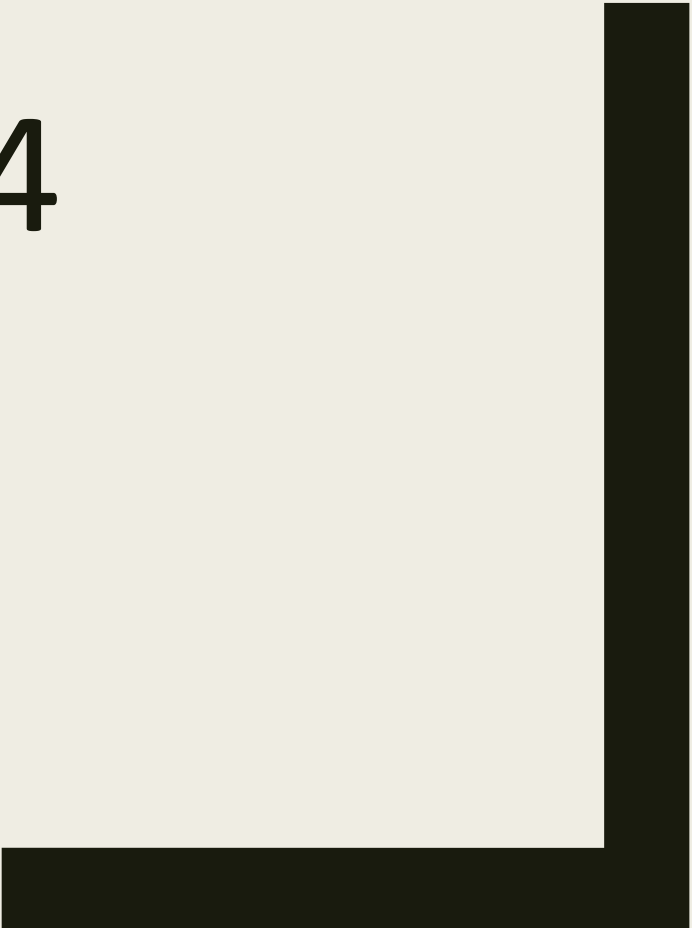




WORKSHOP 3

By Group 5-
Anand Mohan Thakur
Josh Shaji
Poonam Bhaliyan
Poornima Singh
Prateek Ramjanam Singh





PART II: DATA PARTITIONS



Answer 2 and 3, Page 3-46, 3-49

Question:

2- Open the project you worked on in Workshop 1. Open the Retention Diagram. If you don't have access to this project, repeat pages 3-31 to 3-34 of the SAS Advanced Business Analytics course notes (available under BB> Course Information> Resources).

3-*Optional*: Follow instructions on pages 3-46 to 3-49 of the SAS Advanced Business Analytics course notes.

Answer:

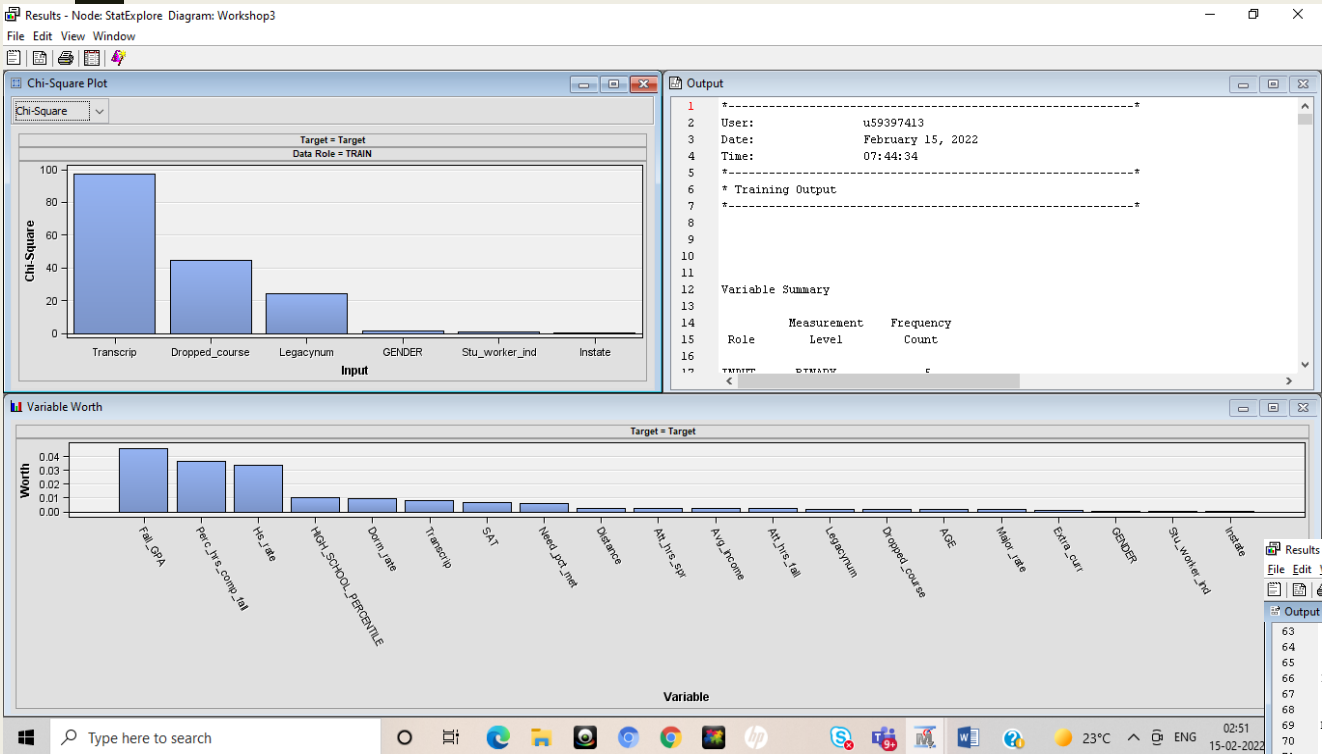
Running the Explore node on Retention data

The screenshot displays the SAS Enterprise Miner - Workshop1 interface. On the left, a tree view shows the project structure: Workshop1, Data Sources, RETENTION, Diagrams, Workshop3, and Model Packages. Below this, a property window for the selected node shows various settings. The main workspace shows a diagram with a 'RETENTION' node connected to a 'StatExplore' node. The bottom status bar indicates 'Diagram Workshop3 opened' and 'Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)'.

Property	Value
General	
Node ID	Stat
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No

Answer 2 and 3, Page 3-46, 3-49

The results and the output:



Results - Node: StatExplore Diagram: Workshop3

File Edit View Window

Output

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
AGE	INPUT	18.55367	0.584676	2626	0	16.85969	18.54346	39.43053	17.50904	619.559
Att_hrs_fall	INPUT	14.61272	1.389914	2626	0	3	15	19	-0.77191	2.502027
Att_hrs_spr	INPUT	14.86596	1.625372	2626	0	4	15	20	-0.36375	0.414561
Avg_income	INPUT	58521.58	23475.95	2469	157	14126	54000	200001	0.853792	0.85756
Distance	INPUT	260.1557	343.9516	2528	98	0.785832	147.8862	3882.192	3.748368	23.05816
Dorm_rate	INPUT	0.836195	0.022936	2626	0	0.6	0.845717	0.878319	-1.85093	9.311383
Extra_curr	INPUT	0.327494	0.609873	2626	0	0	0	4	2.189099	6.096882
Fall_GPA	INPUT	2.946933	0.811767	2626	0	0	3.1	4	-0.99401	0.842477
HIGH_SCHOOL_PERCENTILE	INPUT	79.09107	19.43468	2363	263	0.6	86	100	-1.17481	0.821318
Hs_rate	INPUT	0.836606	0.143239	1821	805	0.5	0.843023	6	25.58073	928.3077
Major_rate	INPUT	0.836452	0.021735	2547	79	0.75	0.831511	0.935484	1.169446	4.451965
Need_pct_met	INPUT	0.917301	0.245779	2626	0	0	1	1	-3.17454	6.722515
Perc_hrs_comp_fall	INPUT	0.935889	0.158619	2626	0	0	1	2.428571	-2.75434	14.58444
SAT	INPUT	1178.313	133.0724	2626	0	750	1180	1600	0.153506	-0.18797

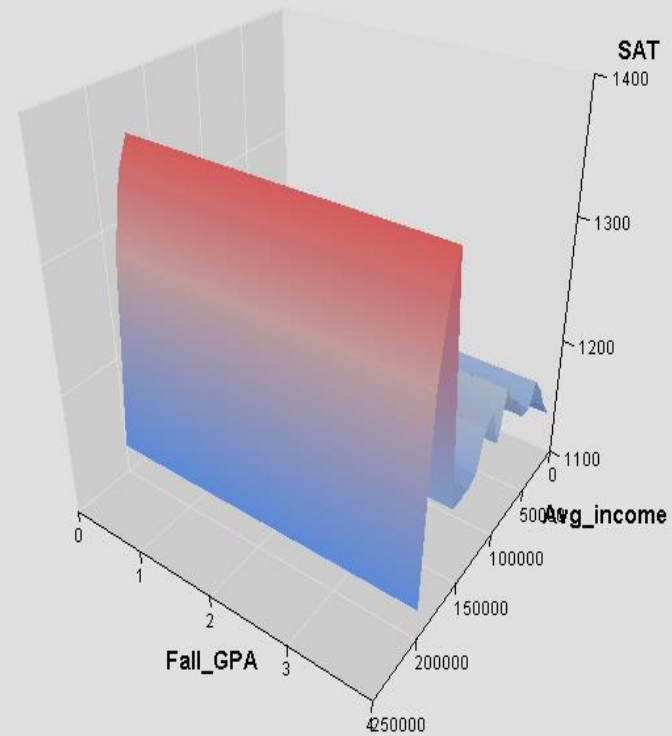
Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Dropped_course

Target	Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Target	Level	Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage

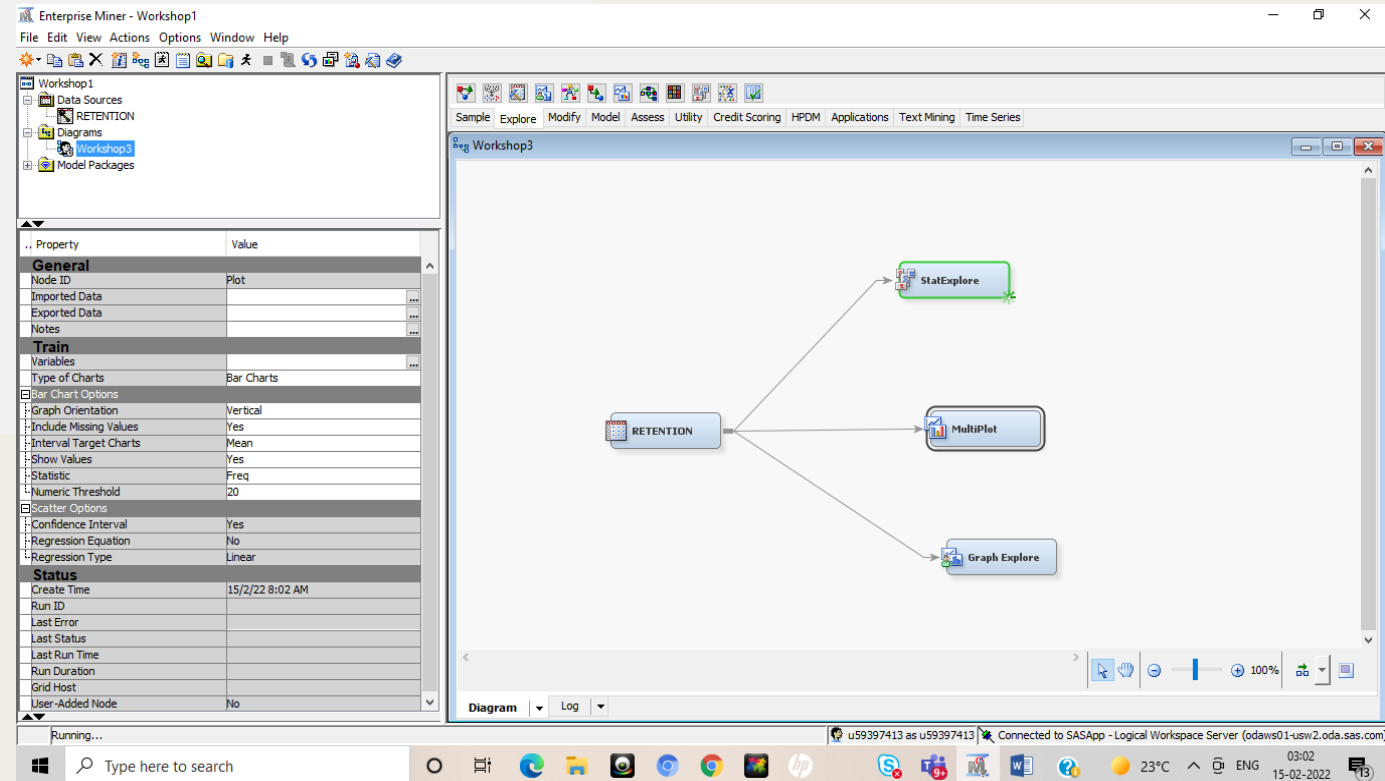
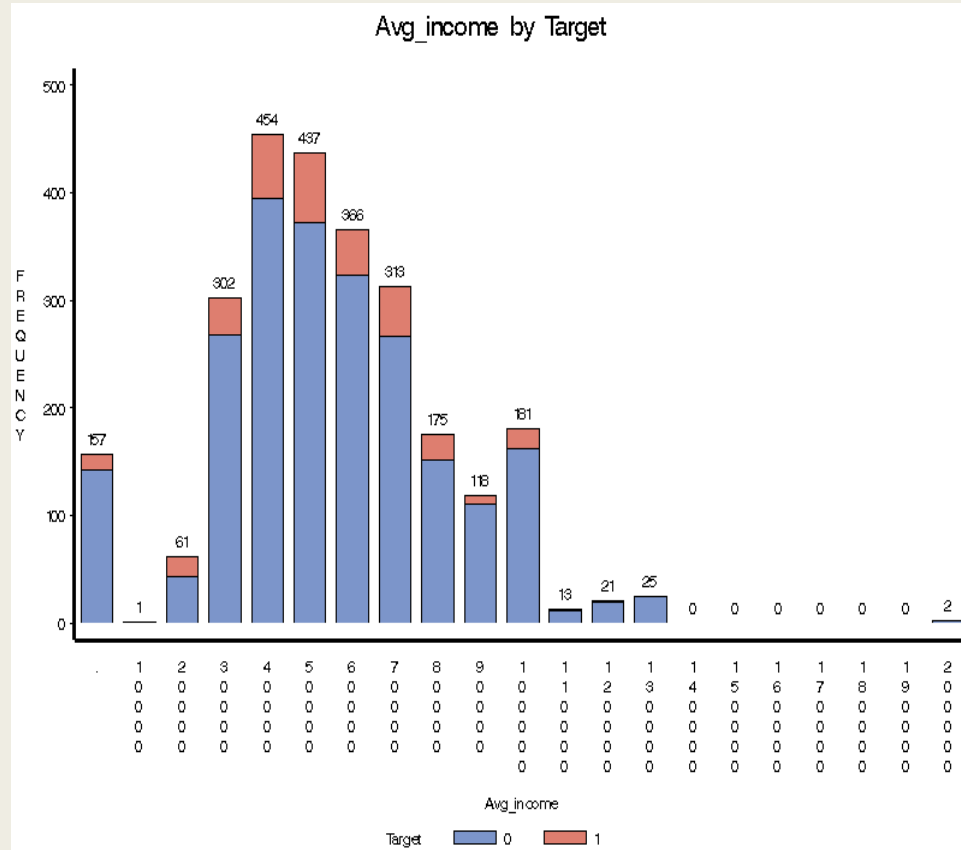
Answer 2 and 3, Page 3-46, 3-49

The Graph Explore:



Answer 2 and 3, Page 3-46, 3-49

The Multi Plot:



Answer 4, Page 3-49

Question:

Follow steps 9 to 10 on page 3-49 of the SAS Advanced Business Analytics course notes to create the data partitions (training and validation sets). Save the results as *WS3_pg3-49.lst* and submit with this workshop.

The screenshot displays the SAS Enterprise Miner - Workshop1 interface. On the left, a tree view shows the project structure: Workshop1, Data Sources, RETENTION, Diagrams, Workshop3, and Model Packages. Below this, a table lists the properties of the selected node (Workshop3).

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	15/2/22 8:06 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a diagram titled 'Workshop3' with a central 'RETENTION' node. Four arrows point from 'RETENTION' to four other nodes: 'StatExplore', 'MultiPlot', 'Data Partition', and 'Graph Explore'. The bottom status bar indicates the user is connected to the SASApp - Logical Workspace Server.

Data Set Allocations

Training	60.0
Validation	40.0
Test	0.0

Answer 4, Page 3-49

Question:

Follow steps 9 to 10 on page 3-49 of the SAS Advanced Business Analytics course notes to create the data partitions (training and validation sets). Save the results as *WS3_pg3-49.lst* and submit with this workshop.

Answer:

Have submitted the *WS3_pg3-49.lst* file on Blackboard

```
Results - Node: Data Partition Diagram: Workshop3
File Edit View Window

Output

1 *-----*
2 User:          u59397413
3 Date:          February 15, 2022
4 Time:          08:07:51
5 *-----*
6 * Training Output
7 *-----*
8
9
10
11
12 Variable Summary
13
14      Measurement   Frequency
15      Role          Level      Count
16
17 INPUT      BINARY      5
18 INPUT      INTERVAL    14
19 INPUT      NOMINAL     1
20 TARGET     BINARY      1
21
22
23
24
25 Partition Summary
26
27      Number of
28      Type      Data Set      Observations
29
30 DATA      EMWS1.Ids_DATA      2626
31 TRAIN      EMWS1.Part_TRAIN     1574
32 VALIDATE   EMWS1.Part_VALIDATE  1052
33
34
35 *-----*
36 * Score Output
37 *-----*
```



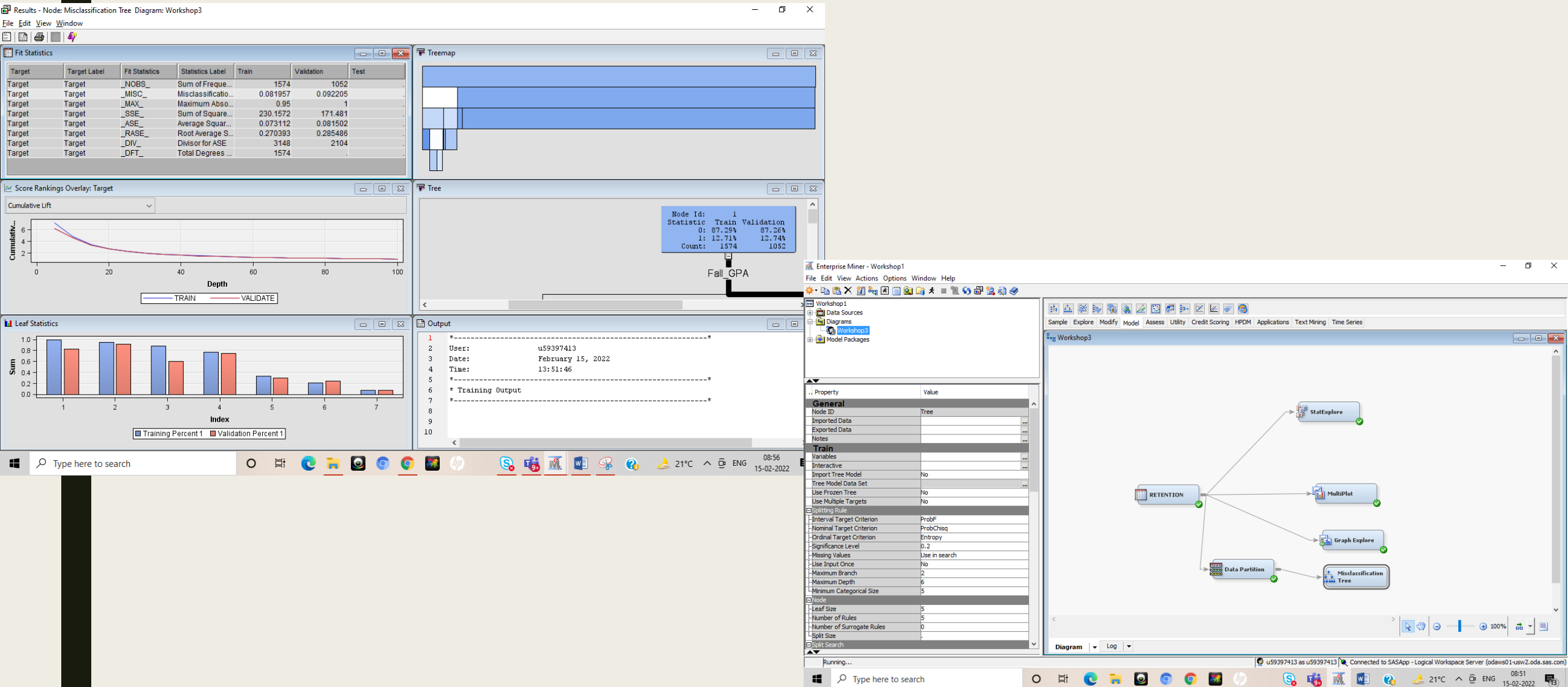

PART III: DECISION TREES



Answer 5, Page 3-59 to 3-61

Question:

Follow the instructions on pages 3-59 to 3-61 of the SAS Advanced Business Analytics course notes. As you go through the steps, answer the following questions.



Answer 5, Page 3-59 to 3-61

Question:

At step 3 on page 3-59, What is the misclassification rate for the validation set?

Answer:

The misclassification rate for the validation set is 0.092205

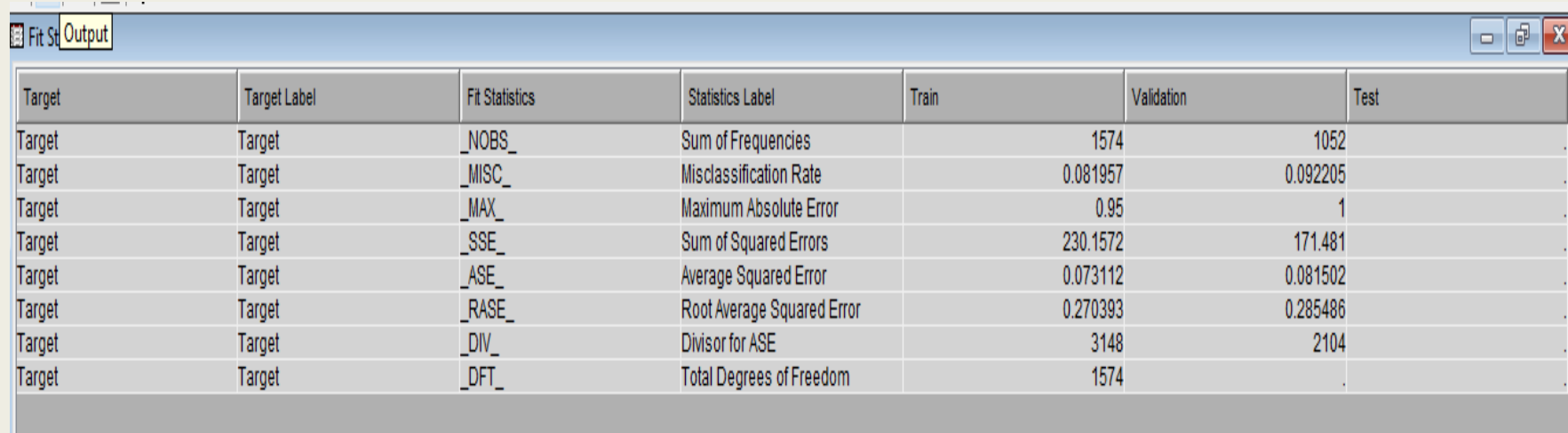
Question:

For the training set, compare sum of squared errors (SSE) with the average squared error (ASE). What is SSE divided by? Is this the same number mentioned as DIV in the Fit Statistics table?

Answer:

The SSE (Sum of Squared Errors) is 171.481 and the ASE (Average Squared Error) is 0.081502.

The SSE is divided by the DFT (total degrees of freedom) which 1574 for training data. Here, the divisor is 3148 for the Train data and 2104 for the validation.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target	Target	_NOBS_	Sum of Frequencies	1574	1052	.
Target	Target	_MISC_	Misclassification Rate	0.081957	0.092205	.
Target	Target	_MAX_	Maximum Absolute Error	0.95	1	.
Target	Target	_SSE_	Sum of Squared Errors	230.1572	171.481	.
Target	Target	_ASE_	Average Squared Error	0.073112	0.081502	.
Target	Target	_RASE_	Root Average Squared Error	0.270393	0.285486	.
Target	Target	_DIV_	Divisor for ASE	3148	2104	.
Target	Target	_DFT_	Total Degrees of Freedom	1574	.	.

Answer 5, Page 3-59 to 3-61

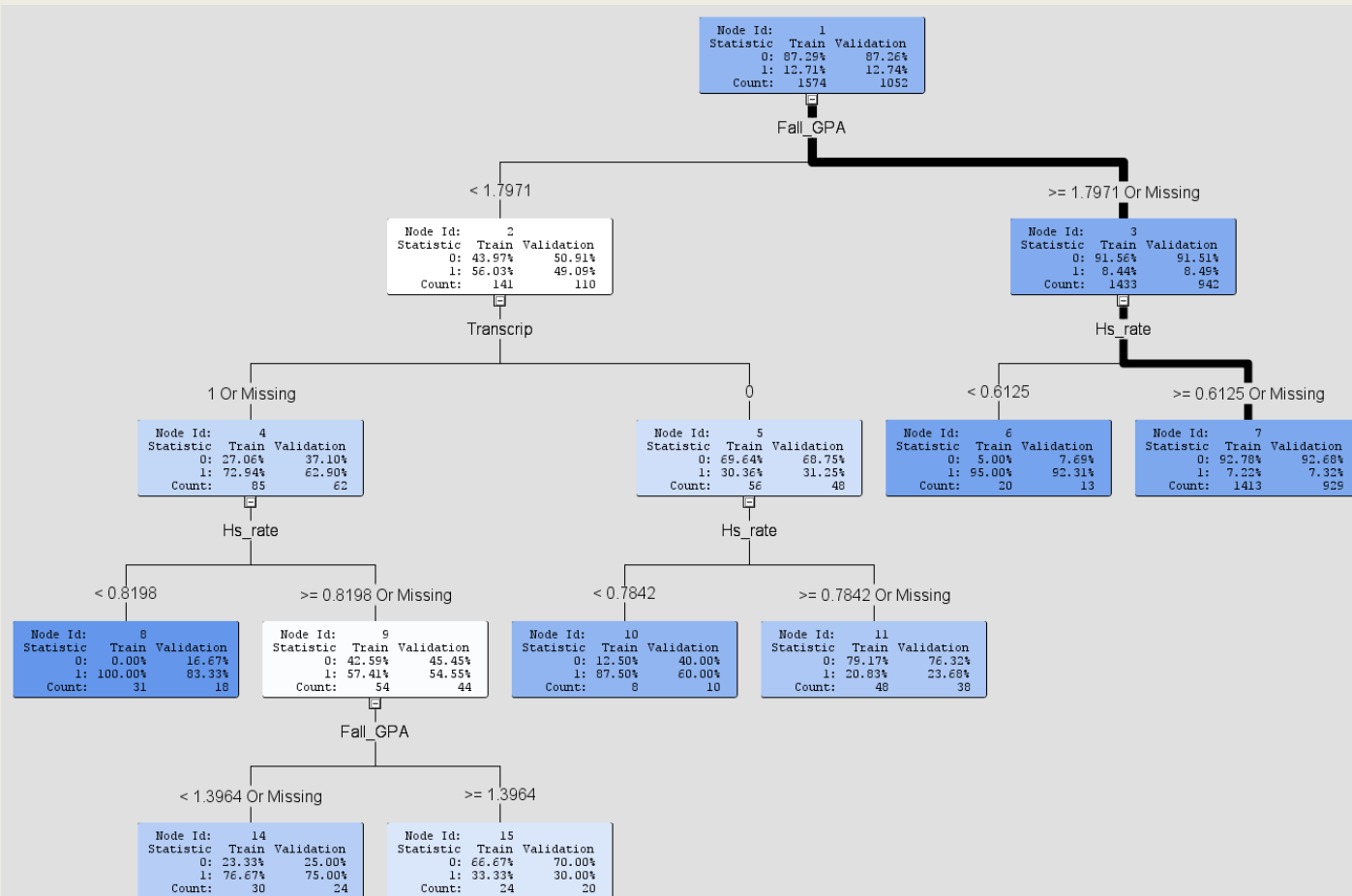
Question:

At step 4 on page 3-59, For the leaves at depth 5, what is the probability of having target = 1? From the menu, choose Edit> View> Fit to page.

Then save the tree as WS3_pg3-59-Tree.bmp and submit with this workshop.

Answer:

In training set for Fall GPA <1.3964 or missing target =1, probability is **76.67%**
Have submitted the file WS3_pg3-59-Tree.bmp on Blackboard



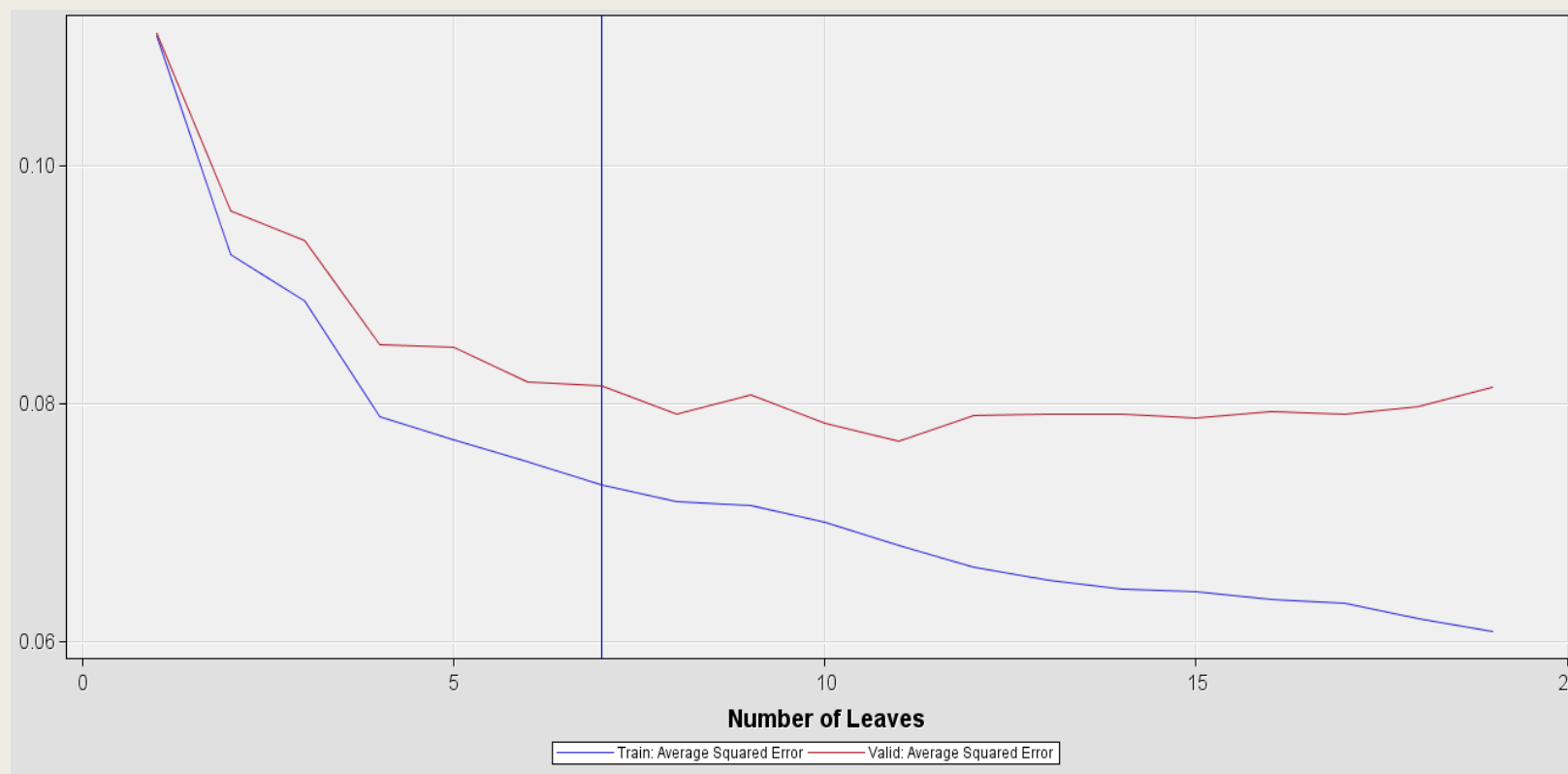
Answer 5, Page 3-59 to 3-61

Question:

At step 5 on page 3-60, look at the average square error plot. What other value for the *number of leaves* looks promising?

Answer:

Decision Trees with 8 to 15 leaves looks promising as the variance between the train and validation model is minimum and this range prevents overfitting and underfitting of the model



Answer 5, Page 3-59 to 3-61

Question:

At step 6 of page 3-60, copy the ‘variable importance’ report and paste here. See page 3-16 and explain which inputs are important for predicting which students will not return to school.

Answer:

Variables which are important for predicting which students will not return to school are as follows:

Fall_GPA: GPA for previous fall semester has a validation importance of 100%

Hs_Rate: Average retention rate for freshmen for the past for their high school, has a validation importance of 77.12%

Transcrip: Binary variable, has 1 if student applied for transcript in fall semester, 0 otherwise, has a validation importance of 38.17%

60						
61	Variable Importance					
62						
63						
64			Number of	Ratio of		
65	Variable		Splitting		Validation	to Training
66	Name	Label	Rules	Importance	Importance	Importance
67						
68	Fall_GPA	Fall_GPA	2	1.0000	1.0000	1.0000
69	Hs_rate	Hs_rate	3	0.8313	0.7712	0.9277
70	Transcrip	Transcrip	1	0.4404	0.3817	0.8669
71						
72						

Answer 5, Page 3-59 to 3-61

Question:

At step 7 of page 3-60, paste the counts of FN, TN, FP, TP for training and validation sets here. You will be using these numbers in Part IV.

Answer:

```
110 Classification Table
111
112 Data Role=TRAIN Target Variable=Target Target Label=Target
113
114           Target      Outcome      Frequency      Total
115 Target      Outcome      Percentage      Percentage      Count      Percentage
116 |
117 0           0           91.9192       99.3450       1365       86.7217
118 1           0           8.0808       60.0000       120        7.6239
119 0           1          10.1124        0.6550         9         0.5718
120 1           1          89.8876       40.0000        80         5.0826
121
122
123 Data Role=VALIDATE Target Variable=Target Target Label=Target
124
125           Target      Outcome      Frequency      Total
126 Target      Outcome      Percentage      Percentage      Count      Percentage
127
128 0           0          91.5907       98.4749       904       85.9316
129 1           0           8.4093       61.9403        83        7.8897
130 0           1          21.5385        1.5251        14         1.3308
131 1           1          78.4615       38.0597        51         4.8479
132
```

```
135
136 Event Classification Table
137
138 Data Role=TRAIN Target=Target Target Label=Target
139
140      False      True      False      True
141 Negative      Negative      Positive      Positive
142
143      120      1365         9         80
144
145
146 Data Role=VALIDATE Target=Target Target Label=Target
147
148      False      True      False      True
149 Negative      Negative      Positive      Positive
150
151      83      904         14         51
152
```

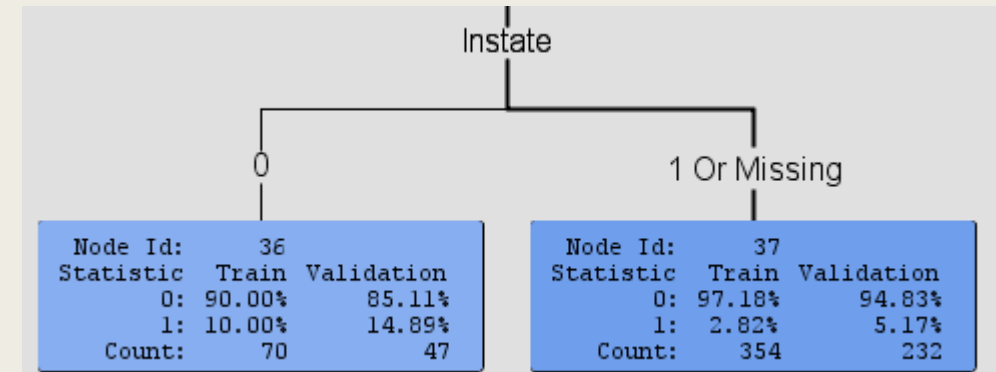
Answer 5, Page 3-59 to 3-61

Question:

At step 11 on page 3-61, copy the 'variable importance' report and paste here. What is the meaning of the rightmost leaf at level 7 (deepest) in plain English?

Answer:

Output						
61	Variable Importance					
62						
63						
64						
65						
66	Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
67						
68	Fall_GPA	Fall_GPA	2	1.0000	1.0000	1.0000
69	Hs_rate	Hs_rate	4	0.8877	0.8262	0.9307
70	Transcrip	Transcrip	2	0.5148	0.5356	1.0403
71	Att_hrs_fall	Att_hrs_fall	1	0.2057	0.0000	0.0000
72	Perc_hrs_comp_fall	Perc_hrs_comp_fall	1	0.2011	0.3440	1.7109
73	Dorm_rate	Dorm_rate	1	0.1938	0.0756	0.3904
74	Instate	Instate	1	0.0976	0.1393	1.4269
75						
76						
77						
78	Tree Leaf Report					
79						
80						
81	Node	Training	Training			
82	Id	Depth	Observations	Percent	Validation	Validation
83				1	Observations	Percent 1
84	13	3	623	0.03	409	0.01
85	37	6	354	0.03	232	0.05
86	18	4	267	0.19	184	0.20
87	28	5	99	0.16	57	0.14
88	36	6	70	0.10	47	0.15
89	25	5	34	0.06	28	0.14
90	8	3	31	1.00	18	0.83
91	14	4	30	0.77	24	0.75
92	15	4	24	0.33	20	0.30
93	6	2	20	0.95	13	0.92
94	16	4	9	0.56	3	0.00
95	10	3	8	0.88	10	0.60
96	24	5	5	0.60	7	0.71



Answer:

The rightmost leaf at level 7 means that if the value of instate is "1 or missing" i.e. if the student lives in the state or the value is unknown, then there is 5.17% probability for the student to return to college and 94.83% probability to not return to college for the validation model.

Answer 5, Page 3-59 to 3-61

Question:

At step 11 of page 3-60, paste the counts of FN, TN, FP, TP for training and validation sets here. You will be using these numbers in Part IV.

Answer:

120	Classification Table					
121						
122	Data Role=TRAIN Target Variable=Target Target Label=Target					
123						
124			Target	Outcome	Frequency	Total
125	Target	Outcome	Percentage	Percentage	Count	Percentage
126						
127	0	0	92.3861	98.9083	1359	86.3405
128	1	0	7.6139	56.0000	112	7.1156
129	0	1	14.5631	1.0917	15	0.9530
130	1	1	85.4369	44.0000	88	5.5909
131						
132						
133	Data Role=VALIDATE Target Variable=Target Target Label=Target					
134						
135			Target	Outcome	Frequency	Total
136	Target	Outcome	Percentage	Percentage	Count	Percentage
137						
138	0	0	92.0164	97.9303	899	85.4563
139	1	0	7.9836	58.2090	78	7.4144
140	0	1	25.3333	2.0697	19	1.8061
141	1	1	74.6667	41.7910	56	5.3232
142						

145				
146	Event Classification Table			
147				
148	Data Role=TRAIN Target=Target Target Label=Target			
149				
150	False	True	False	True
151	Negative	Negative	Positive	Positive
152				
153	112	1359	15	88
154				
155				
156	Data Role=VALIDATE Target=Target Target Label=Target			
157				
158	False	True	False	True
159	Negative	Negative	Positive	Positive
160				
161	78	899	19	56
162				

Answer 5, Page 3-59 to 3-61

Question:

Paste a picture of the retention diagram here.

Answer:

The screenshot displays the SAS Enterprise Miner interface. The main window, titled 'Workshop3', shows a diagram with a central 'RETENTION' node connected to several other nodes: 'StatExplore', 'MultiPlot', 'Graph Explore', 'Data Partition', 'Misclassification Tree', and 'Probability Tree'. The 'Data Partition' node is further connected to 'Misclassification Tree' and 'Probability Tree'. The left pane shows the 'Workshop1' tree structure with 'Workshop3' selected. Below the tree is a table of properties and values.

Property	Value
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5

The bottom status bar indicates the user is connected to the SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com) and shows the system time as 10:17 on 15-02-2022.



PART IV: CLASSIFICATION ASSESSMENT



Answer 6

Question:

Use the numbers you obtained above for the validation set by the “Misclassification Tree” and the “Probability Tree” to fill the following tables. Then calculate Precision, Recall, and F1 for both trees.

Answer:

Misclassification Tree

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 904	FP= 14	FP+TN = 918
Truly 1 (target = 1)	FN= 83	TP= 51	TP+FN = 134
Total	TN+FN= 987	TP+FP= 65	1052

$$\text{Recall (R)} = \text{TP} / (\text{TP} + \text{FN}) = 51/134 = \mathbf{0.3805}$$

$$\text{Precision (P)} = \text{TP} / (\text{TP} + \text{FP}) = 51/65 = \mathbf{0.7846}$$

$$F_1 = 2P.R / (P + R) = (2*0.3805*0.7846)/(0.7846+0.3805) = \mathbf{0.5124}$$

Answer 6

Question:

Use the numbers you obtained above for the validation set by the “Misclassification Tree” and the “Probability Tree” to fill the following tables. Then calculate Precision, Recall, and F1 for both trees.

Answer:

Probability Tree

Probability Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 899	FP=19	FP+TN = 918
Truly 1 (target = 1)	FN=78	TP= 56	TP+FN = 134
Total	TN+FN=977	TP+FP= 75	1052

$$\text{Recall (R)} = TP / (TP + FN) = 56/134 = \mathbf{0.4179}$$

$$\text{Precision (P)} = TP / (TP + FP) = 56/75 = \mathbf{0.7466}$$

$$F_1 = 2P.R / (P + R) = (2*0.7466*0.4179)/(0.7466+0.4179) = \mathbf{0.5358}$$

GROUP WORK DECLARATION

We, **Group 5 (Anand Mohan Thankur, Josh Shaji, Poonam Bhaliyan, Prateek Ramjanam Singh, and Poornima Singh)** declare that the attached assignment is our own work in accordance with the Seneca Academic Policy. We have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. We have not distributed our work to other students.

	Name	Task(s)
1	Anand Mohan Thakur (149200206)	Consolidated the Workshop together on MS Teams
2	Josh Shaji (133557215)	Consolidated the Workshop together on MS Teams
3	Poonam Bhaliyan (121114219)	Consolidated the Workshop together on MS Teams
4	Prateek Ramjanam Singh (124483215)	Consolidated the Workshop together on MS Teams
5	Poornima Singh (125638213)	Consolidated the Workshop together on MS Teams



THANK YOU

