Analysis of Stroke data using Data Mining Methods

Weixiong Junyan Deng

Instructor: Dr. Alen Safer

Spring 2016

Applied Statistics

California State University Long Beach

1 Introduction

Stoke is “brain attack”. It occurs when blood flow to brain is clogged off. Brain cells are isolated from oxygen and begin to die. Stroke is the fifth leading cause of death in the US. Nearly 800,000 people experience stroke each year. More than 2/3 of survivors suffers some type of disability. Some patients can fully recover after being discharged from hospital in six months, but the others will still be at risk of disability or death.

In this study, we use two different stage information, information before treatment and events in the hospital, to define explanatory variable. Information before treatment includes age, gender, time from onset to treatment, presence or absence of atrial fibrillation (AF), aspirin administration within 3 days prior to treatment, systolic blood pressure, level of consciousness and neurological deficit. Events within 14 days include recurrence of stroke, side-effect, and pulmonary embolism. The target variable is the dead or alive form of discharge from hospital. The discharge day is not necessary within in 14 days. The goal of this study is to find a best model classify the patients with lowest misclassification rate. Our interest is how these 14 day events contribute to the death. Decision tree, neural network, and logistic regression are utilized to build models. In addition, ensemble model is built based on decision tree models. After comparing all models, we determine Neural Network Model 3 with Trust region as training technique and 4 hiding layer is the best model. It has 83.68% accuracy and 16.32% misclassification on validation dataset.

2. Data description

This data is come from US National library of Medicine National Institute of Health. In this study, we created 58 variables for analysis. The original variable Died is the target variable. 1 indicates the patient is dead in 6 months follow up. 0 indicates the patient is alive. If the patient is predicted as 1, this suggests that the patient need more help from the hospital after discharge. Table 1 shows part of the data description. More detail is shown on appendix. Table 2 shows the important variables.

Table 1 Description of part of the variables

HOSPNUM	Hospital number
RDELAY	Delay between stroke and randomisation in hours
RCONSC	Conscious state at randomisation (F - fully alert, D - drowsy, U - unconscious)
SEX	M = male; F = female
AGE	Age in years
RSLEEP	Symptoms noted on waking (Y/N)
RATRIAL	Atrial fibrillation (Y/N); not coded for pilot phase - 984 patients
RCT	CT before randomisation (Y/N)
RVISINF	Infarct visible on CT (Y/N)
RHEP24	Heparin within 24 hours prior to randomisation (Y/N)
RASP3	Aspirin within 3 days prior to randomisation (Y/N)
RSBP	Systolic blood pressure at randomisation (mmHg)
RDEF1	Face deficit (Y/N/C=can't assess)
RDEF2	Arm/hand deficit (Y/N/C=can't assess)
RDEF3	Leg/foot deficit (Y/N/C=can't assess)

Table2 Important Variables

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
DASPLT		1	1.0000	1.0000	1.0000
RCONSC		2	0.7882	0.8548	1.0846
REP_AGE	Replacement: AGE	3	0.6398	0.4657	0.7278
REP_ONDRUG	Replacement: ONDRUG	1	0.5113	0.5938	1.1615
STYPE		1	0.2881	0.2273	0.7889
RDEF3		1	0.0975	0.0778	0.7981

There are 14.2% percent of observations have missing values. So we impute the data using distribution method and mean respectively. But these two methods make the model worse than before imputation. So we just replace all missing value into the category of U(unknown).

3. Method

Decision tree, Neural network, and logistic regression method are used in this study.

3.1 Decision Tree

A decision tree splits into children branches from the top. The first parent node is the root node. Every parent node is split by using optimal rule from all variables. The terminal nodes of the tree are leaves. The leaves show the target value predicted by the model. The model is called classification tree if the target variable is categorical.

Classification and Regression Tree (CART) is first developed in 1973. It is easy to interpret.

3.2 Neural Network

Neural Network can depend on a large number of inputs to estimate or make an inference. Message can be exchanged between interconnected neurons. In each neuron, the system consists of a function, weight and bias. An input enters one neuron and after being weighted and transformed by the function, the output will then be passed on to next neuron. Neural Network can have more than one output.

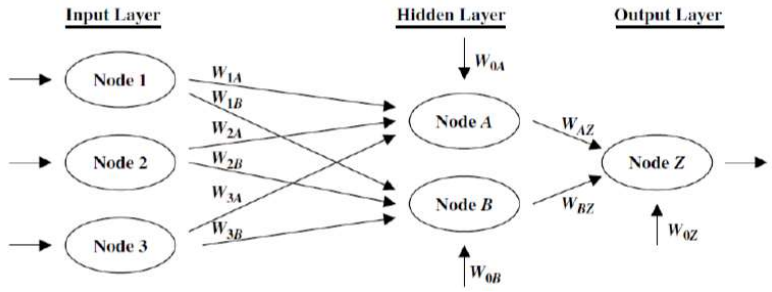

Figure 3.2 Neural Network Plot

3.3 Logistic Regression

Logistic regression estimates the probability using a logistic function in order to explain the relationship between the categorical dependent variable and the independent variables. Because the output variable is categorical, logistic regression can be used in classification. When Y=1, the probability function is,

$$P(Y = 1|X) = \frac{\exp(\beta 0 + \beta' 1X)}{1 + \exp(\beta 0 + \beta' 1X)}$$

P(Y=0|X) =1-P(Y=1|X).

The ratio P(Y=1|X)/P(Y=0|X) is odd ratio. When odd ratio is greater than 1, Y is more possibly to be 1 comparing to 0. When od ratio is less than 1, Y is more possibly to be 0 comparing to 1.

4. Results

4.1 Decision Tree

4.1.1 Model Description

We use different setting for each model in table 4.1.1 to build multiple Decision Tree models. all the model is set on significant at 0.2 level and 4 level of depth.

Table 4.1.1 Model Description

Model No.	Nominal Criterion	Ordinal Criterion	Max branch	Leave size
1	Porb	Entropy	2	3
2	Prob	Gini	2	5
3	Gini	Gini	2	3
4	Gini	Entropy	2	5
5	Entropy	Entropy	2	7
6	Entropy	Gini	2	5
7	Prob	Entropy	3	3
8	Entropy	Gini	3	5
9	Gini	Entropy	3	7

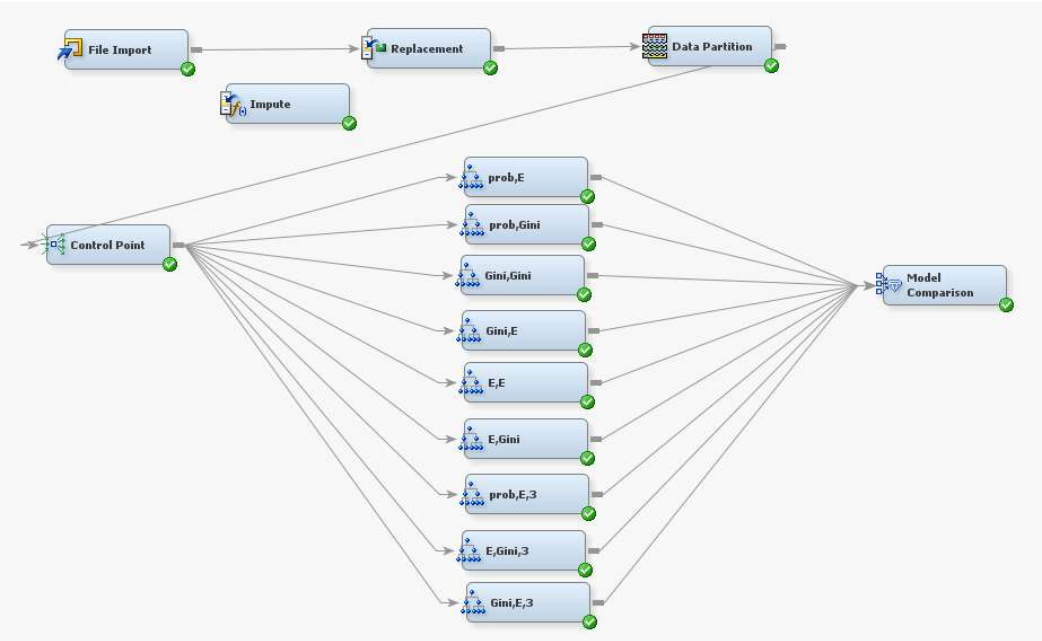

Figure 4.1.1 Flow Chart of Decision Trees

4.1.2. Comparison

From table 4.1.2 we find that tree 8 is the best model with Entropy-Gini criterion for nominal and ordinal criterion. The model has 16.64% misclassification at training set and 16.77% at validation set. The ORC curve also shows that the red line, tree 8, has the best accuracy.

Table 4.1.2 Fit Statistics

Fit Statistics

Predecessor Node	Model Description	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Train: Average Squared Error	Train: Maximum Absolute Error
Tree8	E,Gini,3	0.167732	0.166359	0.122208	0.967718
Tree9	Gini,E,3	0.169914	0.166206	0.126856	0.947368
Tree7	prob,E,3	0.172253	0.164823	0.123718	0.973553
Tree5	E,E	0.178176	0.166283	0.128668	0.900763
Tree6	E,Gini	0.178176	0.166283	0.128668	0.900763
Tree	prob,E	0.179111	0.165591	0.124241	0.925158
Tree2	prob,Gini	0.179111	0.165591	0.124241	0.925158
Tree3	Gini,Gini	0.179111	0.165591	0.124241	0.925158
Tree4	Gini,E	0.179111	0.165591	0.124241	0.925158

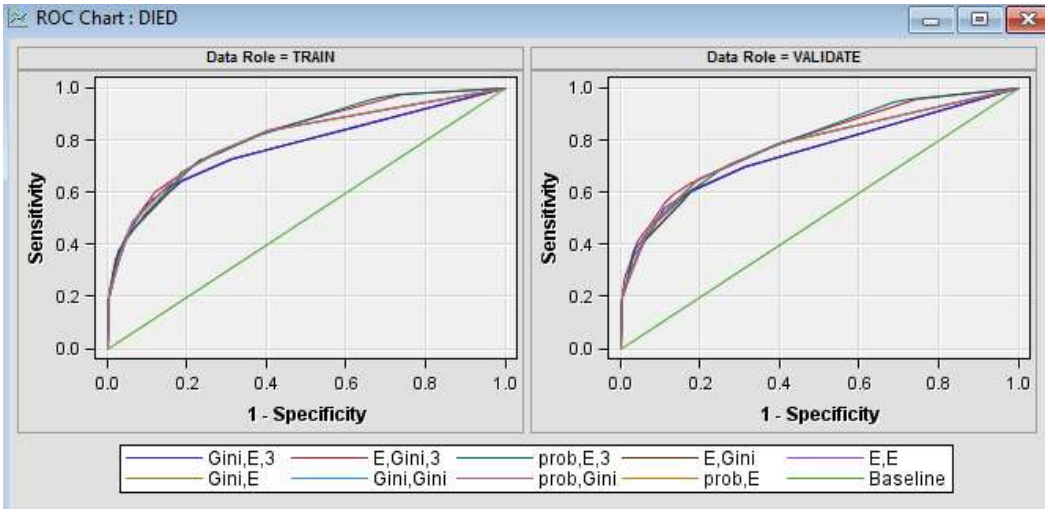

Figure 4.1.2 ORC Chart

4.1.3 Best model

For the best model decision tree 1, table 4.1.3 shows that Dasplt (discharge with long term asplin) is the most important variable, and Rconsc(recurrence of SC stroke) is the second most important variable. The Accuracy is 83.23% and misclassification is 16.77%.

Patients with DASPLT and age less than 67.5 are 95.28% predicted to be alive in 6 months. Patients with DASPLT and age less than 82.5 but greater than 67.5 are 87.45% predicted to be alive in 6 months. Patients without DASPLT and whose RCONSC is unconscious (Conscious state F - fully alert, D - drowsy, unconscious) are 96.08% predict to be dead in six months. Patients without DASPLT, whose RCONSC is DROWSY and ONDEUG (receive trial days) less than 5.5 are 92.18% predicted to be dead.

Table 4.1.3 Important Variables

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
REP_DASPLT	Replacement: DASPLT	1	1.0000	1.0000	1.0000
RCONSC		3	0.8133	0.8772	1.0786
REP_AGE	Replacement: AGE	3	0.6782	0.5841	0.8612
REP_IMP_ONDRUG	Replacement: Imputed ONDRUG	3	0.5436	0.6345	1.1674
STYPE		1	0.2024	0.2456	1.2129
RDEF3		1	0.1080	0.0887	0.8212
REP_DNOSTRK	Replacement: DNOSTRK	1	0.1016	0.0942	0.9267
REP_RDELAY	Replacement: RDELAY	1	0.0383	0.0453	1.1846

Table 4.1.3.1 Classification Summary

Classification Table

Data Role=TRAIN Target Variable=DIED Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	86.2056	93.6193	9449	72.5730
1	0	13.7944	51.6570	1512	11.6129
0	1	31.2773	6.3807	644	4.9462
1	1	68.7227	48.3430	1415	10.8679

Data Role=VALIDATE Target Variable=DIED Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	85.5231	92.5583	4602	71.7381
1	0	14.4769	53.9848	779	12.1434
0	1	35.7834	7.4417	370	5.7677
1	1	64.2166	46.0152	664	10.3507

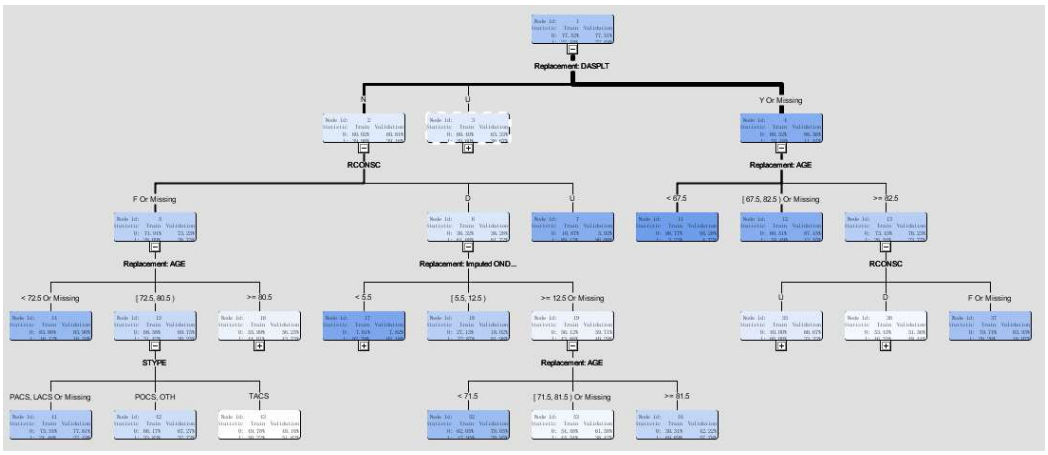

Figure 4.1.3.1 Decision Tree of Model 3

4.2 Neural Network

4.2.1 Model Description

We set different condition for each Neural Network Model as described in Table 4.2.1. All the models use Multilayer Perceptron Architecture. Figure 4.2.1 shows how the models were built in Enterprise Miner. Model

Table 4.2.1 Model Description

Model NO.	Training Technique	NO. of Hidden Unites
1	Trust Region	3
2	Back Drop	3
3	Trust Region	4
4	Back Drop	4

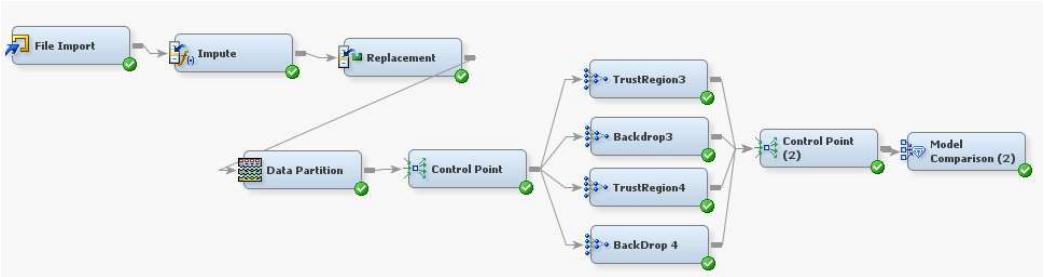

Figure 4.2.1 Flow chart of Neural Network

4.2.2 Model Comparison

Table 4.2.2 Shows that the Neural Network Model 1 is the best model with smallest misclassification rate 16.32 on validation dataset and 14.77% on training dataset. Figure 4.2.2 ROC graphic shows that all the Neural Network have similar performances.

Table 4.2.2 Fit statistics

Fit Statistics

Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Train: Final Prediction Error	Train: Average Squared Error	Train: Akaike's Information Criterion	Train: Maximum Absolute Error	Train: Schwarz's Bayesian Criterion	Train: Total Degrees of Freedom
Neural	TrustRegion,3	0.163211	0.147696	0.110869	0.107135	9445.454	0.995922	11112.21	13020
Neural4	BackDrop,4	0.163523	0.153533	0.116985	0.111767	9939.982	0.999668	12159.83	13020
Neural2	BackDrop,3	0.16477	0.153533	0.115697	0.1118	9821.695	0.997403	11488.45	13020
Neural3	TrustRegion,4	0.167576	0.148387	0.111751	0.106767	9558.782	0.995581	11778.63	13020

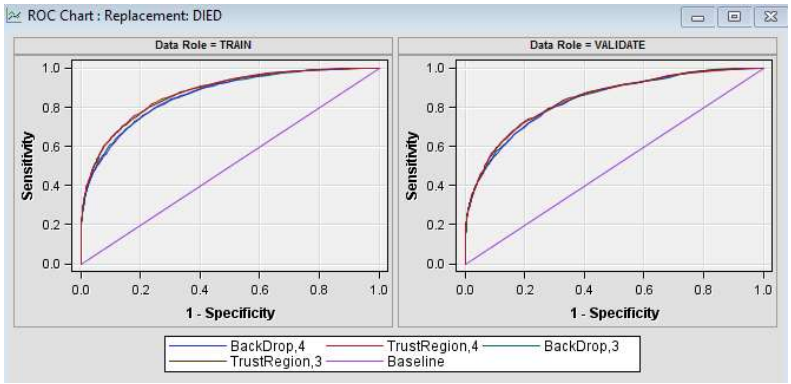

Figure 4.2.2 ORC graphic

4.2.3 Best Model

Table 4.2.2 shows that Neural Network Model 1 with trust region as training technique and 3 hiding lays is the best model. The model has 16.32% misclassification rate on validation data. Figure 4.2.3, Iteration Plot, shows that when training iteration is 6 the misclassification rate is the best for both training and validation data.

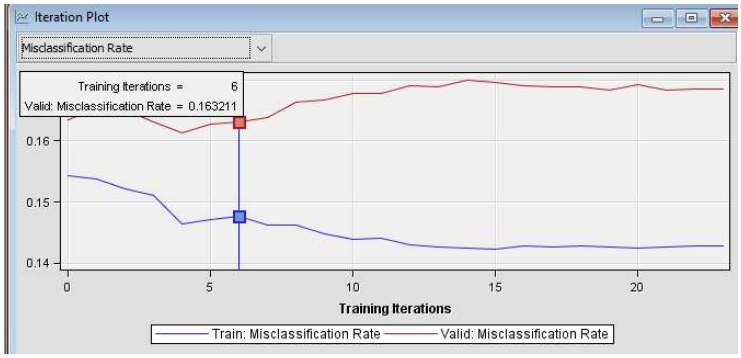

Figure 4.2.3 Iteration Plot

Table 4.2.3 Miss Classification summary

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	86.0928	95.9279	9682	74.3625
1	0	13.9072	53.4335	1564	12.0123
0	1	23.1680	4.0721	411	3.1567
1	1	76.8320	46.5665	1363	10.4685

Data Role=VALIDATE Target Variable=REP_DIED Target Label=Replacement: DIED

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	85.0693	95.1126	4729	73.7178
1	0	14.9307	57.5191	830	12.9384
0	1	28.3879	4.8874	243	3.7880
1	1	71.6121	42.4809	613	9.5557

4.3 Logistic Regression Model

4.3.1 Model Description

We build Logistic Regression Model by setting different criterions.

Table 4.3.1 Model Description

Model NO.	Selection Method	Main Effect	Two-factor Interaction
1	Forward	Yes	No
2	Forward	Yes	Yes
3	Backward	Yes	No
4	Backward	Yes	Yes
5	Stepwise	Yes	No
6	Stepwise	Yes	Yes

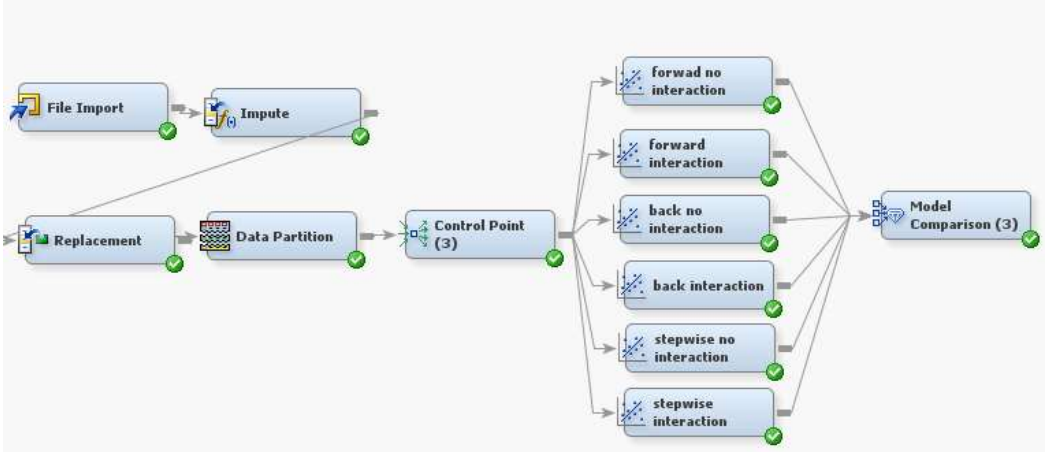

Figure 4.3.1 Flow Chart of Logistic Regression Model

4.3.2 Model Comparison

From Table 4.3.2, we can find out that regression model 1 is the best, with lowest misclassification rate, 17.18%, on validation data. This model shows that forward selection without two-factor interaction, and is the simplest one when comparing to other models with two-factor interaction. Although model 4 (backward with two-factor interaction) performs good at training data, but it performs bad at validation data. Additionally, model 4 with interaction term is more complex when comparing to model 1. This model is over fit on the training data and loses some generalization. So we decide that Model 1 is the best one.

Table 4.3.2.1 Fit Statistics

Fit Statistics

Predecessor Node	Model Description	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Train: Average Squared Error	Train: Average Error Function	Train: Akaike's Information Criterion	Target Label
Reg	forward, no interact	0.171785	0.158756	0.116924	0.377096	9869.59	Replaceme...
Reg3	backward,no interact	0.172564	0.158756	0.117008	0.377281	9870.39	Replaceme...
Reg5	stepwise, no interact	0.172564	0.158756	0.117008	0.377281	9870.39	Replaceme...
Reg6	stepwise interact	0.172564	0.155991	0.114227	0.369346	9723.758	Replaceme...
Reg2	forward, interact	0.173188	0.156989	0.113469	0.36687	9727.302	Replaceme...
Reg9	Backward,interaction	0.176306	0.155069	0.113829	0.368326	9763.201	Replaceme...

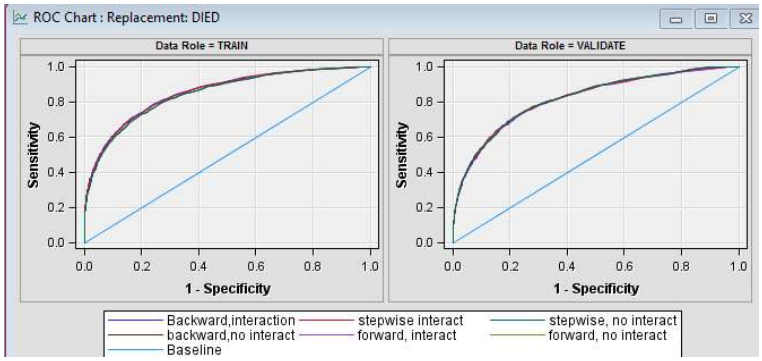

Figure 4.3.2.1 ORC Chart

Table 4.3.2.2 Miss Classification summary

Classification Table

Data Role=TRAIN Target Variable=REP_DIED Target Label=Replacement: DIED

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	86.2708	94.5705	9545	73.3103
1	0	13.7292	51.8961	1519	11.6667
0	1	28.0164	5.4295	548	4.2089
1	1	71.9836	48.1039	1408	10.8141

Data Role=VALIDATE Target Variable=REP_DIED Target Label=Replacement: DIED

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	85.2074	94.1874	4683	73.0008
1	0	14.7926	56.3410	813	12.6734
0	1	31.4472	5.8126	289	4.5051
1	1	68.5528	43.6590	630	9.8207

Model 1 consists of the following significant effects: AGE ONDRUG RDELAY REP_DASPLT REP_DDIAGHA REP_DMAJNCH REP_DOAC REP_DPE REP_DSTER REP_RCONSC REP_RVISINF REP_RXASP REP_STYPE RSBP. The estimates of parameters are shown below.

Table 4.3.2.2

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.0300	0.3786	113.31	<.0001
AGE		1	0.0737	0.00278	703.83	<.0001
ONDRUG		1	-0.0916	0.00610	225.62	<.0001
RDELAY		1	-0.00628	0.00214	8.60	0.0034
REP_DASPLT	N	1	0.5655	0.0605	87.22	<.0001
REP_DASPLT	U	1	0.4550	0.1087	17.52	<.0001
REP_DMAJNCH	N	1	-0.9180	0.4084	5.05	0.0246
REP_DMAJNCH	U	1	1.6961	0.8038	4.45	0.0349
REP_DPE	U	1	-1.5014	0.7483	4.03	0.0448
REP_DSTER	U	1	-1.3831	0.7075	3.82	0.0506
REP_RCONSC	F	1	-1.0282	0.0762	181.84	<.0001
REP_RVISINF	N	1	-0.1414	0.0284	24.72	<.0001
REP_RXASP	N	1	-0.1238	0.0262	22.35	<.0001
RSBP		1	-0.00305	0.000930	10.79	0.0010

5 Ensemble

5.1 Comparison of best models from different method

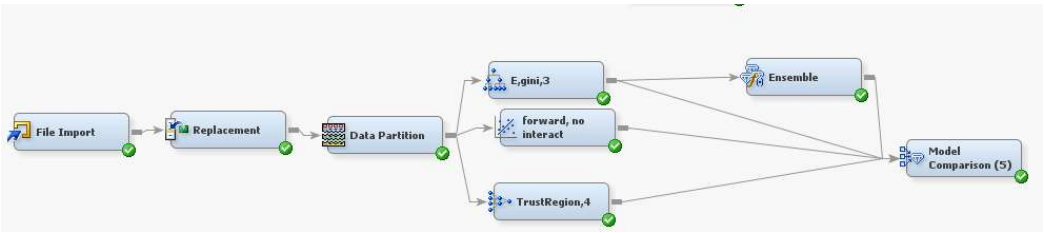

Figure 5.1 Flow Chart

From table 5.1 we can find out that Neural 3 is the best model. This model has lowest misclassification rate on both training data and validation data. The rates change a little bit because we repartition the dataset and run the model comparison again. ROC chart agree with our conclusion that neural 3 is the best model.

Table 5.1 Fit Statistics

Fit Statistics

Predecessor Node	Model Description	Train: Misclassification Rate	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE
Neural3	TrustRegion,4	0.15169	0.167264	0.112065	0.334761	26040
Ensmbl	Ensemble	0.164363	0.169914	0.121641	0.348771	26040
Tree8	E,gini,3	0.164363	0.169914	0.121641	0.348771	26040
Reg	forward, no interact	0.160061	0.175058	0.117363	0.342583	26040

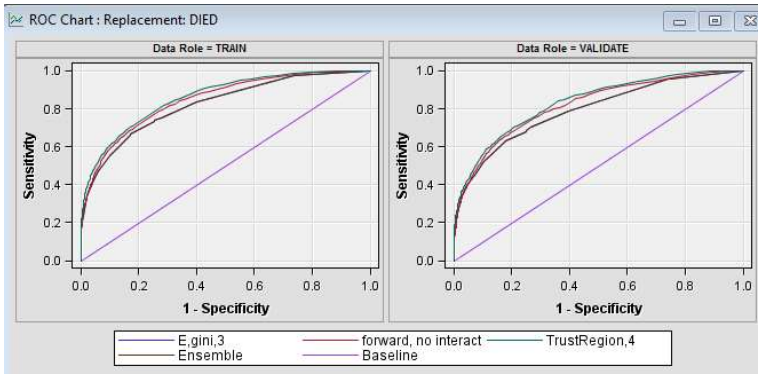

Figure 5.1.1 ROC Chart

6. Conclusion

The Neural 3 model is the best one comparing other methods. When training iteration is 6 the misclassification rate is the best for both training and validation data. The accuracy for this model is 83.27%, and misclassification rate is 16.73%. Those patients who are with long term aspirin diagnostic, younger in age and with fully alert state of consciousness when they are sent to hospital have higher probability to survive after discharge. Those patients who are not diagnosed with long term aspirin, older in age and are unconscious when being sent to hospital are more possible to be dead after discharge from hospital. These patients need more care to be safe.

7 Tableau

7.1 Dash Plot

We create visual analysis on Tableau software. We only generate analysis on important variables. From Figure 7.1, Sheet 1 RCONSC, state of consciousness. We can find out that patients with fully alert have higher probability to be survived in six months after discharge from hospital. Most of the patients are fully alert when they are sent to the hospital. On the other hand, unconscious patients are at high risk to be dead. From sheet 4 DASPLT, discharged with long term aspirin are more possible to be survived.

Stroke can happen in young people. It is a serious disease. The risk is going up as age increases. People between 70 and 75 are at the highest risk of suffering stoke. People between 75 and 88 are at the highest risk of death because of stroke. This can be shown in Figure 7.1.2.

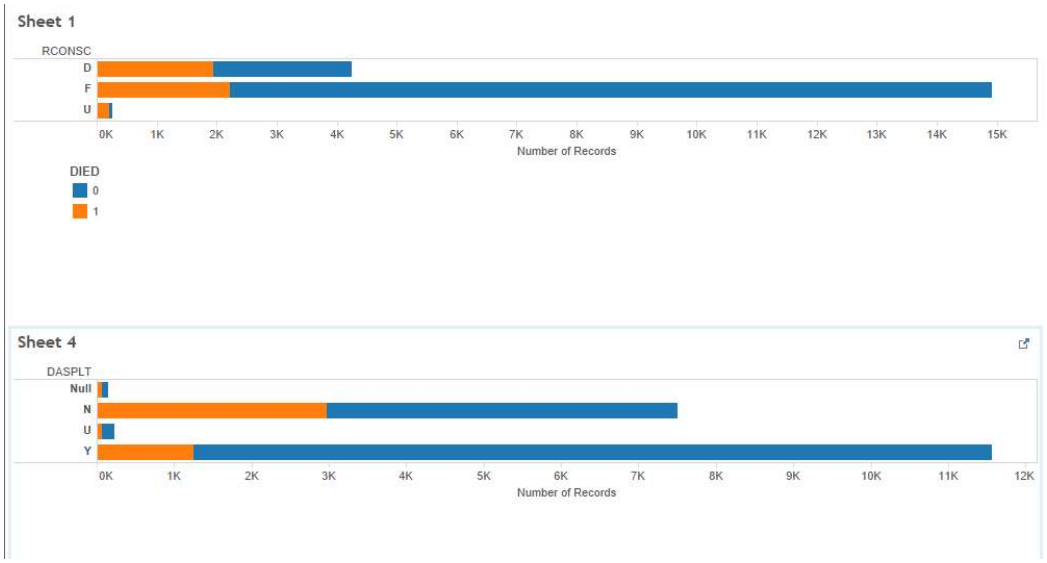

Figure 7.1.1 Dash Plot

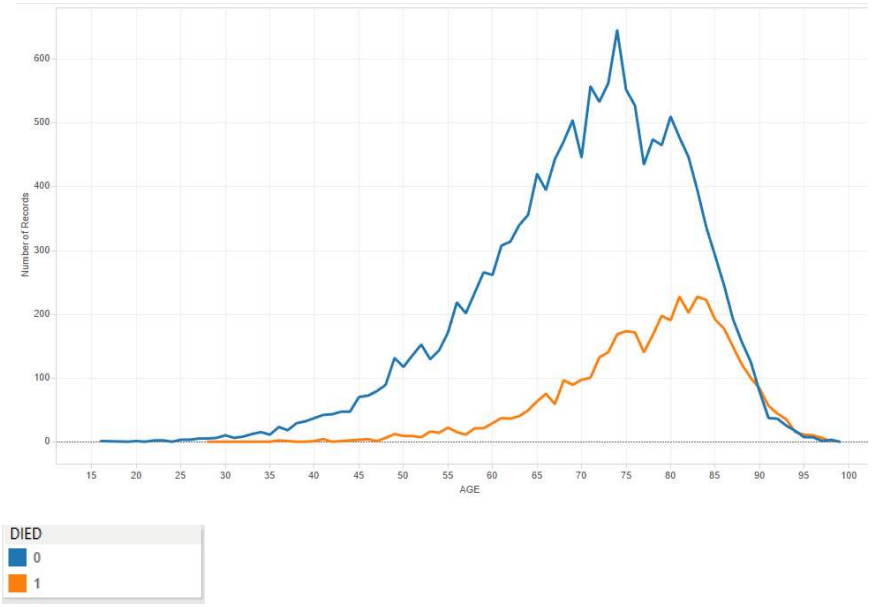

Figure 7.1.2 Age Distribution Plot

Appendix

Table 1 Variable description

Randomisation data	
HOSPNUM	Hospital number
RDELAY	Delay between stroke and randomisation in hours
RCONSC	Conscious state at randomisation (F - fully alert, D - drowsy, U - unconscious)
SEX	M = male; F = female
AGE	Age in years
RSLEEP	Symptoms noted on waking (Y/N)
RATRIAL	Atrial fibrillation (Y/N); not coded for pilot phase - 984 patients
RCT	CT before randomisation (Y/N)
RVISINF	Infarct visible on CT (Y/N)
RHEP24	Heparin within 24 hours prior to randomisation (Y/N)
RASP3	Aspirin within 3 days prior to randomisation (Y/N)
RSBP	Systolic blood pressure at randomisation (mmHg)
RDEF1	Face deficit (Y/N/C=can't assess)
RDEF2	Arm/hand deficit (Y/N/C=can't assess)

RDEF3	Leg/foot deficit (Y/N/C=can't assess)
RDEF4	Dysphasia (Y/N/C=can't assess)
RDEF5	Hemianopia (Y/N/C=can't assess)
RDEF6	Visuospatial disorder (Y/N/C=can't assess)
RDEF7	Brainstem/cerebellar signs (Y/N/C=can't assess)
RDEF8	Other deficit (Y/N/C=can't assess)
STYPE	Stroke subtype (TACS/PACS/POCS/LACS/OTH=other)
RDATE	Year and month of randomisation (yyyy-mm)
HOURLOCAL	Local time - hours (99-missing data) of randomisation
MINLOCAL	Local time - minutes (99-missing data) of randomisation
DAYLOCAL	Estimate of local day of week: 1 - Sunday, 2-Monday, 3-Tuesday, 4-Wednesday, 5-Thursday, 6-Friday, 7-Saturday
RXASP	Trial aspirin allocated (Y/N)

Data collected on 14 day/discharge form about treatments given in hospital

DASP14	Aspirin given for 14 days or till death or discharge (Y/N/U=unknown)
DASPLT	Discharged on long term aspirin (Y/N/U=unknown)
DLH14	Low dose heparin given for 14 days or till death/discharge (Y/N/U=unknown)
DMH14	Medium dose heparin given for 14 days or till death/discharge (Y/N/U=unknown)
DHH14	Medium dose heparin given for 14 days etc in pilot (combine with above; Y/N)
ONDRUG	Estimate of time in days on trial treatment
DSCH	Non trial subcutaneous heparin (Y/N/U=unknown)
DIVH	Non trial intravenous heparin (Y/N/U=unknown)
DAP	Non trial antiplatelet drug (Y/N/U=unknown)
DOAC	Other anticoagulants (Y/N/U=unknown)
DGORM	Glycerol or manitol (Y/N/U=unknown)
DSTER	Steroids (Y/N/U=unknown)
DCAA	Calcium antagonists (Y/N/U=unknown)

DHAEMD	Haemodilution (Y/N/U=unknown)
DCAREND	Carotid surgery (Y/N/U=unknown)
DTHROMB	Thrombolysis (Y/N/U=unknown)
DMAJNCH	Major non-cerebral haemorrhage (Y/N/U=unknown)
DMAJNCHD	Date of above (days elapsed from randomisation)
DMAJNCHX	Comment on above
DSIDE	Other side effect (Y/N/U=unknown)
DSIDED	Date of above (days elapsed from randomisation)
DSIDEX	Comment on above
Final diagnosis of initial event	
DDIAGISC	Ischaemic stroke (Y/N/U=unknown)
DDIAGHA	Haemorrhagic stroke (Y/N/U=unknown)
DDIAGUN	Indeterminate stroke (Y/N/U=unknown)

DNOSTRK	Not a stroke (Y/N/U=unknown)
DNOSTRKX	Comment on above
Recurrent stroke within 14 days	
DRSISC	Ischaemic recurrent stroke (Y/N/U=unknown)
DRSISCD	Date of above (days elapsed from randomisation)
DRSH	Haemorrhagic stroke (Y/N/U=unknown)
DRSHD	Date of above (days elapsed from randomisation)
DRSUNK	Unknown type (Y/N/U=unknown)
DRSUNKD	Date of above (days elapsed from randomisation)
Other events within 14 days	
DPE	Pulmonary embolism; (Y/N/U=unknown)
DPED	Date of above (days elapsed from randomisation)
DALIVE	Discharged alive from hospital (Y/N/U=unknown)

Table2

Final diagnosis of initial event.

	Number
Ischaemic stroke	17398
Haemorrhagic stroke	599
Definite stroke, pathological type unknown	992
Not a stroke	420
Uncertain diagnosis	26
Total	19435

Reference

1. An Introduction to DATA MINING by Daniel T. Larose.

2 Data comes from The International Stroke Trial database.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104487/