ISE465 HW#3 Report

Bolun Xu ***box215@lehigh.edu***

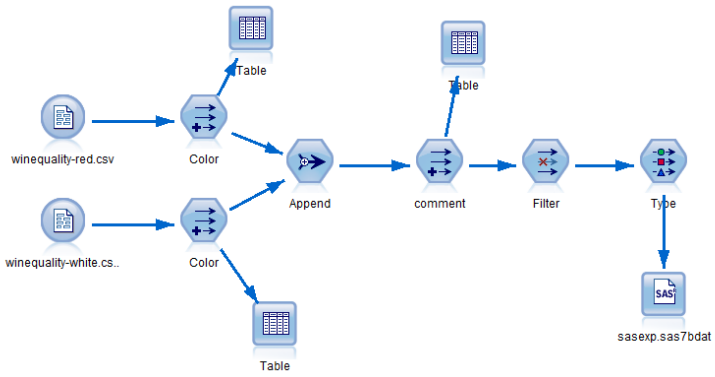
2.A:

(1). Use IBM SPSS to make the data fit our requirement.

i) Derive node is used to create a new variable 'Color'. First data set is red, and second is white.

ii) Append node combines the two given data sets. Then used another Derive node to create the new variable 'Comment', set quality>=7 to be T and else to be F.

iii) Import the new data set as SAS file.



(2). Use SAS Enterprise Miner to model.

i) Use SAS Code to see the correlation of variables. We get this :

Correlation Coefficients |r| > 0.5: Density, alcohol; residual_sugar, Density; free_sulfur_dioxide, total_sulfur_dioxide;

So we drop Density and free_sulfur_dioxide.

ii) Partition data into 70/30 training/validation data set and used decision tree to establish model.

iii) Model types: 3 Common different types of nominal criterion; 3 decision trees After dropping the high correlative variables; 3 decision trees after Variable Selection node to eliminated variables that not highly correlated to the target. Total 9 models.

iv) Comparison node is used to compare the decision trees described above.

3.A:

From StatExplore Node, through the high correlation, we can find alcohol and sugar may partly decide the density of wine.

From the results of Graphboard node in SPSS and MultiPlot Node in Enterprise Miner, we get proper relationship between variables and the target:

High alcohol ~ high quality; Low chlorides ~ high quality; High or low pH ~ low quality; low volatile acidity wine ~ high quality; Low residual sugar ~ high quality; Other variables are not quite related.

Our goal is to predict which wine will be of high quality. Thus, high-quality wine is what we care about. We we should use precision to evaluate the model, meanwhile we need keep accuracy in an acceptable range. From the chart/table below we can see the precision, sensitivity, specificity and accuracy are shown below.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree3	Ginil	TRAIN	comment		612	3530	127	279
Tree3	Ginil	VALIDATE	comment		271	1505	58	115
Tree2	Entropy1	TRAIN	comment		694	3585	72	197
Tree2	Entropy1	VALIDATE	comment		298	1524	39	88
Tree	Prob1	TRAIN	comment		688	3564	93	203
Tree	Prob1	VALIDATE	comment		302	1518	45	84
Tree5	Prob2	TRAIN	comment		596	3506	151	295
Tree5	Prob2	VALIDATE	comment		270	1491	72	116
Tree4	Entropy2	TRAIN	comment		553	3501	156	338
Tree4	Entropy2	VALIDATE	comment		242	1485	78	144
Tree6	Gini2	TRAIN	comment		599	3537	120	292
Tree6	Gini2	VALIDATE	comment		275	1506	57	111
Tree7	Prob3	TRAIN	comment		731	3558	99	160
Tree7	Prob3	VALIDATE	comment		320	1524	39	66
Tree8	Entropy3	TRAIN	comment		743	3596	61	148
Tree8	Entropy3	VALIDATE	comment		317	1538	25	69
Tree9	Gini3	TRAIN	comment		730	3570	87	161
Tree9	Gini3	VALIDATE	comment		319	1526	37	67

Comparison Node result with 9 models ↑

Model description	FN	TN	FP	TP	sensitivity	specificity	precision	accuracy
Gini1	271	1505	58	115	0.2979275	0.9628919	0.66474	**0.831195**
Entropy1	298	1524	39	88	0.2279793	0.975048	0.692913	0.827091
Prob1	302	1518	45	84	0.2176166	0.9712092	0.651163	0.82196
Prob2	270	1491	72	116	**0.3005181**	0.9539347	0.617021	0.824525
Entropy2	242	1485	78	14	0.0546875	0.950096	0.152174	0.824079
Gini2	275	1506	57	111	0.2875648	0.9635317	0.660714	0.829656
Prob3	320	1524	39	66	0.1709845	0.975048	0.628571	0.815803
Entropy3	317	1538	25	69	0.1787565	**0.9840051**	**0.734043**	0.824525
Gini3	319	1526	37	67	0.1735751	0.9763276	0.644231	0.817342

The table above shows that the accuracy of all decision trees are similar to each other, while the Gini1 decision tree built based on manual dropped data has the highest accuracy. Yet I think Entropy with variable selection is the best. It has the highest precision value and its accuracy is also acceptable (>80%). The result shows that by using this model (Entropy decision tree with variable selection), about 68% of the wine that I predicted to be high quality are true high quality wine. With highest sensitivity model above, its precision is much lower. That shows the correctness of the high quality wine prediction is worse than the chosen one.