Wine Data SAS Analysis: Rita Dicarlo, Katie Clewett, Chang Guo
Classification

The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

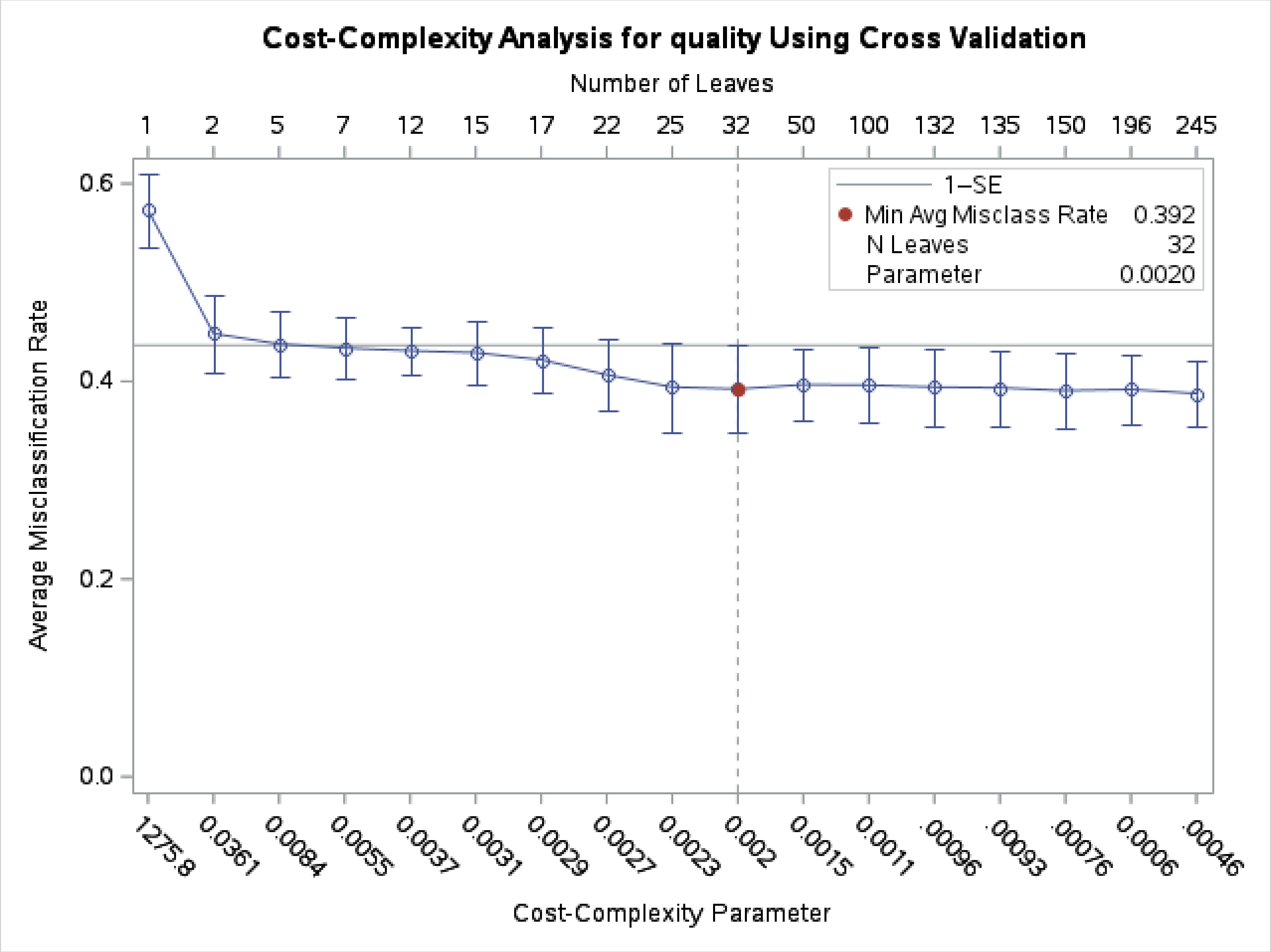
Data Access Information			
Data	**Engine**	**Role**	**Path**
WORK.WINE	V9	Input	On Client

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	9
Number of Leaves Before Pruning	274
Number of Leaves After Pruning	43

Number of Observations Read	1599
Number of Observations Used	1599

Wine Data SAS Analysis: Rita Dicarlo, Katie Clewett, Chang Guo
Classification

The HPSPLIT Procedure



10-Fold Cross Validation Assessment of Model

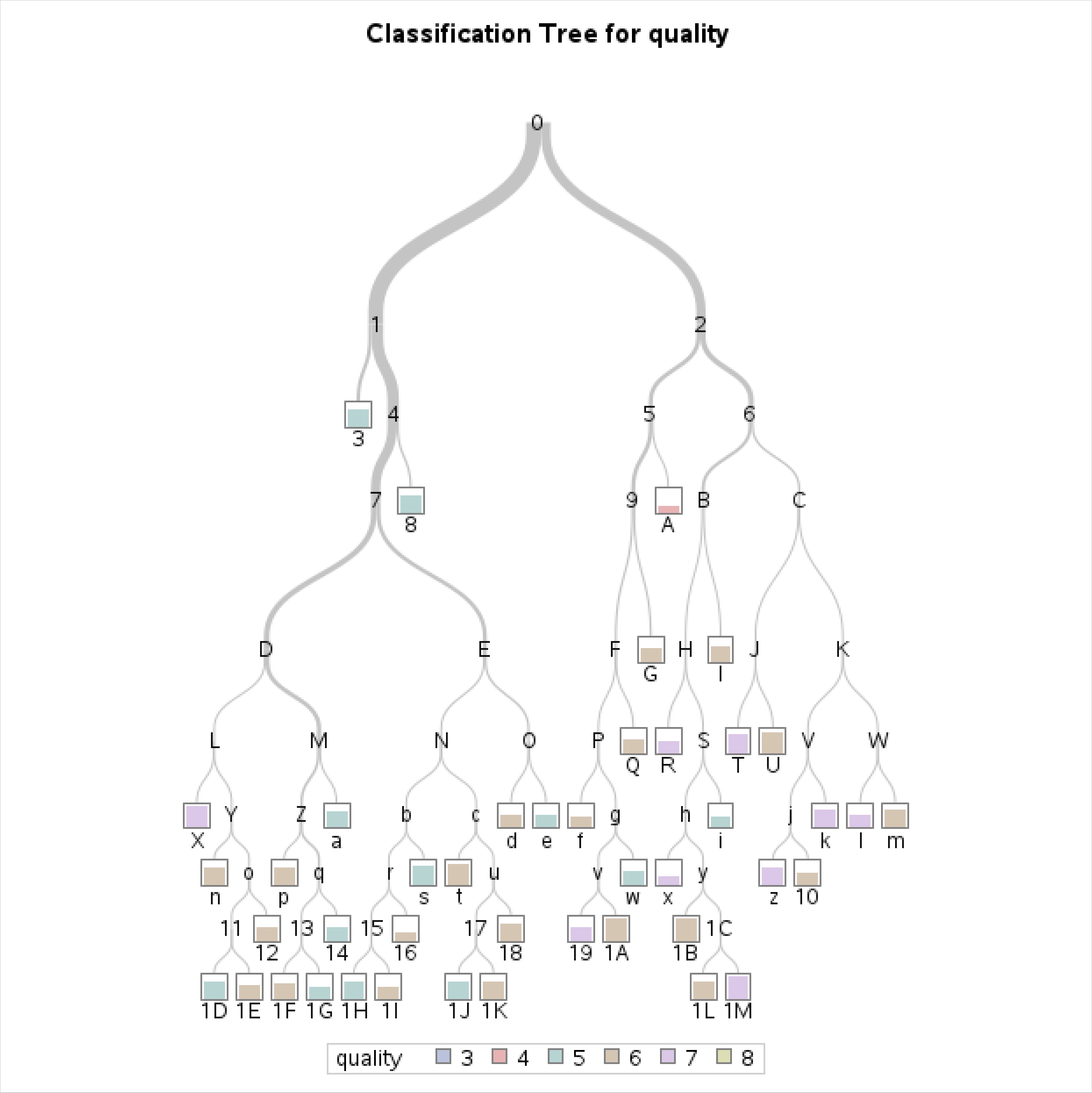
	Average Square Error				Number of Leaves			Misclassification Rate			
N Leaves	Min	Avg	Standard Error	Max	Min	Median	Max	Min	Avg	Standard Error	Max
34	0.0813	0.0918	0.00810	0.1088	29	34.5	53	0.3224	0.3870	0.0406	0.4882

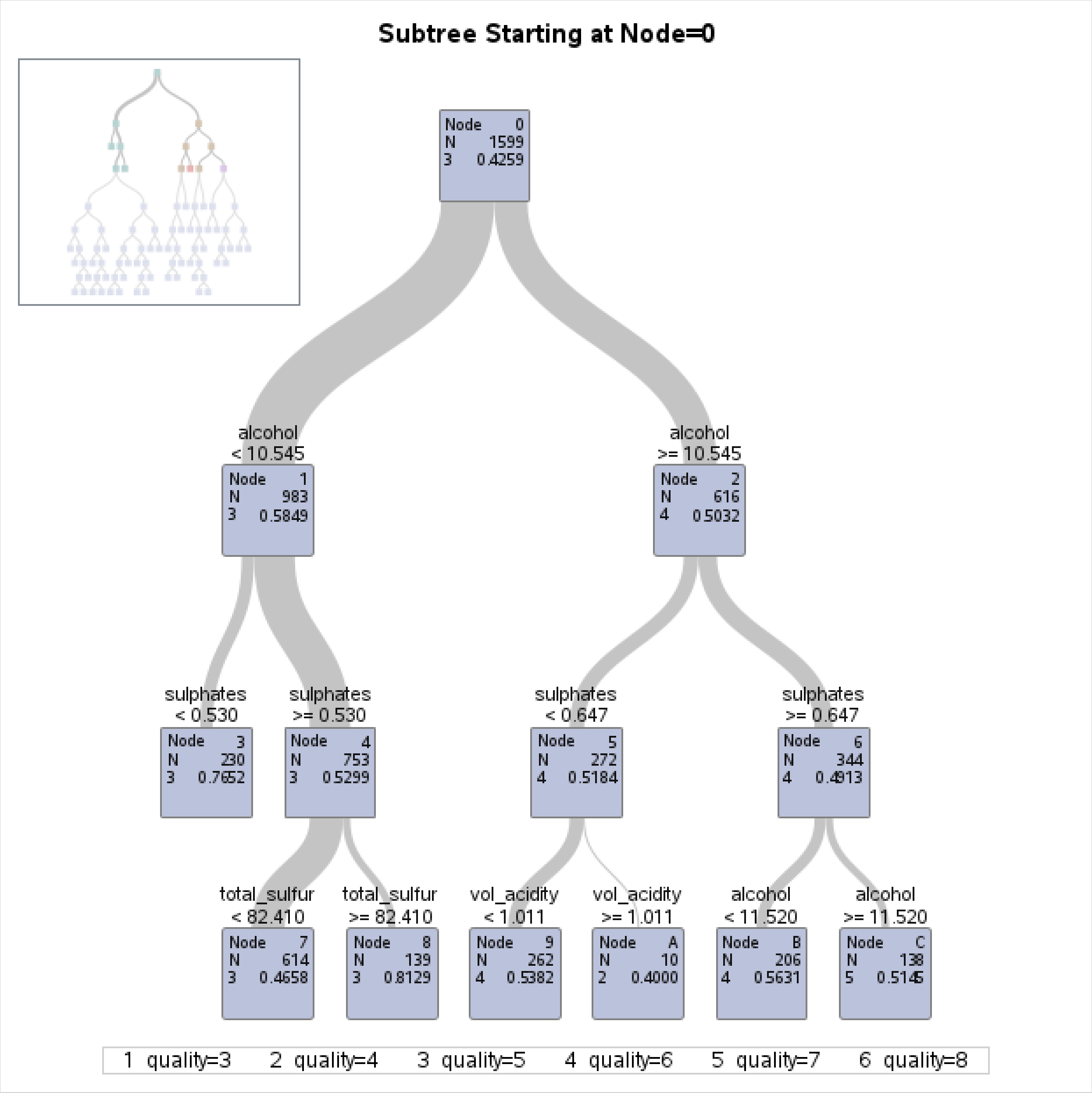
10-Fold Cross Validation Confusion Matrix

	Predicted						
Actual	3	4	5	6	7	8	Error Rate
3	1	3	3	2	1	0	0.9000
4	2	1	31	19	0	0	0.9811
5	1	2	507	165	6	0	0.2555
6	0	1	202	390	45	0	0.3887
7	0	1	17	99	81	1	0.5930
8	0	0	0	7	11	0	1.0000

Wine Data SAS Analysis: Rita Dicarlo, Katie Clewett, Chang Guo
Classification

The HPSPLIT Procedure


Classification Tree for quality
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
10
11
12
13
14
15
16
17
18
19
1A
1B
1C
1D
1E
1F
1G
1H
1I
1J
1K
1L
1M
quality
3
4
5
6
7
8



Wine Data SAS Analysis: Rita Dicarlo, Katie Clewett, Chang Guo
Classification

The HPSPLIT Procedure

Confusion Matrices

		Predicted						Error
	Actual	3	4	5	6	7	8	Rate
Model Based	3	0	3	4	3	0	0	1.0000
	4	0	4	37	12	0	0	0.9245
	5	0	3	542	126	10	0	0.2041
	6	0	0	143	468	27	0	0.2665
	7	0	0	10	85	104	0	0.4774
	8	0	0	0	10	8	0	1.0000
Cross Validation	3	1	3	3	2	1	0	0.9000
	4	2	1	31	19	0	0	0.9811
	5	1	2	507	165	6	0	0.2555
	6	0	1	202	390	45	0	0.3887

Confusion Matrices								
		Predicted						Error
	Actual	3	4	5	6	7	8	Rate
	7	0	1	17	99	81	1	0.5930
	8	0	0	0	7	11	0	1.0000

Fit Statistics for Selected Tree						
	N Leaves	ASE	Mis-class	Entropy	Gini	RSS
Model Based	43	0.0727	0.3008	1.1063	0.4363	697.7
Cross Validation	34	0.0918	0.3870			

Variable Importance			
	Training		
Variable	Relative	Importance	Count
alcohol	1.0000	11.7329	8
sulphates	0.6146	7.2115	7
total_sulfur	0.5678	6.6624	7
vol_acidity	0.5534	6.4936	9
free_sulfur	0.4257	4.9947	5
pH	0.4091	4.7998	5
chlorides	0.2038	2.3910	1

Wine Data SAS Analysis: Rita Dicarlo, Katie Clewett, Chang Guo
Classification

The HPFOREST Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.WINE	V9	Input	On Client

Model Information		
Parameter	Value	
Variables to Try	3	(Default)
Maximum Trees	100	
Actual Trees	100	
Inbag Fraction	0.3	
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Leaf Fraction	0.00001	(Default)
Leaf Size Setting	1	(Default)
Leaf Size Used	1	
Category Bins	30	(Default)
Interval Bins	100	
Minimum Category Size	5	(Default)

Model Information		
Parameter	**Value**	
Node Size	100000	(Default)
Maximum Depth	20	(Default)
Alpha	1	(Default)
Exhaustive	5000	(Default)
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Gini
Preselection Method	.	BinnedSearch
Missing Value Handling	.	Valid value

Number of Observations	
Type	**N**
Number of Observations Read	1599
Number of Observations Used	1599

Baseline Fit Statistics	
Statistic	**Value**
Average Square Error	0.107
Misclassification Rate	0.574
Log Loss	1.185

Fit Statistics

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (OOB)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Log Loss (Train)	Log Loss (OOB)
1	170	0.1059	0.1512	0.3177	0.454	7.315	10.44
2	338	0.0697	0.1248	0.2996	0.435	3.270	7.90
3	510	0.0587	0.1102	0.2364	0.419	1.832	6.15
4	680	0.0521	0.0996	0.2095	0.407	1.233	4.70
5	853	0.0488	0.0942	0.1870	0.398	0.876	3.98
6	1018	0.0464	0.0903	0.1795	0.381	0.738	3.38
7	1192	0.0453	0.0885	0.1595	0.386	0.644	2.91
8	1370	0.0440	0.0868	0.1595	0.387	0.536	2.64
9	1546	0.0427	0.0852	0.1457	0.378	0.493	2.39
10	1720	0.0423	0.0840	0.1401	0.372	0.491	2.25
11	1885	0.0417	0.0825	0.1426	0.366	0.461	2.04
12	2058	0.0413	0.0824	0.1338	0.367	0.459	2.01
13	2230	0.0409	0.0820	0.1382	0.360	0.445	1.96
14	2415	0.0404	0.0814	0.1295	0.357	0.442	1.92
15	2589	0.0402	0.0812	0.1238	0.356	0.441	1.84
16	2751	0.0401	0.0807	0.1238	0.360	0.441	1.75
17	2923	0.0399	0.0802	0.1232	0.353	0.439	1.73
18	3086	0.0398	0.0800	0.1257	0.357	0.440	1.70
19	3270	0.0397	0.0800	0.1232	0.354	0.440	1.67
20	3431	0.0395	0.0796	0.1207	0.354	0.439	1.63
21	3600	0.0395	0.0797	0.1176	0.352	0.440	1.59
22	3759	0.0394	0.0795	0.1238	0.355	0.439	1.56
23	3936	0.0392	0.0791	0.1151	0.355	0.439	1.55
24	4095	0.0392	0.0790	0.1213	0.355	0.438	1.50

Fit Statistics

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (OOB)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Log Loss (Train)	Log Loss (OOB)
25	4265	0.0390	0.0787	0.1119	0.350	0.439	1.47
26	4456	0.0389	0.0786	0.1151	0.350	0.438	1.43
27	4631	0.0386	0.0781	0.1151	0.343	0.437	1.41
28	4793	0.0386	0.0780	0.1169	0.342	0.437	1.38
29	4958	0.0385	0.0777	0.1182	0.340	0.436	1.37
30	5125	0.0384	0.0774	0.1163	0.346	0.435	1.35
31	5301	0.0383	0.0774	0.1132	0.346	0.436	1.34
32	5482	0.0382	0.0772	0.1144	0.343	0.435	1.31
33	5656	0.0381	0.0770	0.1101	0.343	0.435	1.30
34	5833	0.0381	0.0768	0.1113	0.338	0.435	1.29
35	6004	0.0380	0.0767	0.1107	0.334	0.434	1.27
36	6167	0.0380	0.0767	0.1069	0.331	0.434	1.25
37	6339	0.0380	0.0768	0.1094	0.336	0.434	1.25
38	6503	0.0379	0.0766	0.1057	0.337	0.434	1.24
39	6685	0.0378	0.0764	0.1044	0.337	0.434	1.23
40	6843	0.0378	0.0763	0.1069	0.336	0.434	1.22
41	7012	0.0377	0.0763	0.1063	0.338	0.434	1.22
42	7190	0.0377	0.0762	0.1063	0.335	0.434	1.22
43	7367	0.0375	0.0760	0.1026	0.334	0.432	1.22
44	7525	0.0375	0.0760	0.1051	0.335	0.432	1.21
45	7689	0.0375	0.0759	0.1051	0.336	0.432	1.21
46	7867	0.0375	0.0758	0.1063	0.336	0.432	1.21
47	8041	0.0375	0.0758	0.1038	0.335	0.432	1.21
48	8207	0.0375	0.0760	0.1069	0.340	0.433	1.21
49	8363	0.0375	0.0759	0.1026	0.338	0.433	1.18
50	8533	0.0374	0.0758	0.1044	0.335	0.432	1.17
51	8700	0.0374	0.0758	0.1057	0.334	0.432	1.17
52	8870	0.0374	0.0757	0.1026	0.332	0.432	1.17
53	9036	0.0373	0.0756	0.1038	0.331	0.431	1.16
54	9201	0.0374	0.0756	0.1019	0.333	0.432	1.13
55	9368	0.0373	0.0756	0.1051	0.333	0.432	1.12
56	9539	0.0373	0.0755	0.1019	0.333	0.432	1.12
57	9715	0.0373	0.0756	0.1019	0.332	0.432	1.12
58	9878	0.0373	0.0756	0.0994	0.331	0.432	1.11
59	10045	0.0373	0.0756	0.1019	0.333	0.432	1.11
60	10219	0.0373	0.0755	0.1007	0.330	0.432	1.11
61	10374	0.0373	0.0755	0.1001	0.333	0.433	1.11
62	10533	0.0373	0.0755	0.1013	0.331	0.432	1.11
63	10708	0.0373	0.0755	0.1057	0.331	0.433	1.11
64	10883	0.0373	0.0755	0.1038	0.331	0.432	1.11
65	11061	0.0373	0.0754	0.1032	0.328	0.432	1.11
66	11229	0.0372	0.0754	0.1019	0.328	0.432	1.11
67	11395	0.0372	0.0753	0.1001	0.330	0.432	1.11
68	11572	0.0372	0.0753	0.0976	0.330	0.432	1.10
69	11746	0.0371	0.0753	0.0982	0.325	0.432	1.08
70	11922	0.0371	0.0753	0.1001	0.329	0.432	1.07

Fit Statistics							
Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (OOB)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Log Loss (Train)	Log Loss (OOB)
71	12093	0.0371	0.0753	0.0976	0.327	0.432	1.07
72	12254	0.0371	0.0754	0.0938	0.327	0.433	1.07
73	12421	0.0371	0.0753	0.0932	0.326	0.432	1.07
74	12582	0.0371	0.0753	0.0944	0.330	0.432	1.07
75	12748	0.0371	0.0754	0.0976	0.330	0.433	1.08
76	12916	0.0371	0.0753	0.0969	0.327	0.433	1.08
77	13082	0.0371	0.0753	0.0988	0.325	0.433	1.08
78	13253	0.0371	0.0753	0.0969	0.325	0.433	1.08
79	13424	0.0371	0.0752	0.0988	0.325	0.433	1.08
80	13586	0.0370	0.0752	0.0976	0.326	0.432	1.08
81	13771	0.0370	0.0751	0.0988	0.326	0.432	1.08
82	13936	0.0370	0.0751	0.0982	0.328	0.432	1.06
83	14088	0.0370	0.0751	0.0976	0.326	0.433	1.06
84	14261	0.0370	0.0750	0.0951	0.329	0.432	1.06
85	14429	0.0370	0.0750	0.0988	0.329	0.432	1.06
86	14597	0.0370	0.0750	0.1001	0.326	0.432	1.06
87	14771	0.0369	0.0749	0.0994	0.327	0.432	1.06
88	14948	0.0369	0.0749	0.1001	0.327	0.432	1.06
89	15125	0.0369	0.0749	0.0994	0.325	0.432	1.06
90	15294	0.0369	0.0750	0.0988	0.326	0.432	1.05
91	15462	0.0369	0.0749	0.0988	0.328	0.432	1.05
92	15632	0.0368	0.0748	0.1001	0.329	0.432	1.05
93	15799	0.0368	0.0748	0.0988	0.330	0.432	1.05
94	15972	0.0368	0.0749	0.1013	0.329	0.432	1.05
95	16148	0.0368	0.0749	0.1001	0.329	0.432	1.05
96	16315	0.0368	0.0749	0.1007	0.329	0.432	1.05
97	16482	0.0369	0.0749	0.1001	0.331	0.432	1.03
98	16646	0.0369	0.0749	0.1013	0.328	0.432	1.03
99	16814	0.0369	0.0750	0.1026	0.329	0.432	1.03
100	16994	0.0368	0.0749	0.0994	0.326	0.432	1.02

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
alcohol	3524	0.157763	-0.01210	0.238010	0.06514
chlorides	1077	0.036705	-0.02801	0.063058	-0.00104
vol_acidity	1831	0.080796	-0.03610	0.124482	0.01138
total_sulfur	2509	0.096738	-0.04559	0.167776	0.02654
free_sulfur	1781	0.060638	-0.04641	0.106865	0.00134
sulphates	3296	0.117046	-0.04851	0.195679	0.02624
pH	2876	0.090969	-0.07045	0.159729	-0.00283

High Quality Wine Classification

The LOGISTIC Procedure

Model Information

Data Set	WORK.WINE
Response Variable	high_quality
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1599
Number of Observations Used	1599

Response Profile

Ordered Value	high_quality	Total Frequency
1	0	1382
2	1	217

Probability modeled is high_quality=0.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1271.921	981.754
SC	1277.298	997.886
-2 Log L	1269.921	975.754

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	294.1666	2	<.0001
Score	307.2879	2	<.0001
Wald	212.6273	2	<.0001

Analysis of Maximum Likelihood Estimates

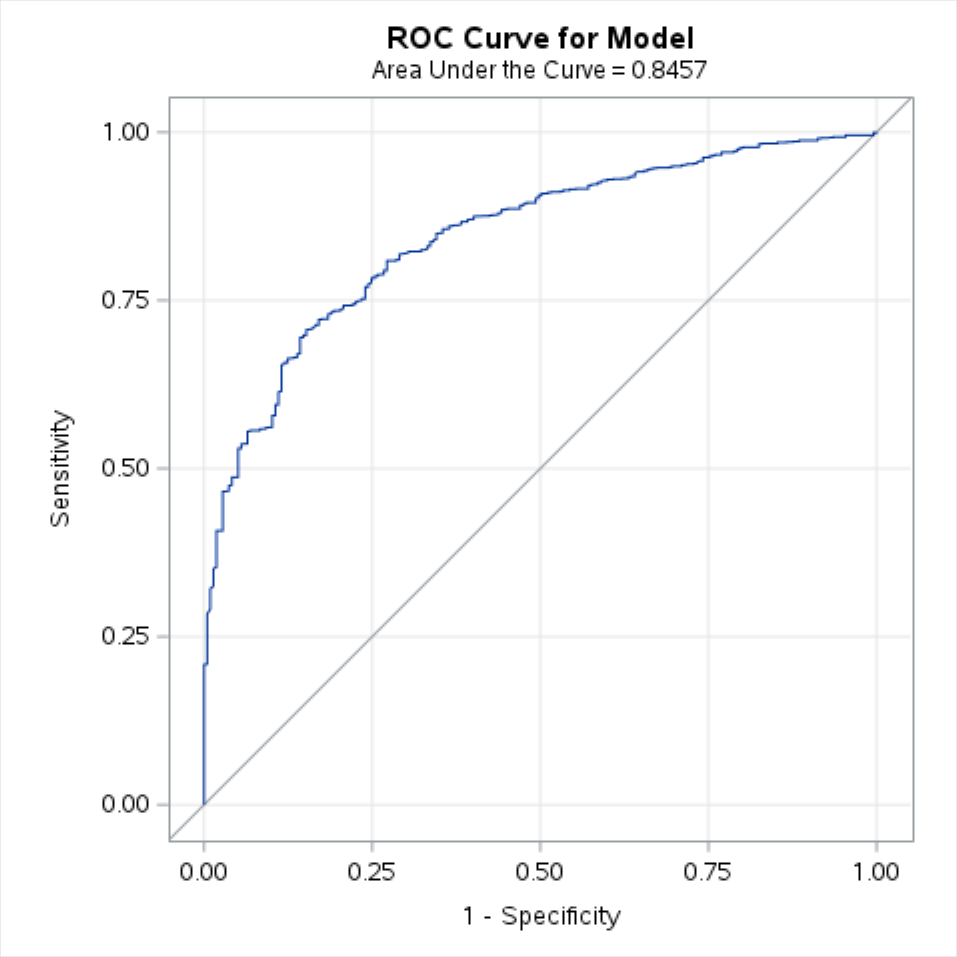
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.8598	0.9942	254.4881	<.0001
alcohol	1	-1.0903	0.0792	189.7042	<.0001
sulphates	1	-3.1444	0.4262	54.4441	<.0001

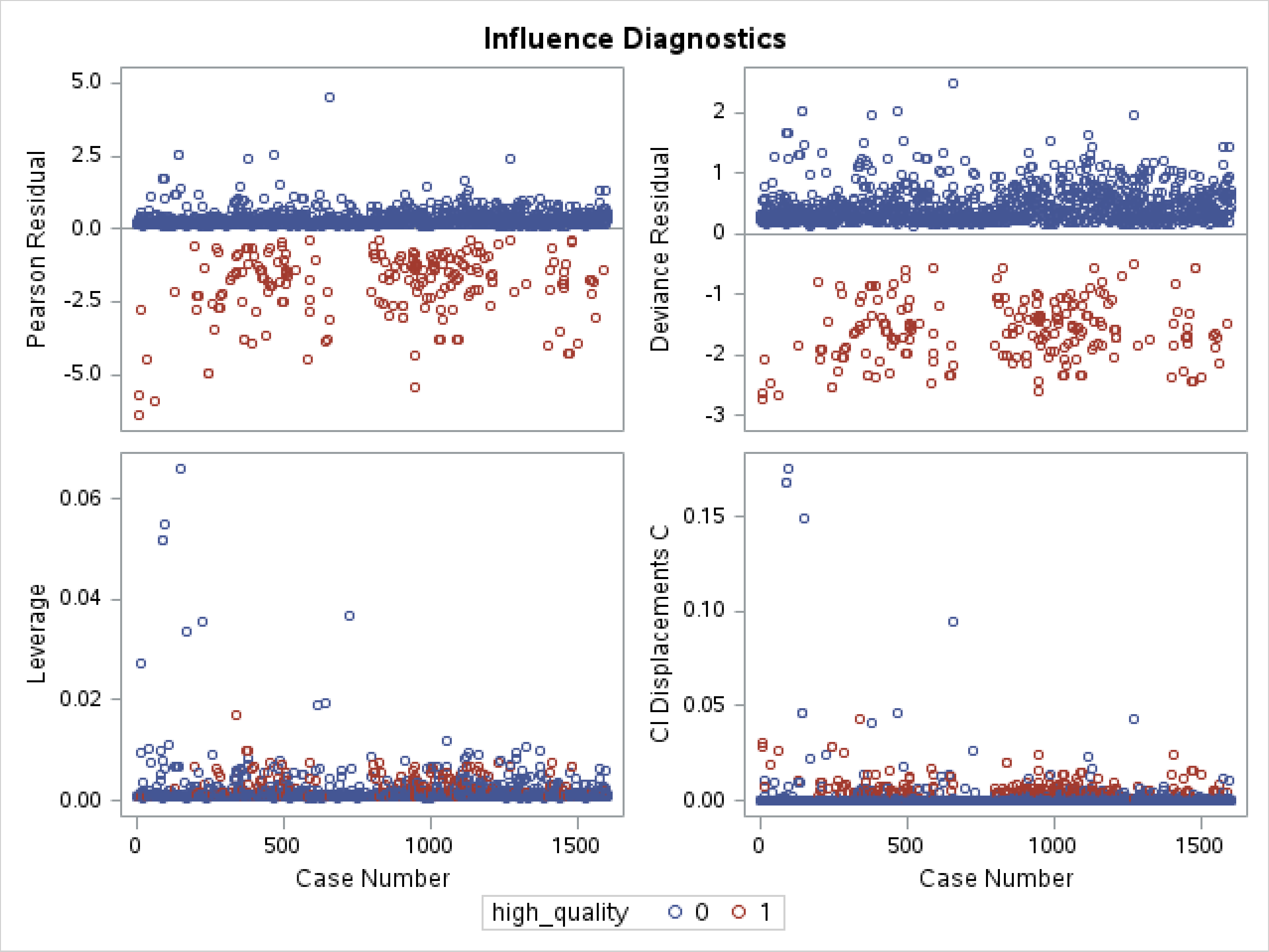
Odds Ratio Estimates

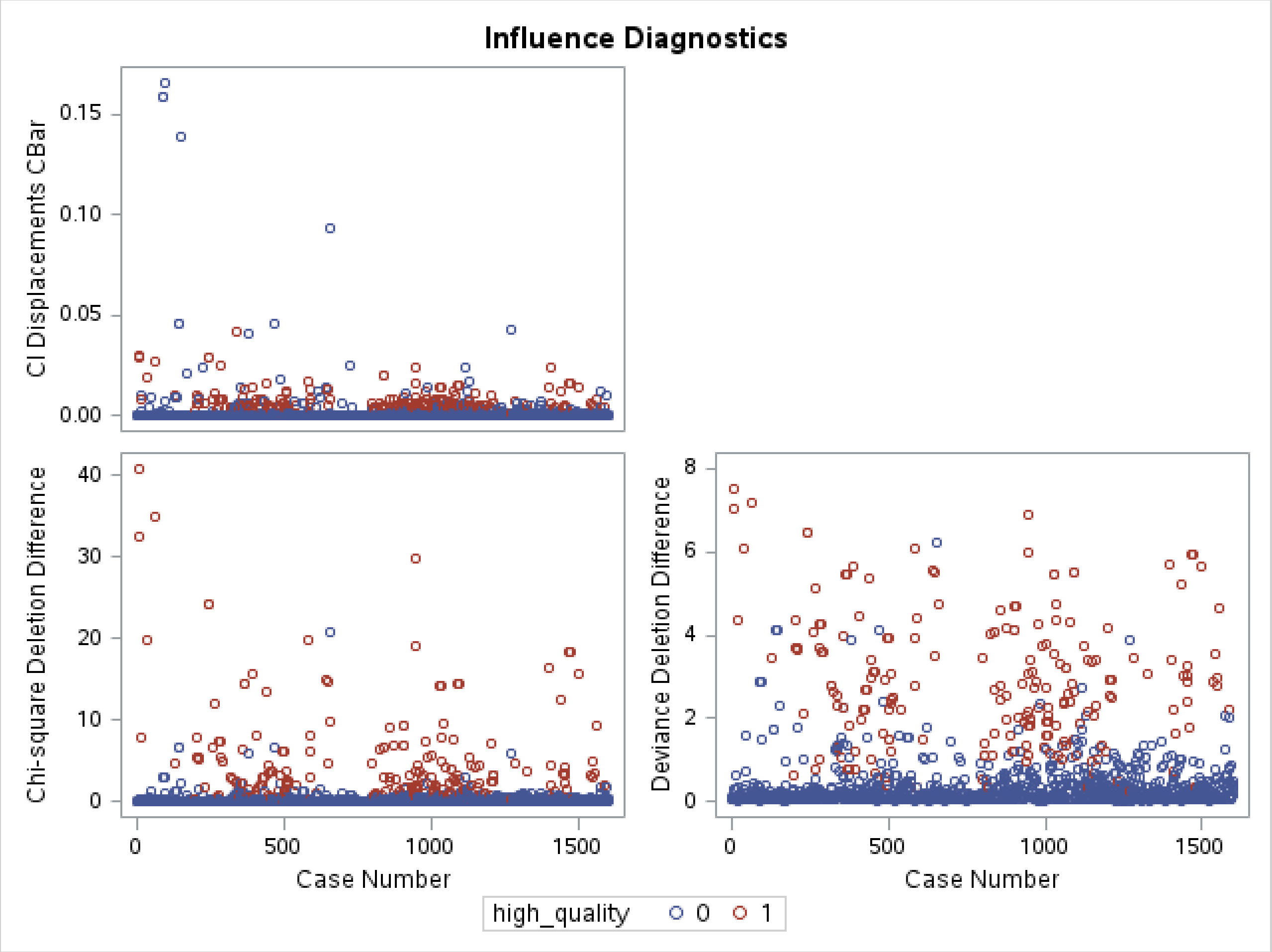
Effect	Point Estimate	95% Wald Confidence Limits	
alcohol	0.336	0.288	0.393
sulphates	0.043	0.019	0.099

Association of Predicted Probabilities and Observed Responses

Percent Concordant	84.6	Somers' D	0.691
Percent Discordant	15.4	Gamma	0.692
Percent Tied	0.0	Tau-a	0.162
Pairs	299894	c	0.846


ROC Curve for Model
Area Under the Curve = 0.8457
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity


Influence Diagnostics
Pearson Residual
5.0
2.5
0.0
-2.5
-5.0
Deviance Residual
2
1
0
-1
-2
-3
Leverage
0.06
0.04
0.02
0.00
CI Displacements C
0.15
0.10
0.05
0.00
0
500
1000
1500
Case Number
0
500
1000
1500
Case Number
high_quality ○ 0 ○ 1



High Quality Wine Classification

The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

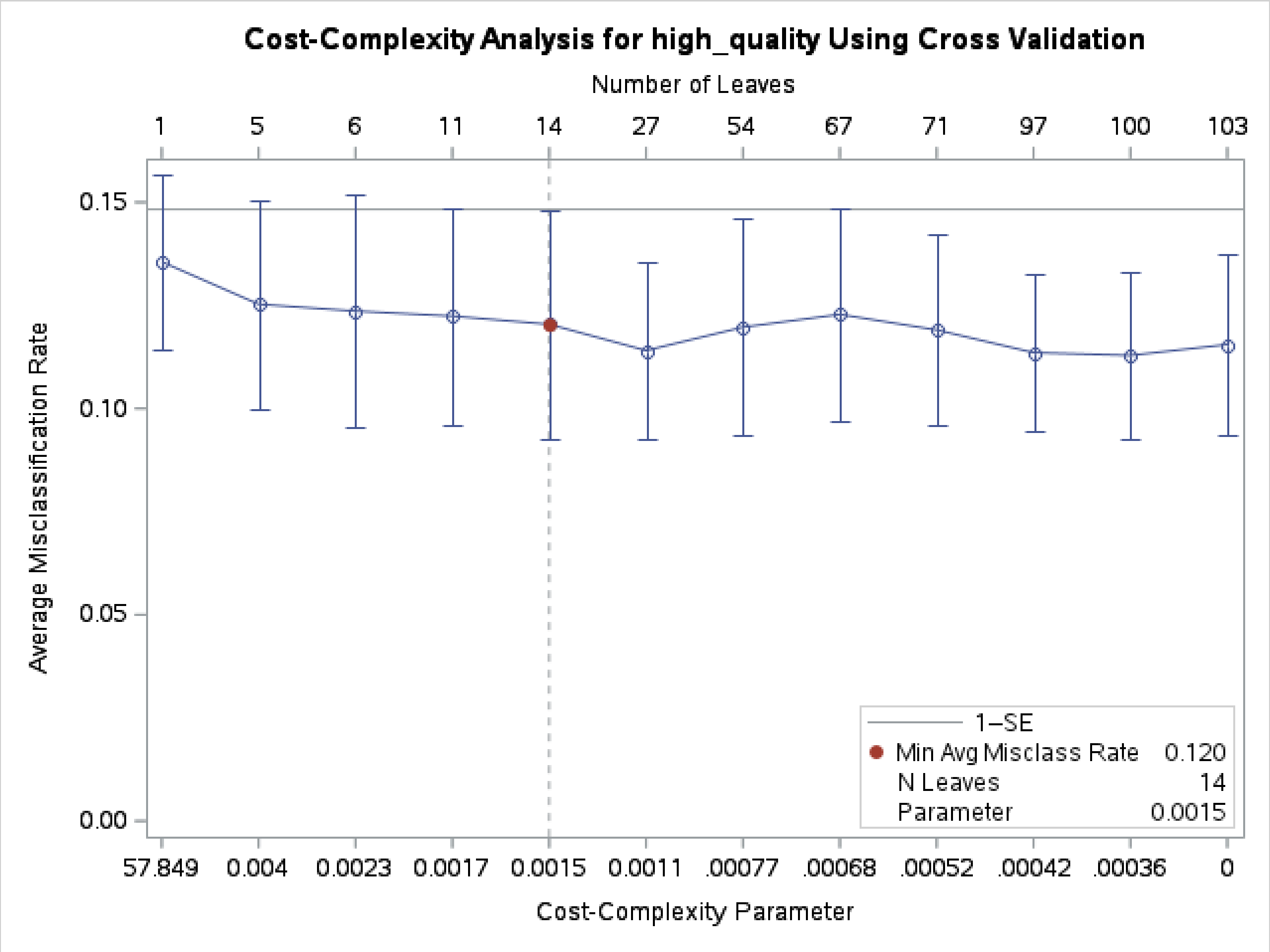
Data Access Information			
Data	Engine	Role	Path
WORK.WINE	V9	Input	On Client

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	7
Number of Leaves Before Pruning	114
Number of Leaves After Pruning	15
Model Event Level	0

Number of Observations Read	1599
Number of Observations Used	1599

High Quality Wine Classification

The HPSPLIT Procedure



10-Fold Cross Validation Assessment of Model

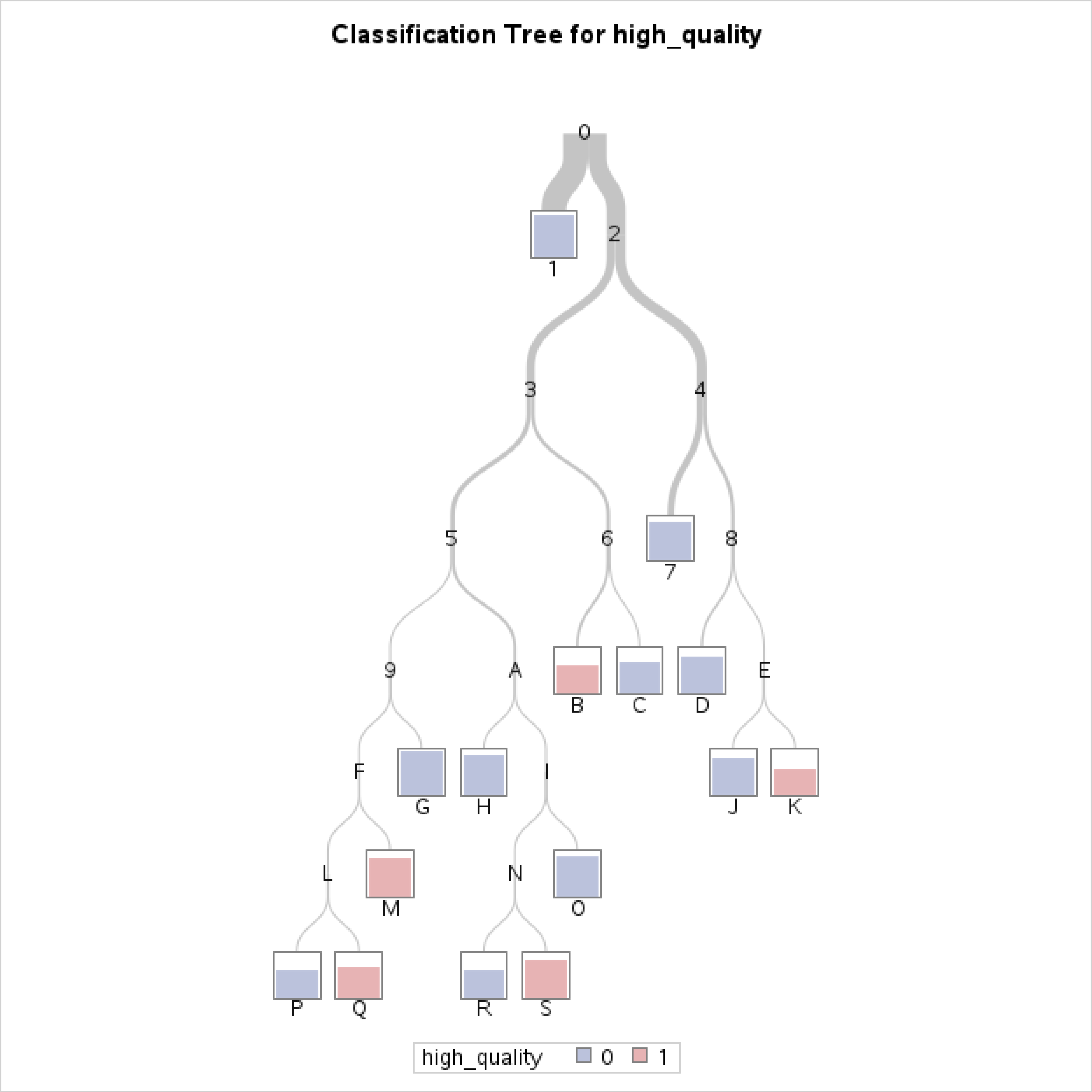
	Average Square Error				Number of Leaves			Misclassification Rate			
N Leaves	Min	Avg	Standard Error	Max	Min	Median	Max	Min	Avg	Standard Error	Max
14	0.0714	0.0948	0.0200	0.1387	7	14.5	24	0.0909	0.1216	0.0249	0.1694

10-Fold Cross Validation Confusion Matrix

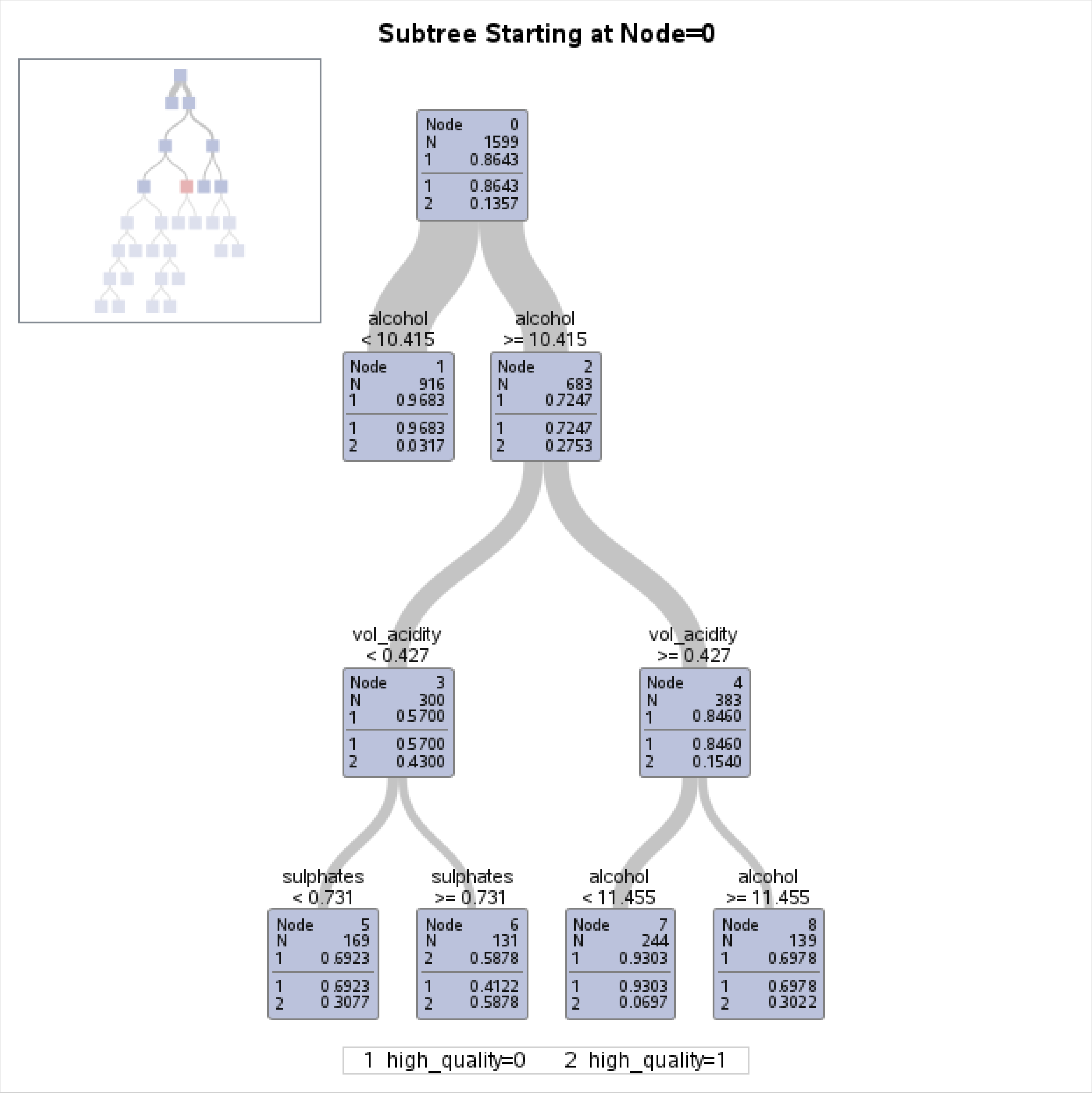
	Predicted		
Actual	0	1	Error Rate
0	1315	67	0.0485
1	130	87	0.5991

High Quality Wine Classification

The HPSPLIT Procedure



Classification Tree for high_quality



High Quality Wine Classification

The HPSPLIT Procedure

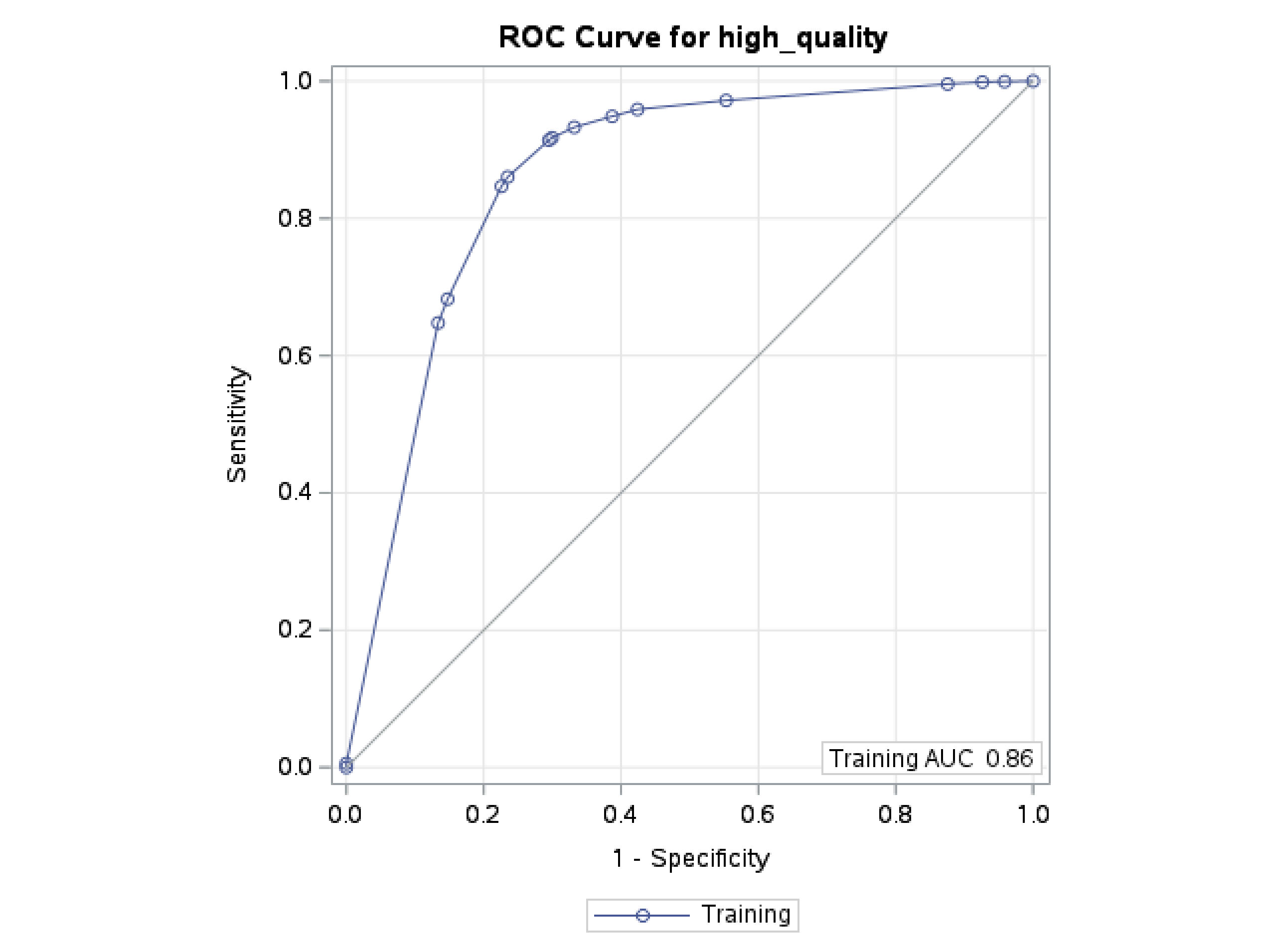
Confusion Matrices

		Predicted		Error Rate
	Actual	**0**	**1**	
Model Based	**0**	1325	57	0.0412
	1	92	125	0.4240
Cross Validation	**0**	1315	67	0.0485
	1	130	87	0.5991

Fit Statistics for Selected Tree

	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Model Based	15	0.0729	0.0932	0.9588	0.5760	0.3702	0.1459	233.2	0.8611

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Cross Validation	14	0.0948	0.1216	0.9515	0.4009				



Variable Importance			
	Training		
Variable	Relative	Importance	Count
alcohol	1.0000	7.9124	4
sulphates	0.6735	5.3289	4
vol_acidity	0.6397	5.0617	1
total_sulfur	0.4852	3.8394	2
chlorides	0.2784	2.2029	1
pH	0.2388	1.8898	1
free_sulfur	0.1820	1.4402	1