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

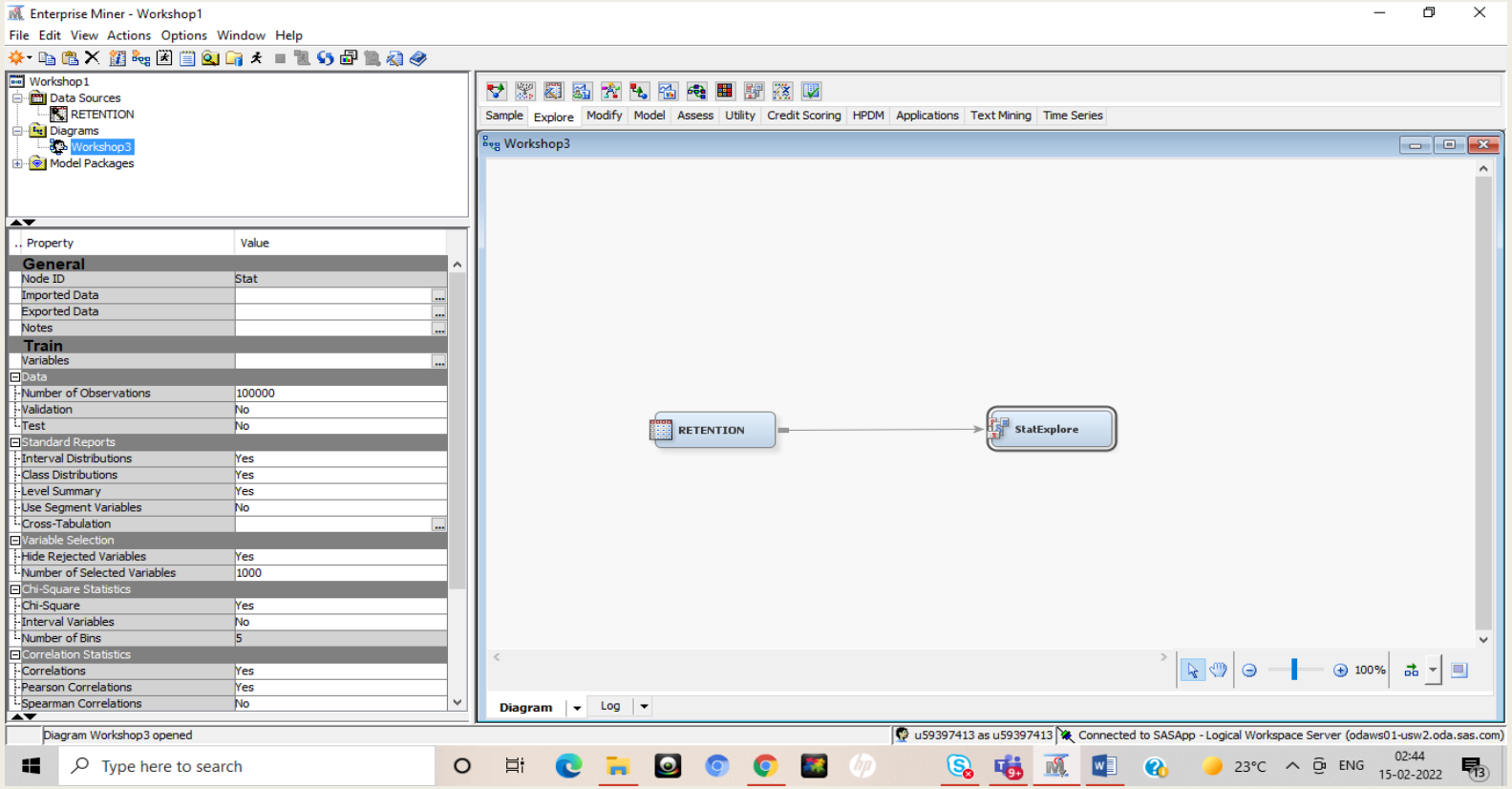
Question:

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.

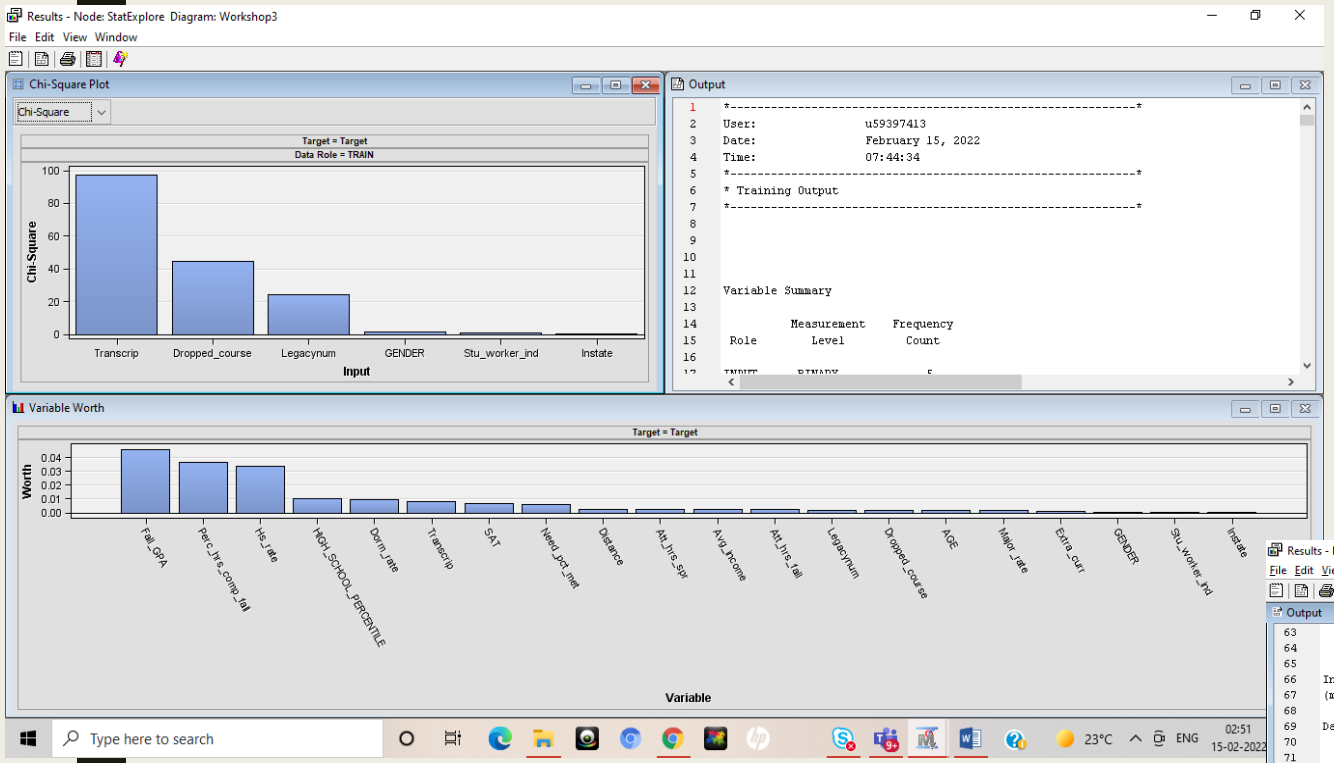
Answer:

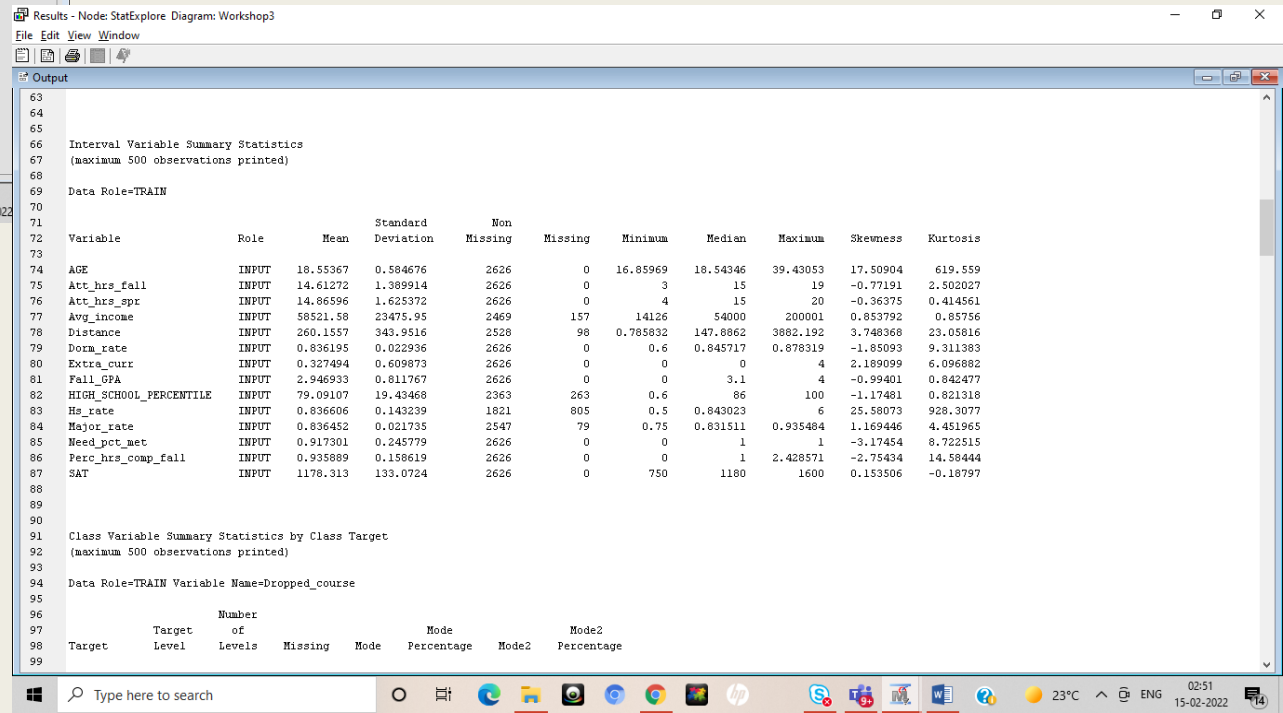
Running the Explore node on Retention data



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

The results and the output:



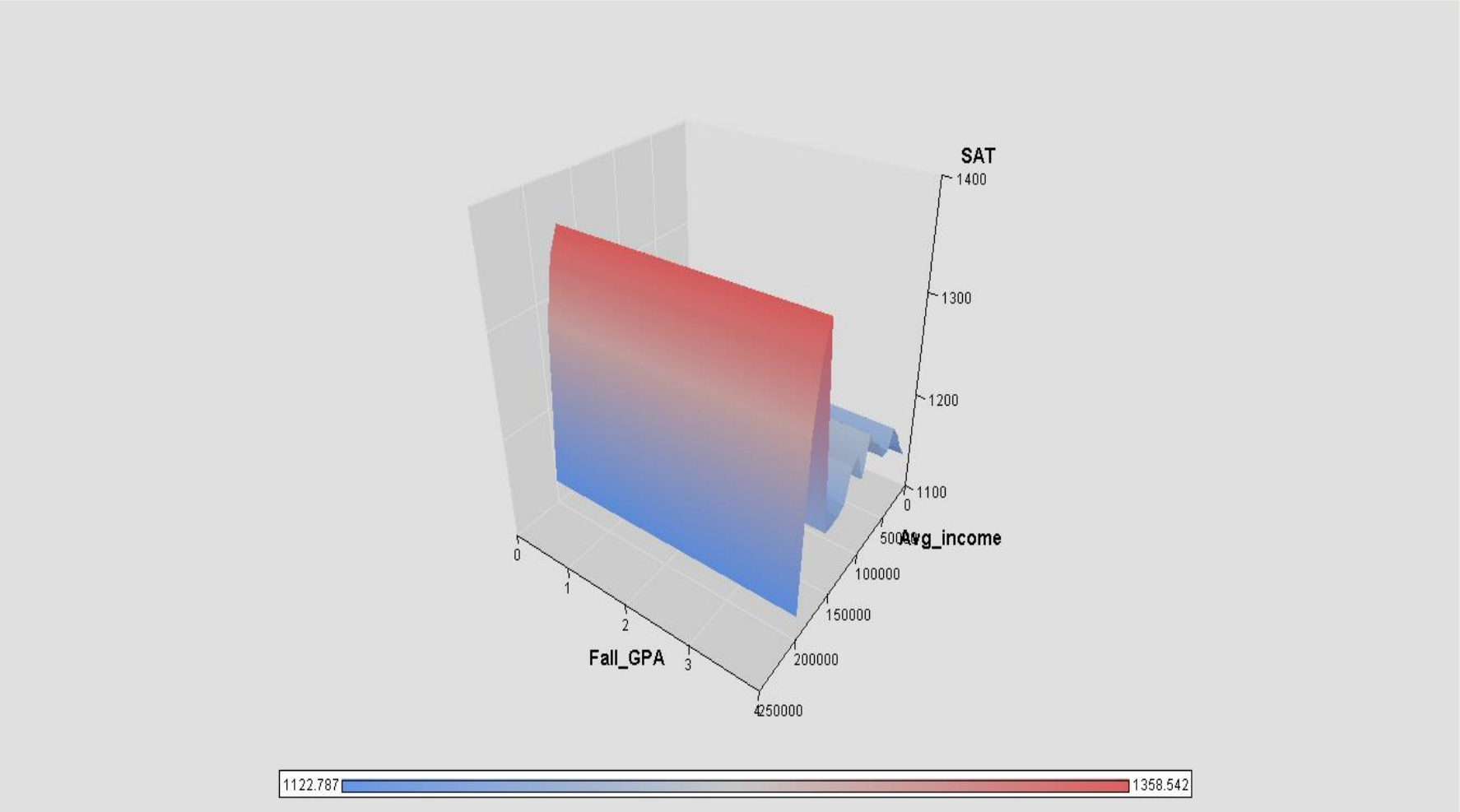


Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
AGE	INPUT	18.55367	0.584676	2626	0	16.85969	18.54346	39.43053	17.50904	619.559
Att_hrs_fall	INPUT	14.61272	1.389914	2626	0	3	15	19	-0.77191	2.502027
Att_hrs_spr	INPUT	14.86596	1.625372	2626	0	4	15	20	-0.36375	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	8.722515
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

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

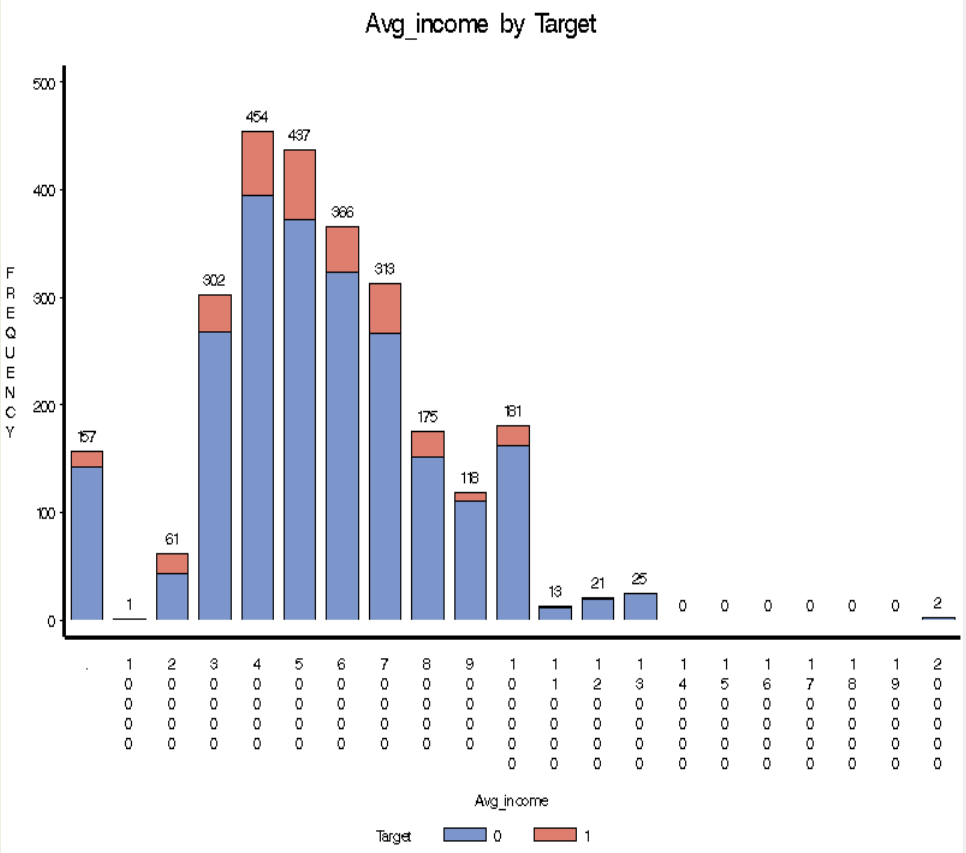
Answer 2 and 3, Page 3-46, 3-49

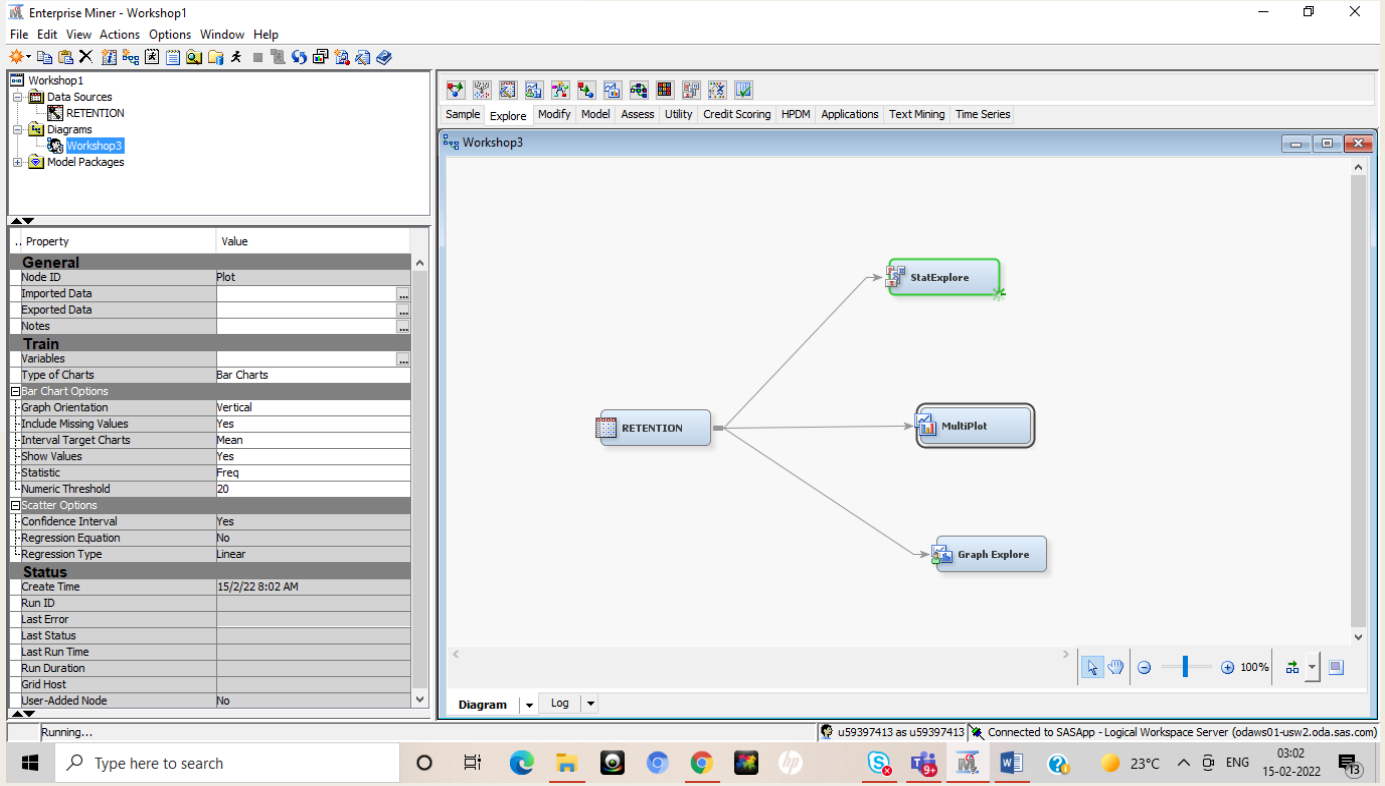
The Graph Explore:



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

The Multi Plot:

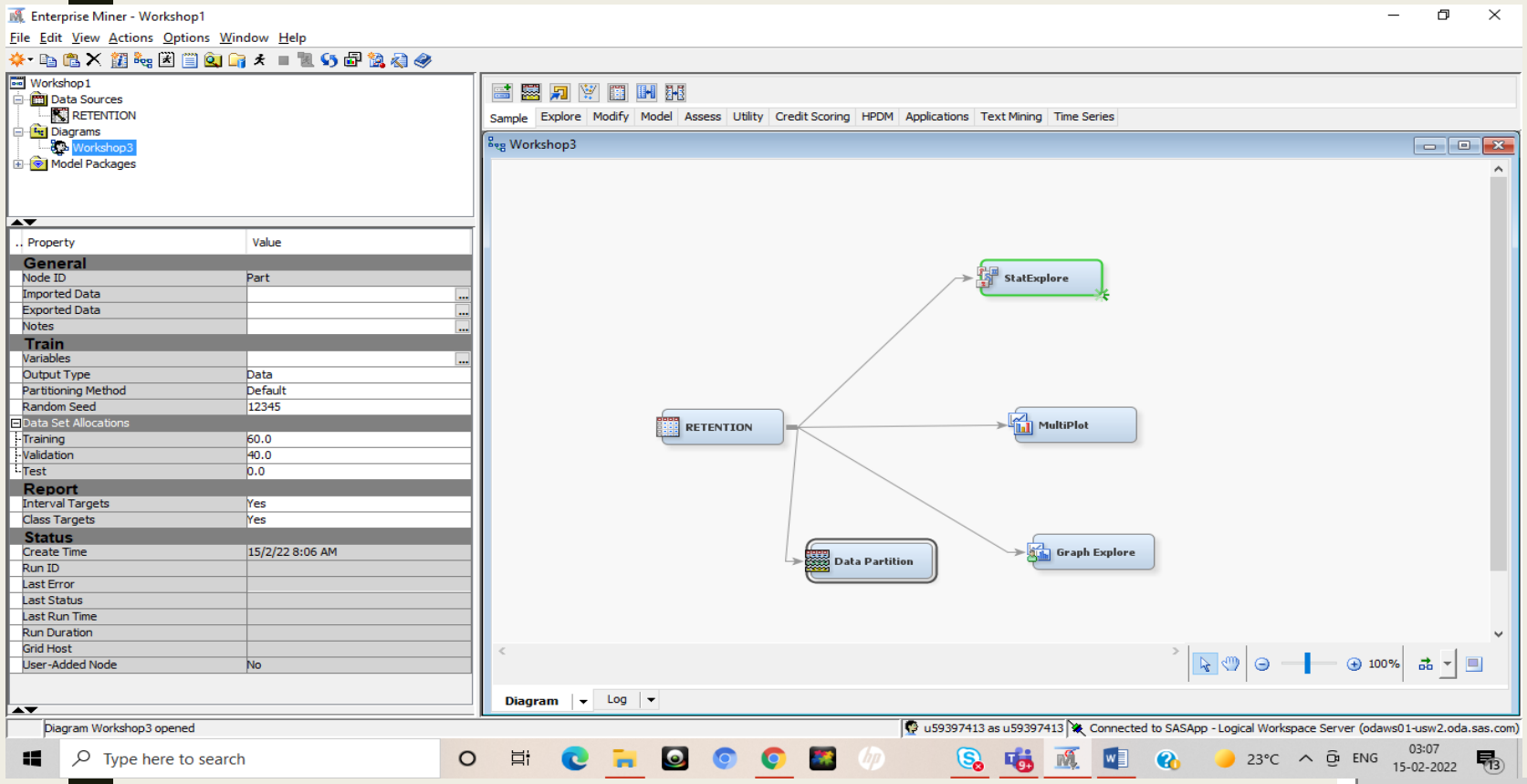


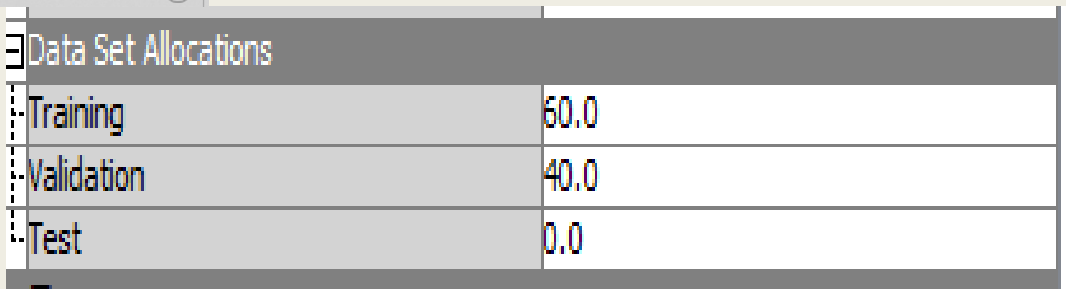


Answer 4, Page 3-49

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.





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

```
*------------------------------------------------------------*
User:                   u59397413
Date:                   February 15, 2022
Time:                   08:07:51
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
 Role          Level         Count

INPUT       BINARY             5
INPUT       INTERVAL          14
INPUT       NOMINAL            1
TARGET      BINARY             1




Partition Summary

                                        Number of
Type              Data Set            Observations

DATA        EMWS1.Ids_DATA              2626
TRAIN       EMWS1.Part_TRAIN            1574
VALIDATE    EMWS1.Part_VALIDATE         1052


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*
```

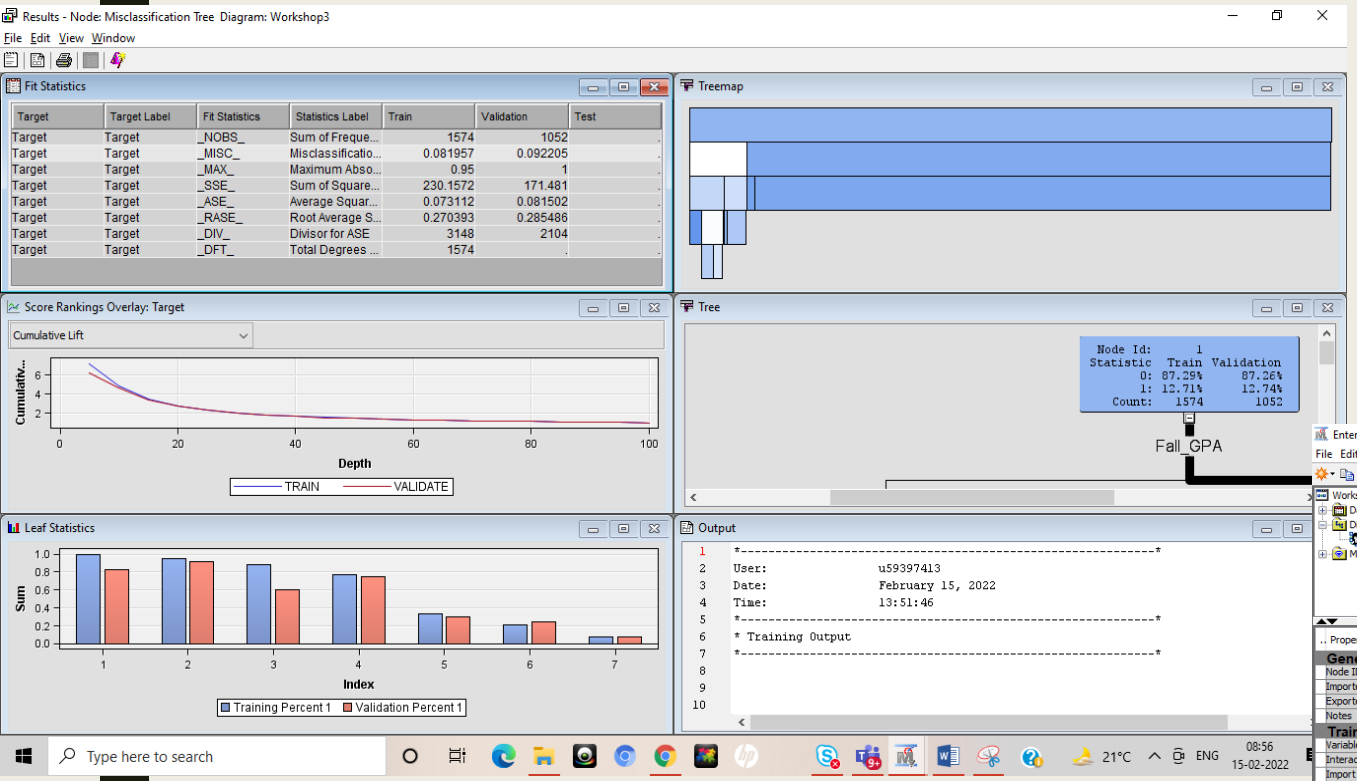
Type here to search 24°C ENG 04:06 15-02-2022

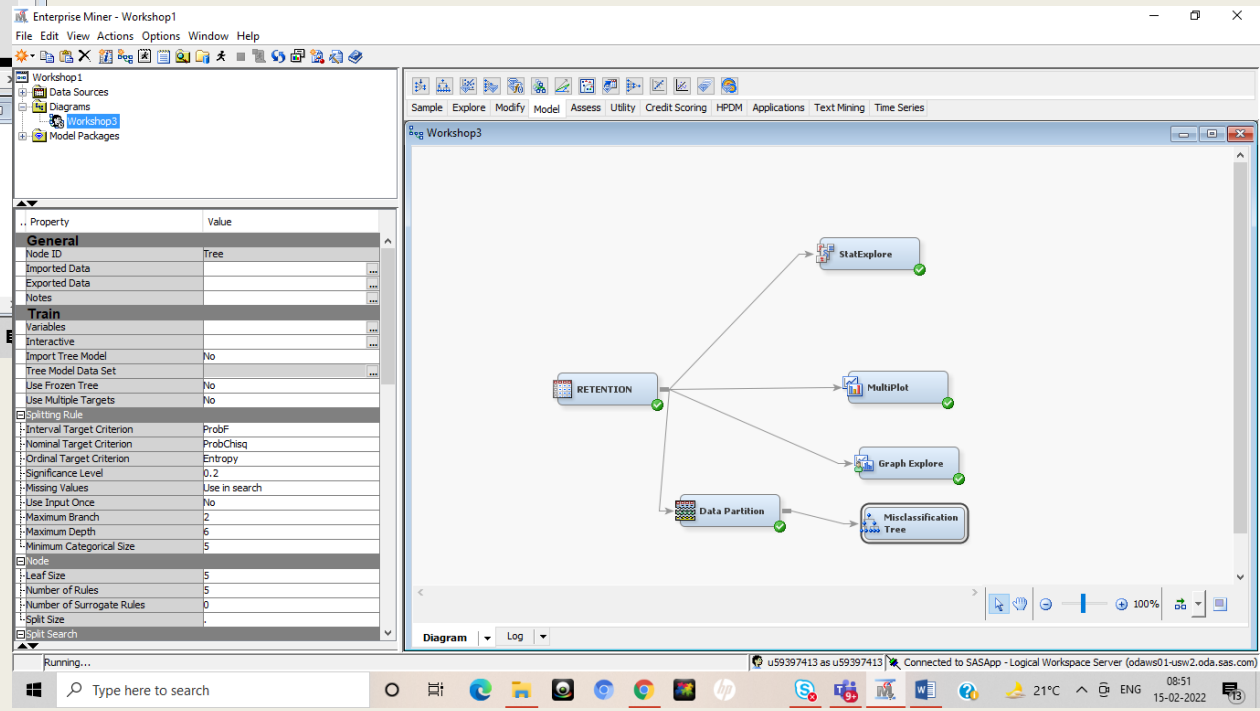
PART III: DECISION TREES

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.

Fit St Output

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:

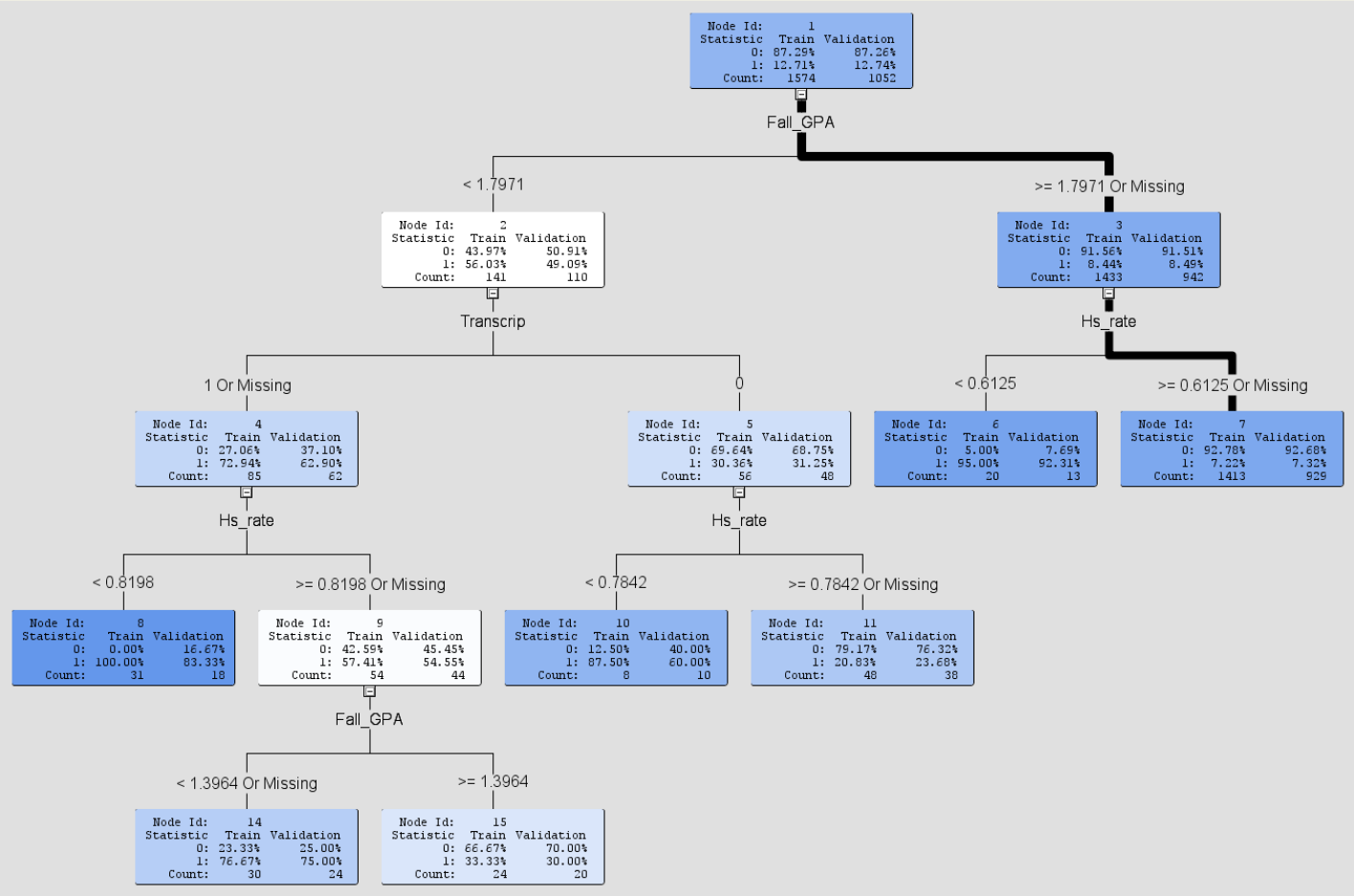
At step 4 on page 3-59,For the leaves at depth 5, what is the probability of having target = 1? From the menu, choose Edit> View> Fit to page.

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

Answer:

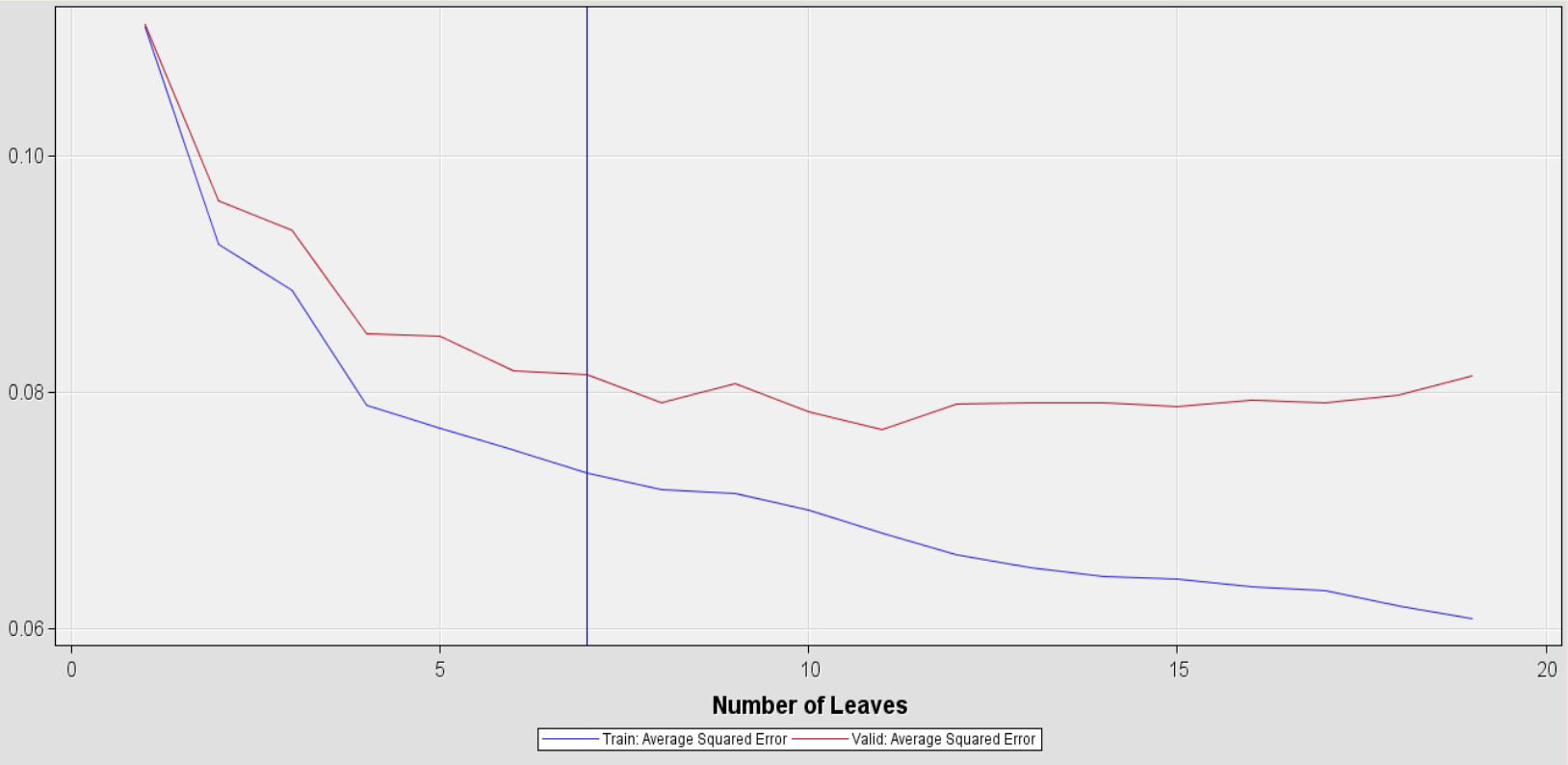
In training set for Fall GPA <1.3964 or missing target =1, probability is **76.67%**

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%

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Fall_GPA	Fall_GPA	2	1.0000	1.0000	1.0000
Hs_rate	Hs_rate	3	0.8313	0.7712	0.9277
Transcrip	Transcrip	1	0.4404	0.3817	0.8669

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:

Classification Table

Data Role=TRAIN Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	91.9192	99.3450	1365	86.7217
1	0	8.0808	60.0000	120	7.6239
0	1	10.1124	0.6550	9	0.5718
1	1	89.8876	40.0000	80	5.0826

Data Role=VALIDATE Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	91.5907	98.4749	904	85.9316
1	0	8.4093	61.9403	83	7.8897
0	1	21.5385	1.5251	14	1.3308
1	1	78.4615	38.0597	51	4.8479

Event Classification Table

Data Role=TRAIN Target=Target Target Label=Target

False Negative	True Negative	False Positive	True Positive
120	1365	9	80

Data Role=VALIDATE Target=Target Target Label=Target

False Negative	True Negative	False Positive	True Positive
83	904	14	51

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:

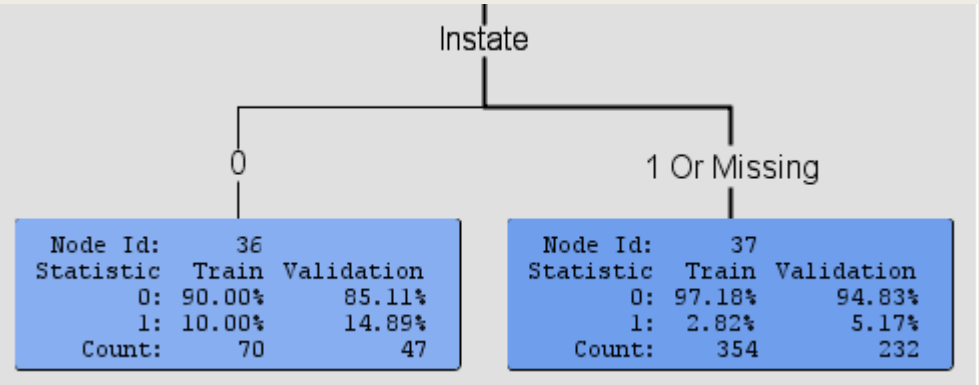
Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Fall_GPA	Fall_GPA	2	1.0000	1.0000	1.0000
Hs_rate	Hs_rate	4	0.8877	0.8262	0.9307
Transcrip	Transcrip	2	0.5148	0.5356	1.0403
Att_hrs_fall	Att_hrs_fall	1	0.2057	0.0000	0.0000
Perc_hrs_comp_fall	Perc_hrs_comp_fall	1	0.2011	0.3440	1.7109
Dorm_rate	Dorm_rate	1	0.1938	0.0756	0.3904
Instate	Instate	1	0.0976	0.1393	1.4269

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1	Validation Observations	Validation Percent 1
13	3	623	0.03	409	0.01
37	6	354	0.03	232	0.05
18	4	267	0.19	184	0.20
28	5	99	0.16	57	0.14
36	6	70	0.10	47	0.15
25	5	34	0.06	28	0.14
8	3	31	1.00	18	0.83
14	4	30	0.77	24	0.75
15	4	24	0.33	20	0.30
6	2	20	0.95	13	0.92
16	4	9	0.56	3	0.00
10	3	8	0.88	10	0.60
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:

Classification Table

Data Role=TRAIN Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	92.3861	98.9083	1359	86.3405
1	0	7.6139	56.0000	112	7.1156
0	1	14.5631	1.0917	15	0.9530
1	1	85.4369	44.0000	88	5.5909

Data Role=VALIDATE Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	92.0164	97.9303	899	85.4563
1	0	7.9836	58.2090	78	7.4144
0	1	25.3333	2.0697	19	1.8061
1	1	74.6667	41.7910	56	5.3232

Event Classification Table

Data Role=TRAIN Target=Target Target Label=Target

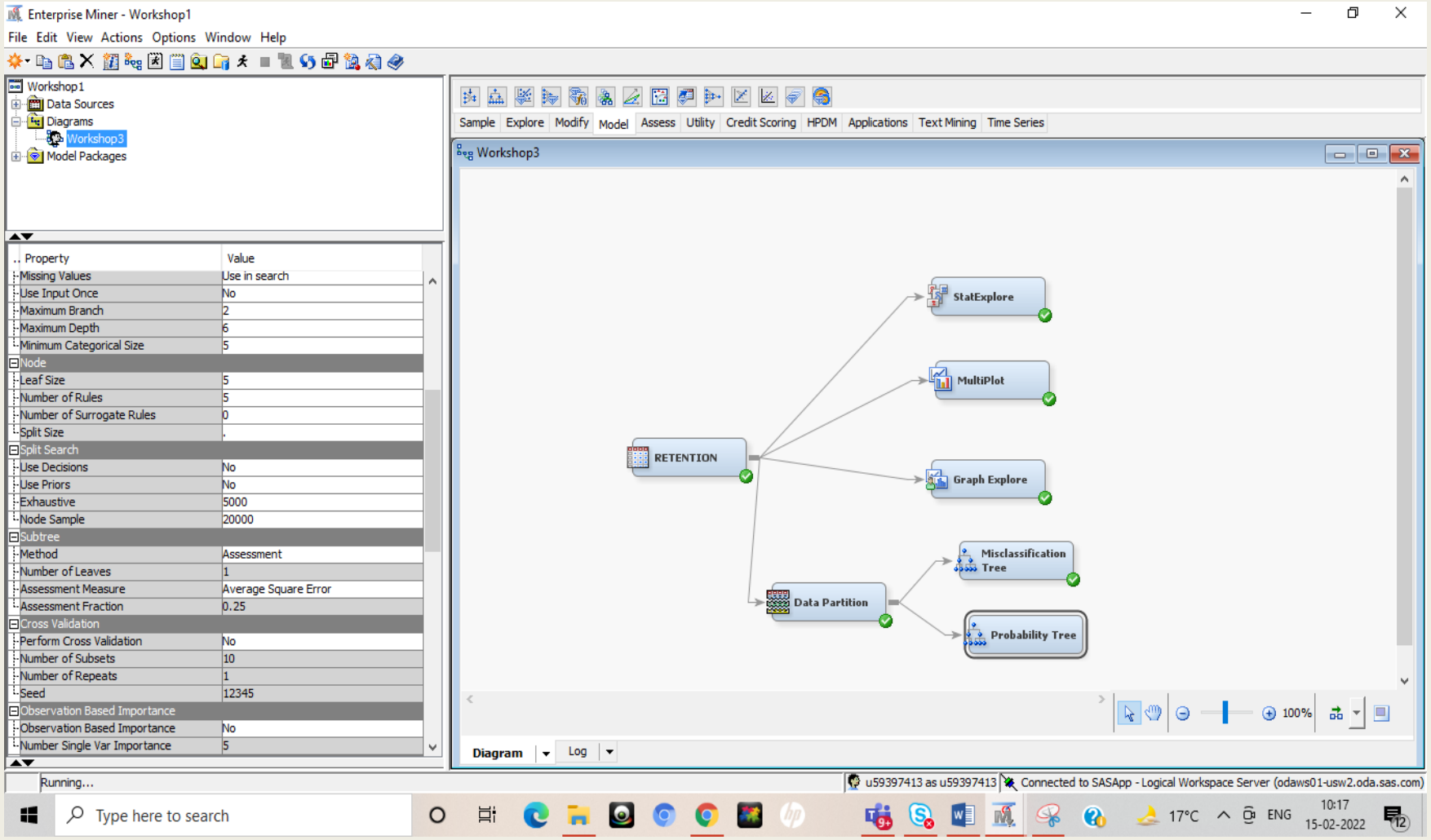
False Negative	True Negative	False Positive	True Positive
112	1359	15	88

Data Role=VALIDATE Target=Target Target Label=Target

False Negative	True Negative	False Positive	True Positive
78	899	19	56

Answer 5, Page 3-59 to 3-61

Question:
Paste a picture of the retention diagram here.
Answer:



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

Recall (R) = TP / (TP + FN) = 51/134 = **0.3805**

Precision (P) = TP / (TP + FP) = 51/65 = **0.7846**

F_1= 2P.R / (P + R) = (2*0.3805*0.7846)/(0.7846+0.3805) = **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

Recall (R) = TP / (TP + FN) = 56/134 = **0.4179**

Precision (P) = TP / (TP + FP) = 56/75 = **0.7466**

F_1= 2P.R / (P + R) = (2*0.7466*0.4179)/(0.7466+0.4179) = **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