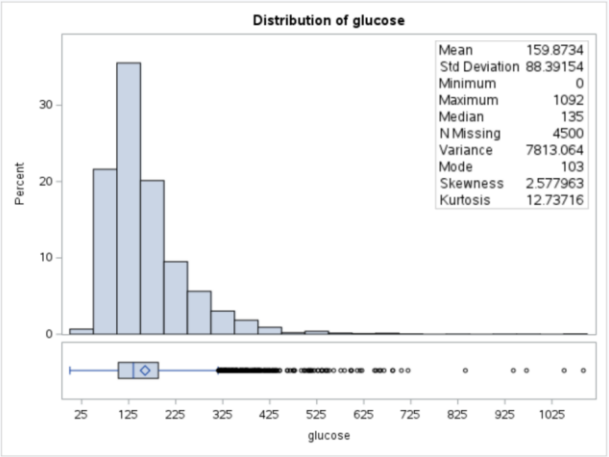

Figure 7. Distribution plot and basic statistics of glucose

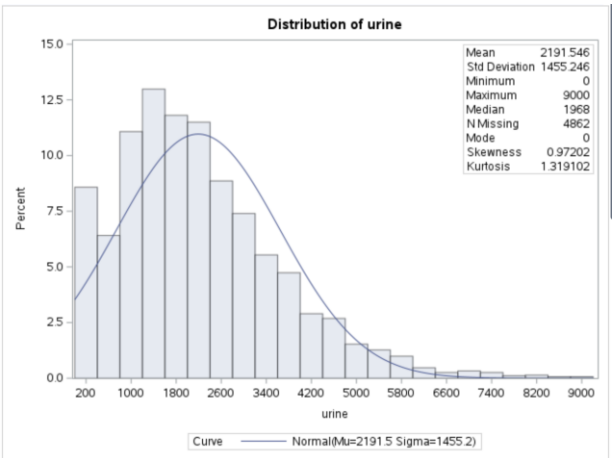

Figure 8. Distribution plot and basic statistics of urine output

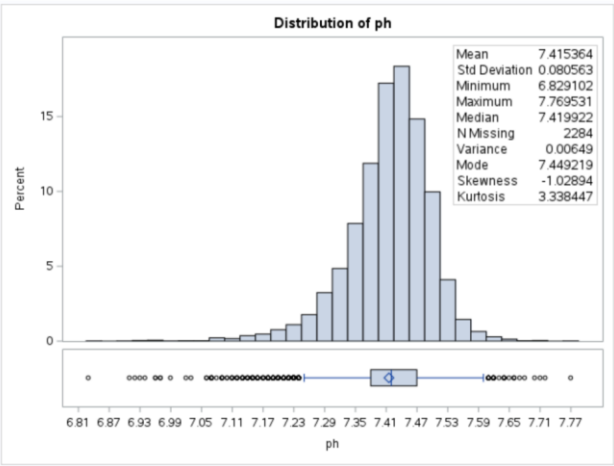

Figure 9. Distribution plot and basic statistics of pH

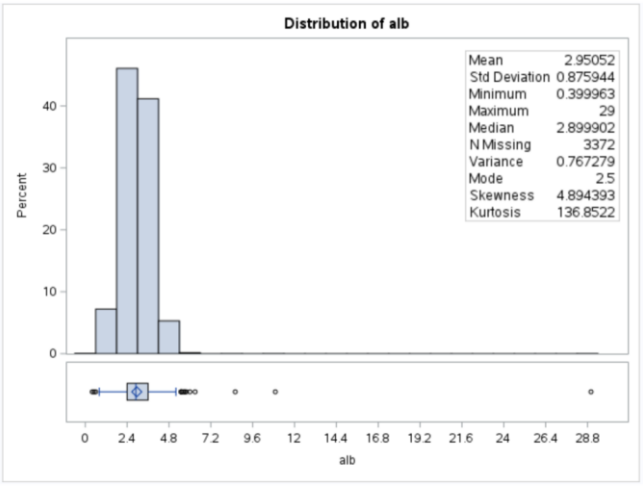

Figure 10. Distribution plot and basic statistics of albumin

Model Information	
Split Criterion Used	Chi-Square
Pruning Method	Entropy
Subtree Evaluation Criterion	Entropy
Number of Branches	2
Maximum Tree Depth Requested	5
Maximum Tree Depth Achieved	5
Tree Depth	5
Number of Leaves Before Pruning	30
Number of Leaves After Pruning	7
Model Event Level	0

Number of Observations Read	9105
Number of Observations Used	9105
Number of Training Observations Used	7285
Number of Validation Observations Used	1820

Figure 11. Model 1's key information

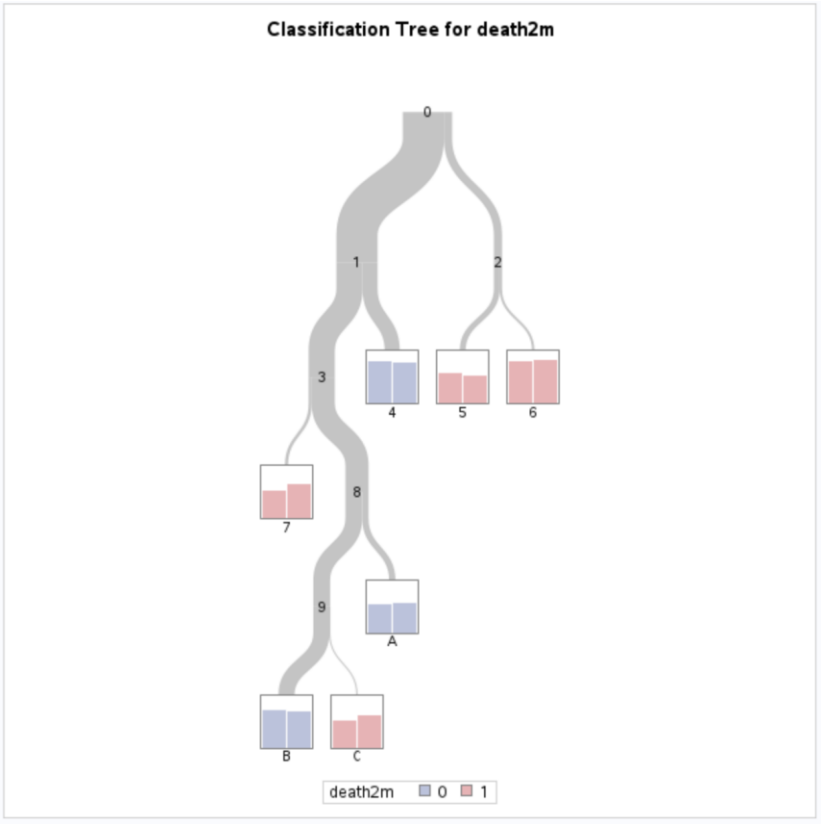

Figure 12. Model 1's basic tree structure

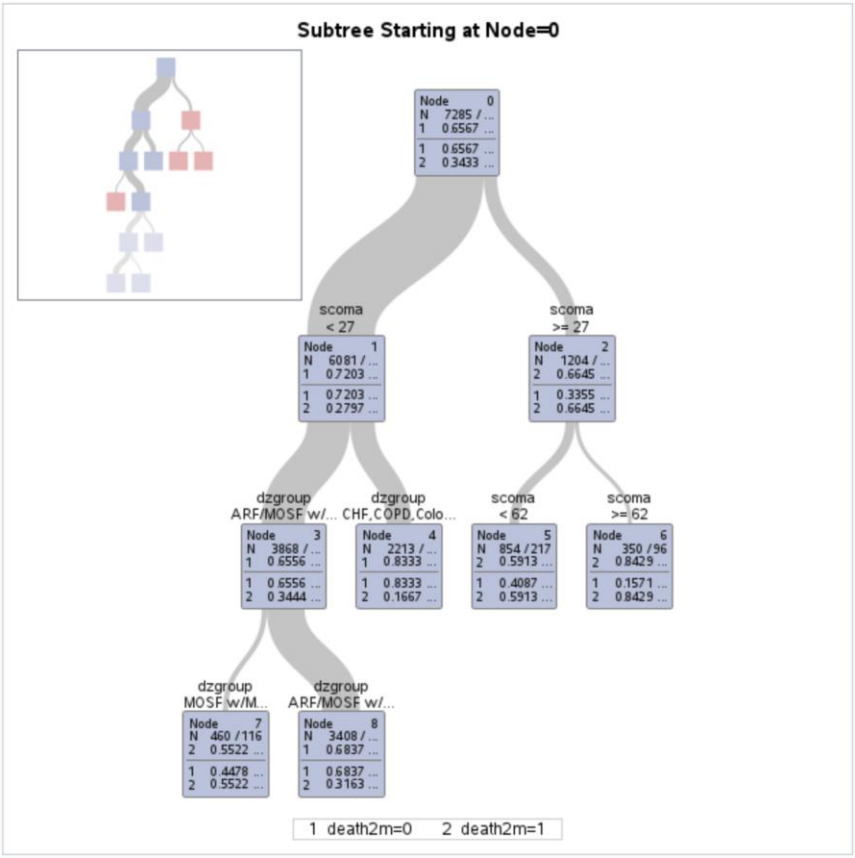

Figure 13. Model 1's top most nodes

Confusion Matrices

	Actual	Predicted 0	Predicted 1	Error Rate
Training	0	4117	667	0.1394
	1	1381	1120	0.5522
Validation	0	1015	158	0.1347
	1	350	297	0.5410

Fit Statistics for Selected Tree

	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	7	0.1905	0.2811	0.8606	0.4478	0.8156	0.3810	2775.5	0.7183
Validation	7	0.1927	0.2791	0.8653	0.4590	0.8192	0.3828	701.5	0.7159

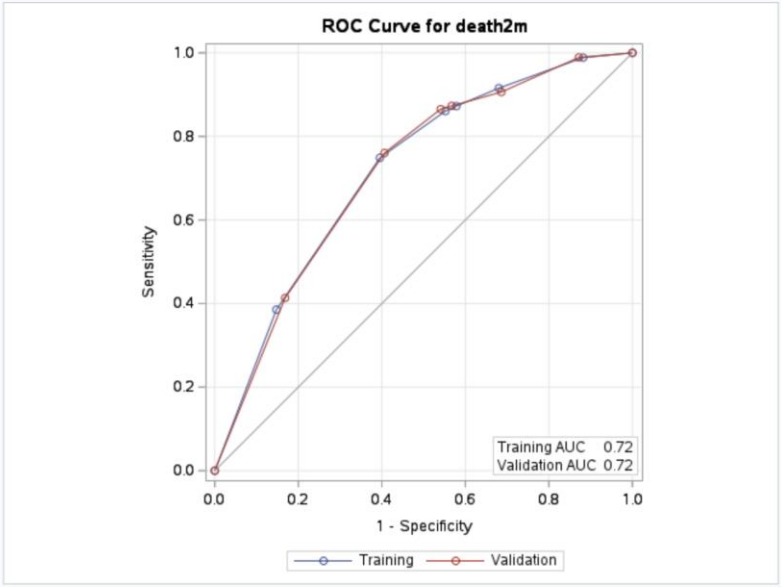

Figure 14. Model 1's confusion matrices, fit statistics and ROC curve

Variable Importance

Variable	Variable Label	Training Relative	Training Importance	Validation Relative	Validation Importance	Relative Ratio	Count
scoma	scoma	1.0000	18.1363	1.0000	8.7749	1.0000	2
dzgroup	dzgroup	0.6381	11.5721	0.7644	6.7078	1.1980	2
bili	bili	0.2307	4.1843	0.2915	2.5580	1.2635	1
age	age	0.2965	5.3773	0.2411	2.1158	0.8132	1

Figure 15. Model 1's top variables

Model Information	
Split Criterion Used	FastCHAID
Pruning Method	C4.5-Style
Subtree Evaluation Criterion	C4.5-Style
Number of Branches	2
Maximum Tree Depth Requested	5
Maximum Tree Depth Achieved	5
Tree Depth	5
Number of Leaves Before Pruning	32
Number of Leaves After Pruning	10
Model Event Level	0

Number of Observations Read	9087
Number of Observations Used	9087
Number of Training Observations Used	7269
Number of Validation Observations Used	1818

Figure 16. Model 2's key information

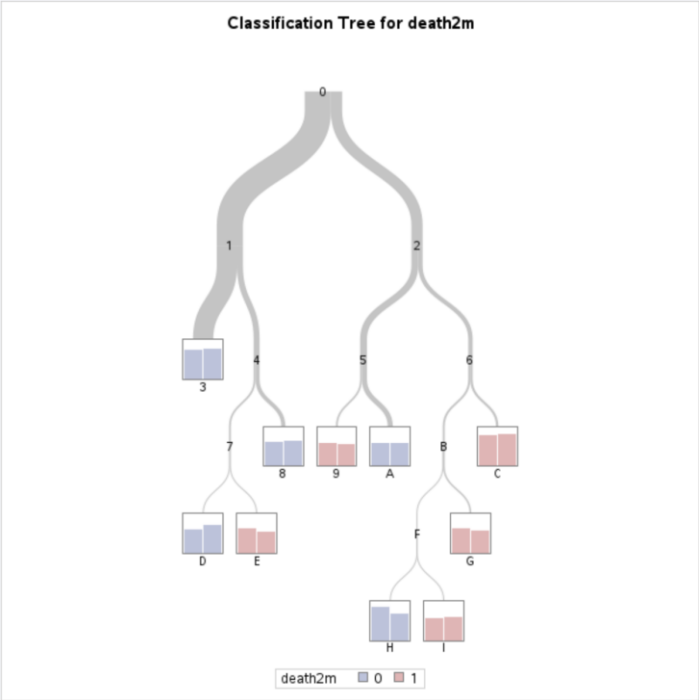

Figure 17. Model 2's basic tree structure

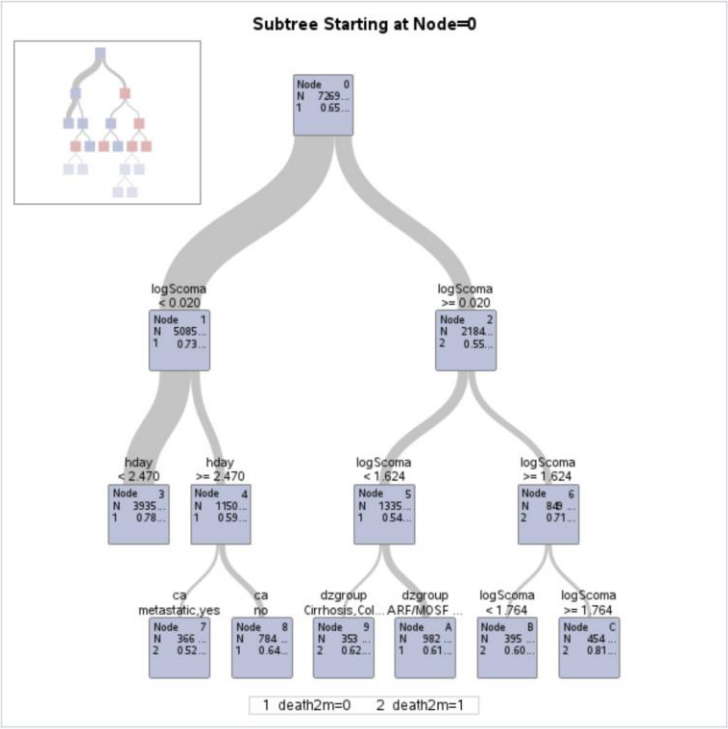

Figure 18. Model 2's top most nodes

Variable Importance							
		Training		Validation			
Variable	Variable Label	Relative	Importance	Relative	Importance	Relative Ratio	Count
logScoma		1.0000	18.7493	1.0000	10.0216	1.0000	3
hday	hday	0.4208	7.8903	0.3471	3.4786	0.8248	1
dzgroup	dzgroup	0.2859	5.3609	0.2234	2.2391	0.7814	1
Prin1		0.2064	3.8697	0.1655	1.6586	0.8019	1
age	age	0.1574	2.9515	0.1194	1.1966	0.7585	1
ca	ca	0.2071	3.8831	0.1170	1.1728	0.5650	1
num.co	num.co	0.1768	3.3142	0.0000	0	0.0000	1

Figure 19. Model 2's top variables

Confusion Matrices

	Actual	Predicted 0	Predicted 1	Error Rate
Training	0	4319	413	0.0873
	1	1582	955	0.6236
Validation	0	1097	114	0.0941
	1	370	237	0.6096

Fit Statistics for Selected Tree

	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	10	0.1933	0.2745	0.9127	0.3764	0.8271	0.3867	2810.8	0.6988
Validation	10	0.1890	0.2662	0.9059	0.3904	0.8096	0.3760	687.1	0.6973

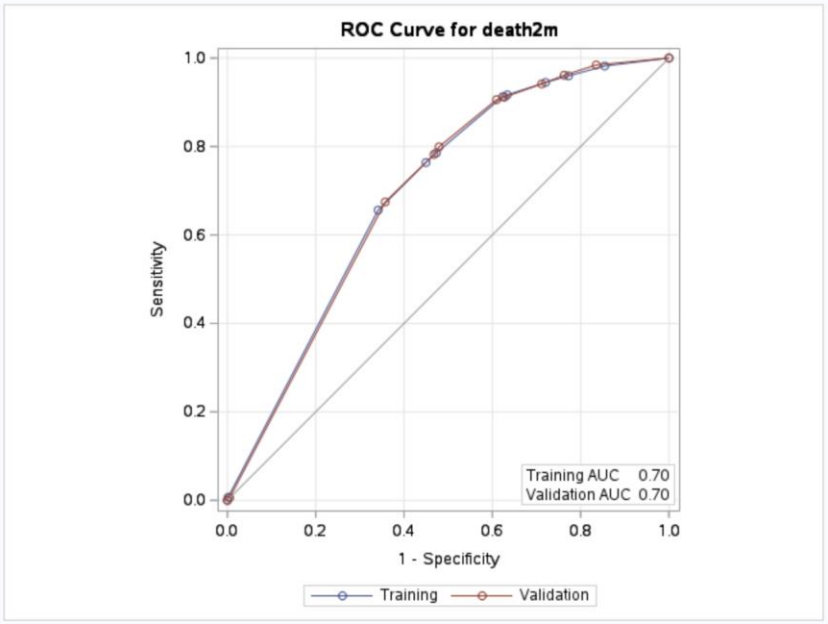

Figure 20. Model 2's confusion matrices, fit statistics and ROC curve