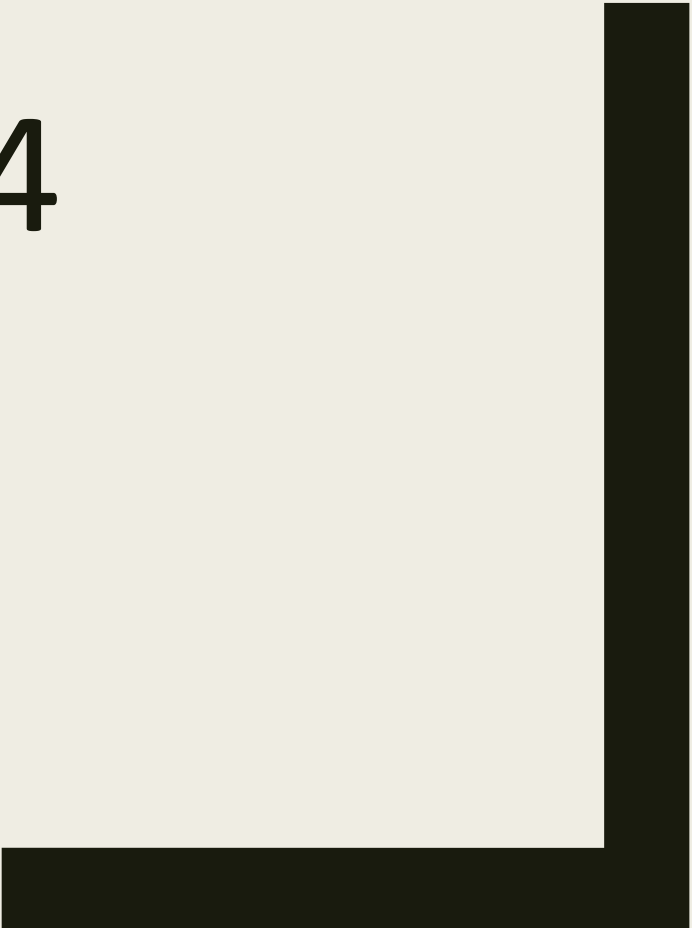




WORKSHOP 5

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

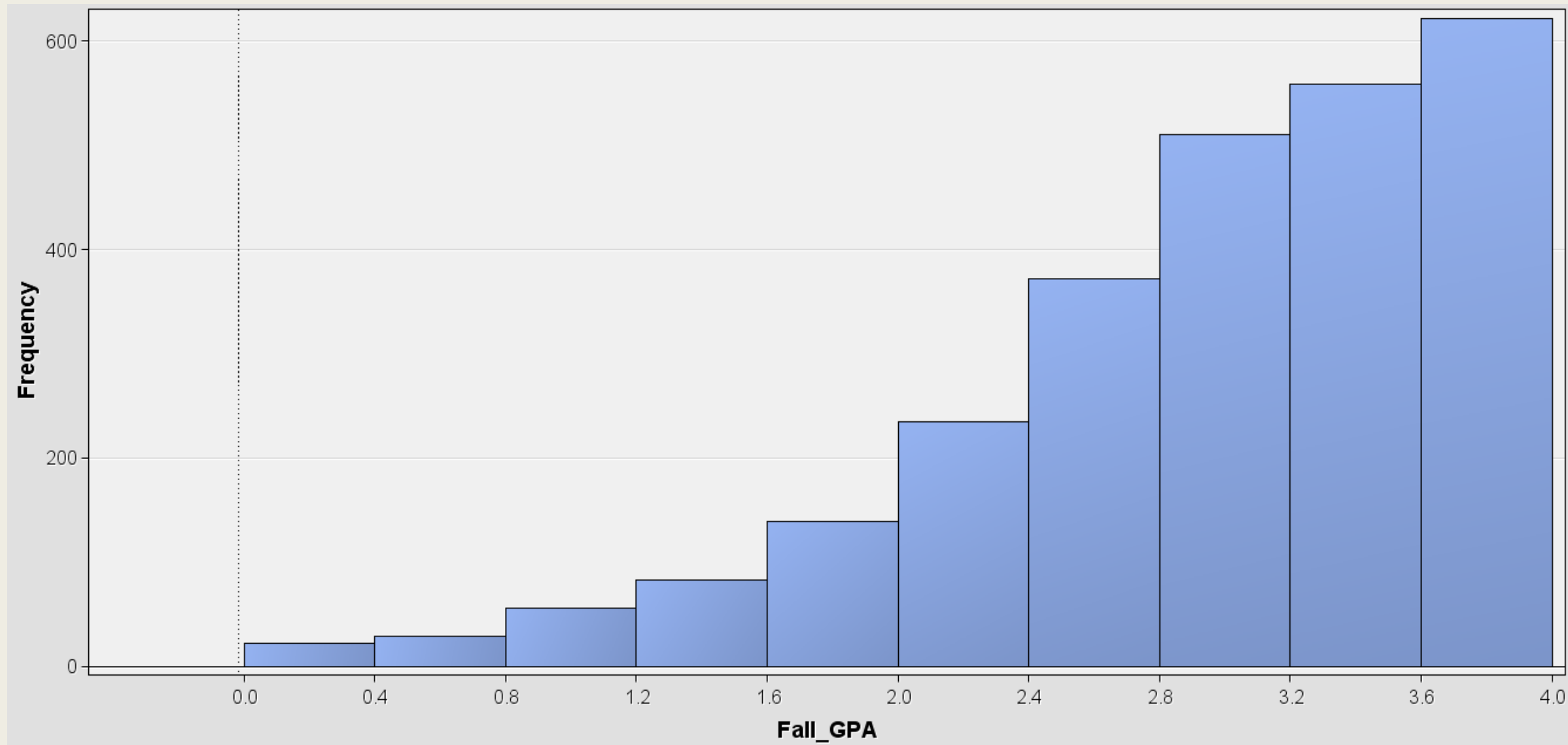




PART II: LOGISTIC REGRESSION



Answer 3: Step 2 - The Fall_GPA Plot



Answer 3: Step 5 – Transformation of Fall_GPA

The formula used for transformation of Fall_GPA is: $\exp(\max(\text{Fall_GPA}-0, 0.0)/4)$

The skewness for Fall_GPA is -0.99401 whereas for the transformed Fall_GPA is -0.49838

Results - Node: Transform Variables Diagram: Workshop5

File Edit View Window

Output

```
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 Computed Transformations
25 (maximum 500 observations printed)
26
27      Input
28      Input Name      Role      Level      Name      Level      Formula
29
30      AGE      INPUT      INTERVAL    SQRT_AGE      INTERVAL    sqrt(max(AGE-16.859685147, 0.0)/22.570841889)
31      Att_hrs_fall      INPUT      INTERVAL    PWR_Att_hrs_fall      INTERVAL    (max(Att_hrs_fall-3, 0.0)/16)**4
32      Avg_income      INPUT      INTERVAL    SQRT_Avg_income      INTERVAL    sqrt(max(Avg_income-14126, 0.0)/185875)
33      Distance      INPUT      INTERVAL    SQRT_Distance      INTERVAL    sqrt(max(Distance-0.7858319691, 0.0)/3881.4065468)
34      Dorm_rate      INPUT      INTERVAL    PWR_Dorm_rate      INTERVAL    (max(Dorm_rate-0.6, 0.0)/0.278319)**4
35      Extra_curr      INPUT      INTERVAL    LOG_Extra_curr      INTERVAL    log(max(Extra_curr-0, 0.0)/4 + 1)
36      Fall_GPA      INPUT      INTERVAL    EXP_Fall_GPA      INTERVAL    exp(max(Fall_GPA-0, 0.0)/4)
37      HIGH_SCHOOL_PERCENTILE      INPUT      INTERVAL    PWR_HIGH_SCHOOL_PERCENTILE      INTERVAL    (max(HIGH_SCHOOL_PERCENTILE-0.6, 0.0)/99.4)**4
38      Hs_rate      INPUT      INTERVAL    LOG_Hs_rate      INTERVAL    log(max(Hs_rate-0.5, 0.0)/5.5 + 1)
39      Need_pct_met      INPUT      INTERVAL    PWR_Need_pct_met      INTERVAL    (max(Need_pct_met-0, 0.0))**4
40
41
42 *-----*
43 * Score Output
44 *-----*
45
46
```

Windows taskbar: Type here to search, 16:56, 04-04-2022

Answer 3: Step 5 – Transformation of Distance

The formula used for transformation of Distance is: $\sqrt{\max(\text{Distance}-0.7858319691, 0.0)/3881.4065468}$

The skewness for Distance is 3.748368 whereas for the transformed Distance is 1.755574. The skewness has reduced

Results - Node: Transform Variables Diagram: Workshop5

File Edit View Window

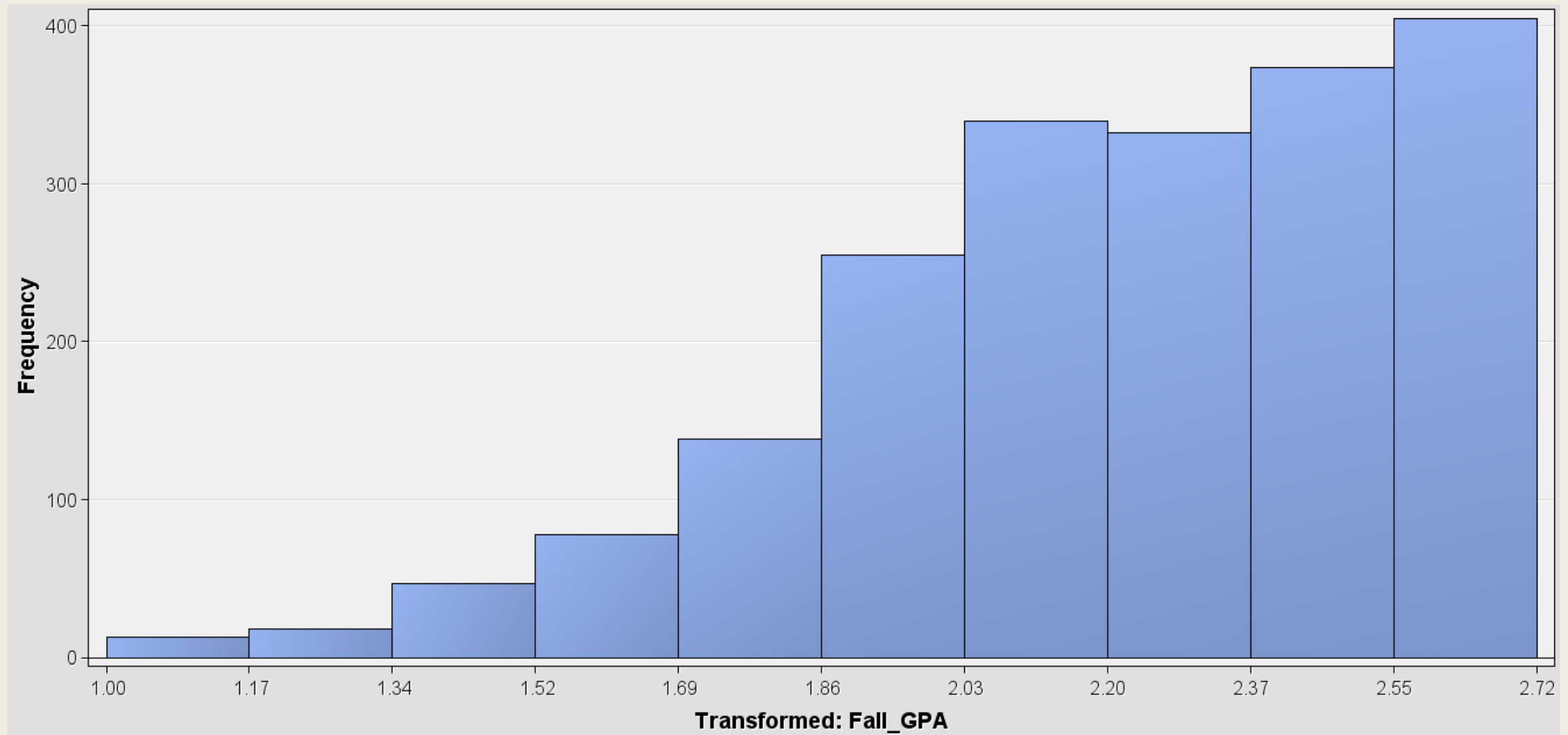
Output

```
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 Computed Transformations
25 (maximum 500 observations printed)
26
27      Input
28      Name      Role      Level      Name      Level      Formula
29
30 AGE            INPUT      INTERVAL    SQRT_AGE      INTERVAL    sqrt(max(AGE-16.859685147, 0.0)/22.570841889)
31 Att_hrs_fall    INPUT      INTERVAL    PWR_Att_hrs_fall    INTERVAL    (max(Att_hrs_fall-3, 0.0)/16)**4
32 Avg_income      INPUT      INTERVAL    SQRT_Avg_income      INTERVAL    sqrt(max(Avg_income-14126, 0.0)/185875)
33 Distance        INPUT      INTERVAL    SQRT_Distance      INTERVAL    sqrt(max(Distance-0.7858319691, 0.0)/3881.4065468)
34 Dorm_rate       INPUT      INTERVAL    PWR_Dorm_rate      INTERVAL    (max(Dorm_rate-0.6, 0.0)/0.278319)**4
35 Extra_curr      INPUT      INTERVAL    LOG_Extra_curr      INTERVAL    log(max(Extra_curr-0, 0.0)/4 + 1)
36 Fall_GPA        INPUT      INTERVAL    EXP_Fall_GPA      INTERVAL    exp(max(Fall_GPA-0, 0.0)/4)
37 HIGH_SCHOOL_PERCENTILE    INPUT      INTERVAL    PWR_HIGH_SCHOOL_PERCENTILE    INTERVAL    (max(HIGH_SCHOOL_PERCENTILE-0.6, 0.0)/99.4)**4
38 Hs_rate         INPUT      INTERVAL    LOG_Hs_rate      INTERVAL    log(max(Hs_rate-0.5, 0.0)/5.5 + 1)
39 Need_pct_met    INPUT      INTERVAL    PWR_Need_pct_met    INTERVAL    (max(Need_pct_met-0, 0.0))**4
40
41
42 *-----*
43 * Score Output
44 *-----*
45
46
```

Windows taskbar: 16:56 04-04-2022

Answer 3: Step 7 – Histogram of EXP Fall_GPA

In the previous graph of Fall_GPA the values range from 0 to 4, whereas in the below graph with exponential values of Fall_GPA, the values range from 1 to 2.72, making the tail less flatter



Answer 3: Step 9 – Tree Surrogate

Tree Surrogate — The Tree Surrogate setting is used to replace missing interval variable values by using the same algorithm as Tree Imputation (in Tree Imputation Use the Tree setting to replace missing interval variable values with replacement values that are estimated by analysing each input as a target. The remaining input and rejected variables are used as predictors), except with the addition of surrogate splitting rules. A surrogate rule is a backup to the main splitting rule. When the main splitting rule relies on an input whose value is missing, the next surrogate is invoked. If missing values prevent the main rule and all the surrogates from applying to an observation, the main rule assigns the observation to the branch that is assigned to receive missing values.

The screenshot displays the SAS Enterprise Miner - Workshop1 interface. On the left, the 'Workshop1' tree shows 'Data Sources' (RETENTION), 'Diagrams' (Workshop3, Workshop5), and 'Model Packages'. The 'Properties' pane for 'Workshop5' is open, showing the 'Default Input Method' set to 'Tree Surrogate'. Below this, the 'Score' section shows 'Hide Original Variables' set to 'Yes'. The 'Report' section shows 'Validation and Test Data' and 'Distribution of Missing' both set to 'No'. The main workspace shows a workflow diagram with three steps: 'RETENTION' (Data Source), 'Transform Variables' (Process), and 'Impute' (Process). A 'Data Partition' step is also visible, connected to the 'Transform Variables' step. The bottom status bar indicates the user is 'u59397413 as u59397413' and is 'Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)'.

Property	Value
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input
Report	
Validation and Test Data	No
Distribution of Missing	No

Default Input Method
Specifies the imputation method for class input variables.

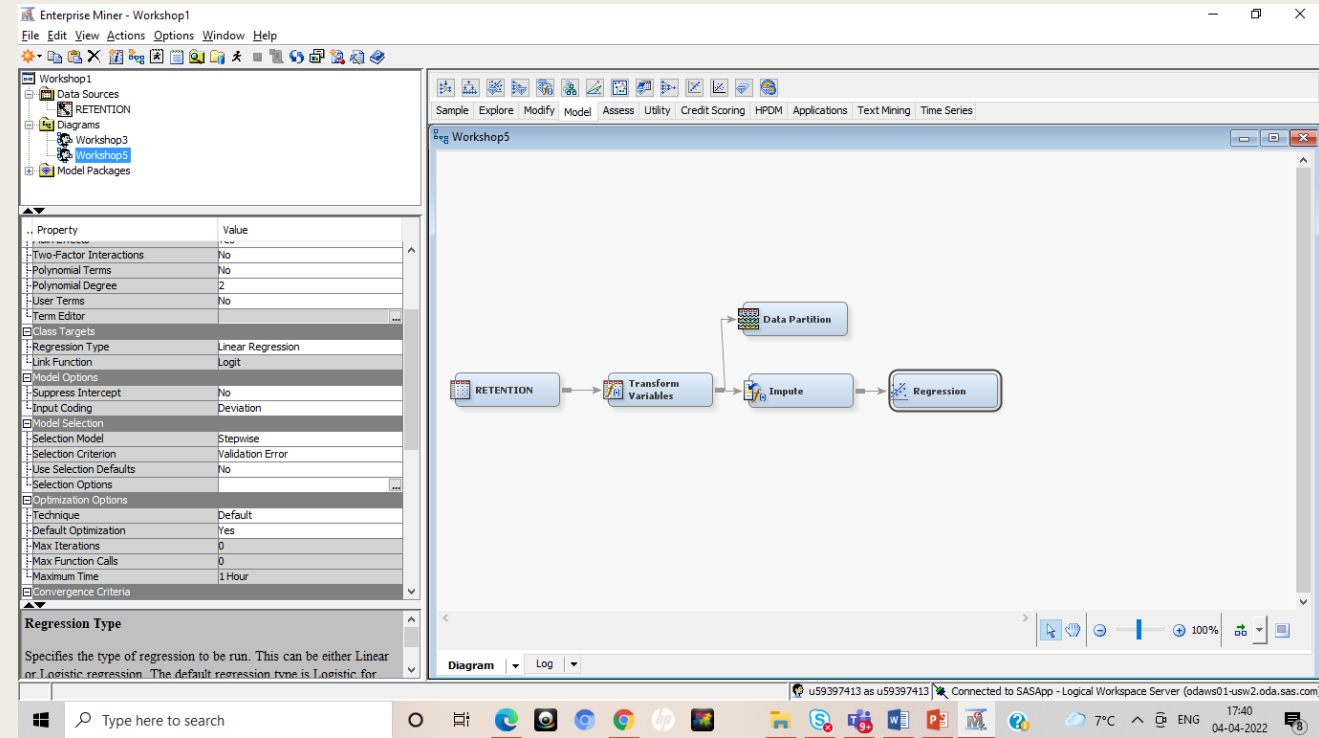
```
graph LR; RETENTION[RETENTION] --> Transform[Transform Variables]; Transform --> Impute[Impute]; Transform --> DataPartition[Data Partition];
```

Answer 3: Step 11 – Stepwise Regression

Stepwise Regression is a variable selection method.

Stepwise: In Stepwise Regression selection begins, by default, with no candidate effects in the model and then systematically adds effects that are significantly associated with the target. However, after an effect is added to the model, Stepwise can remove any effect already in the model that is not significantly associated with the target. This stepwise process continues until one of the following occurs:

- No other effect in the model meets the Stay Significance Level.
- The Max Steps criterion is met. If you choose the Stepwise selection method, then you can specify a Max Steps to put a limit on the number of steps before the effect selection process stops. The default value is set to the number of effects in the model. If you add interactions via the Interaction Builder, the Max Steps is automatically updated to include these terms.
- An effect added in one step is the only effect deleted in the next step



Answer 4: Step 19 – Classification Table for Validation Set

Please find the Classification Table for the Validation Set as follows:

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 903	FP= 87	FP+TN = 990
Truly 1 (target = 1)	FN= 15	TP= 47	TP+FN = 62
Total	TN+FN= 918	TP+FP= 134	

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 47/62 = 0.758$$

$$\text{Precision (P)} = TP/(TP+FP) = 47/134 = 0.351$$

$$F1 = 2P.R/(P+R) = (2*(0.758*0.351))/(0.758+0.351) = 0.4978$$

The Precision value is 0.351, which signifies that only 35.1% of the validation dataset is correctly identified as True Positives which is a really low number. The F1 value of 49.78% is low too

Answer 5: To include Quadratic Polynomial Term in the model

The screenshot displays the SAS Enterprise Miner - Workshop1 interface. The left pane shows the project tree with 'Workshop5' selected. The middle pane shows the 'Train' tab with the following configuration:

Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error

The right pane shows the 'Workshop5' diagram with the following workflow:

```
graph LR; RETENTION[RETENTION] --> TV[Transform Variables]; TV --> DP[Data Partition]; DP --> IMP[Impute]; IMP --> REG[Regression];
```

The bottom status bar indicates 'Diagram Workshop5 opened' and 'Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)'.

Answer 6: F1 value after including Quadratic Polynomial Term

Please find the Classification Table for the Validation Set as follows:

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 892	FP= 26	FP+TN = 918
Truly 1 (target = 1)	FN= 75	TP= 59	TP+FN = 134
Total	TN+FN= 967	TP+FP= 85	

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 59/134 = 0.440$$

$$\text{Precision (P)} = TP/(TP+FP) = 59/85 = 0.694$$

$$F1 = 2P.R/(P+R) = (2*(0.694*0.440))/(0.694+0.440) = 0.538$$

The Precision value is 0.694, which signifies that only 69.4% of the validation dataset is correctly identified as True Positives which is a better model than the previous model. The F1 score is 53.8% which is higher than 49.78% in the previous model



PART III: NEURAL NETWORKS



Answer 7: The number of variables selected in Variable Selection node

Variables with low R square values are rejected. 12 variables are selected

Results - Node: Variable Selection Diagram: Workshop5Part2

File Edit View Window

Variable Selection

Variable Name	Role ▲	Measurement Level	Type	Label	Reasons for Rejection
G_Dropped_course	Input	Nominal	Numeric	Grouped Levels for Dropped_course	
LG10_Dorm_rate	Input	Interval	Numeric	Transformed: Dorm_rate	
LG10_Fall_GPA	Input	Interval	Numeric	Transformed: Fall_GPA	
LG10_IMP_HIGH_SCHOOL_PERCEN...	Input	Interval	Numeric	Transformed: Imputed: HIGH_SCHOO...	
LG10_IMP_Hs_rate	Input	Interval	Numeric	Transformed: Imputed: Hs_rate	
LG10_Need_pct_met	Input	Interval	Numeric	Transformed: Need_pct_met	
Legacynum	Input	Binary	Character	Legacynum	
OPT_Att_hrs_spr	Input	Nominal	Character	Transformed: Att_hrs_spr	
OPT_IMP_Avg_income	Input	Nominal	Character	Transformed: Imputed: Avg_income	
OPT_Perc_hrs_comp_fall	Input	Nominal	Character	Transformed: Perc_hrs_comp_fall	
SAT	Input	Interval	Numeric	SAT	
Transcrip	Input	Binary	Character	Transcrip	
Dropped_course	Rejected	Nominal	Character	Dropped_course	Varsel:Small R-square value, Group va...
GENDER	Rejected	Binary	Character	GENDER	Varsel:Small R-square value
Instate	Rejected	Binary	Character	Instate	Varsel:Small R-square value
LG10_AGE	Rejected	Interval	Numeric	Transformed: AGE	Varsel:Small R-square value
LG10_Att_hrs_fall	Rejected	Interval	Numeric	Transformed: Att_hrs_fall	Varsel:Small R-square value
LG10_Extra_curr	Rejected	Interval	Numeric	Transformed: Extra_curr	Varsel:Small R-square value
LG10_IMP_Distance	Rejected	Interval	Numeric	Transformed: Imputed: Distance	Varsel:Small R-square value
OPT_IMP_Major_rate	Rejected	Nominal	Character	Transformed: Imputed: Major_rate	Varsel:Small R-square value
Stu_worker_ind	Rejected	Binary	Character	Stu_worker_ind	Varsel:Small R-square value

Type here to search

19:03 04-04-2022

Answer 8: Neural Network Configuration

The screenshot displays the Enterprise Miner interface with a workflow diagram and a configuration dialog for a Neural Network.

Workflow Diagram:

```
graph LR; A[Variable Selection] --> B[Data Partition]; B --> C[Neural Network];
```

Enterprise Miner - Workshop1

File Edit View Actions Options Window Help

Workshop1

- Data Sources
 - RETENTION
- Diagrams
 - Workshop3
 - Workshop5
 - Workshop5Part2
- Model Packages

Property Value

Property	Value
General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	4/4/22 11:14 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Workshop5Part2

Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series

Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	Yes
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

Number of Hidden Units

Specifies the number of n hidden units you want in the hidden layer. Permissible values are integers between 1 and 64. The default value is 3.

OK Cancel

Diagram Log

19:15 04-04-2022

Answer 9: F1 Measure for the Neural Network model

477 Event Classification Table

478

479 Data Role=TRAIN Target=Target Target Label=Target

480

481	False	True	False	True
482	Negative	Negative	Positive	Positive

483

484	137	1806	27	130
-----	-----	------	----	-----

485

486

487 Data Role=VALIDATE Target=Target Target Label=Target

488

489	False	True	False	True
490	Negative	Negative	Positive	Positive

491

492	32	456	3	35
-----	----	-----	---	----

Answer 9: F1 Measure for the Neural Network model

Please find the Classification Table for the Validation Set as follows:

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 456	FP= 3	FP+TN = 459
Truly 1 (target = 1)	FN= 32	TP= 35	TP+FN = 67
Total	TN+FN= 488	TP+FP= 38	

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 35/67 = 0.522$$

$$\text{Precision (P)} = TP/(TP+FP) = 35/38 = 0.921$$

$$F1 = 2P.R/(P+R) = (2*(0.522*0.921))/(0.522+0.921) = 0.667$$

The Precision value is 0.921, which signifies that 92.1% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 66.7% which is higher and better than 53.8% obtained in the regression model

Answer 10: F1 Measure for the Neural Network model

Enterprise Miner - Workshop1

File Edit View Actions Options Window Help

Workshop1

- Data Sources
 - RETENTION
- Diagrams
 - Workshop3
 - Workshop5
 - Workshop5Part2
- Model Packages

Property	Value
ID	EMWS3
Name	Workshop5Part2
Status	Open
Notes	
History	
Create Date	4/4/22 10:24 PM
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA
Native OS	Yes

Workshop5Part2

Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series

RETENTION → Impute → Transform Variables → Variable Selection → Data Partition → Neural Network

Diagram Log

Running...

u59397413 as u59397413 Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

Type here to search

19:19 04-04-2022



PART IV: KNN



Answer 11, 12 and 13: New Project, KNN Diagram and File Import

The screenshot displays the SAS Enterprise Miner - Workshop5KNN interface. The main workspace shows a KNN diagram with a 'File Import' button. The left pane contains a tree view with 'Workshop5KNN', 'Data Sources', 'Diagrams', 'KNN', and 'Model Packages'. The bottom pane shows the properties of the KNN diagram, categorized into General, Train, Score, Report, and Status.

General

Property	Value
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	

Train

Variables	
Import File	C:\Users\Poornima Singh\Desktop\Se...
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advisor	No
Rerun	No

Score

Role	Train
------	-------

Report

Summarize	No
-----------	----

Status

Create Time	5/4/22 3:49 PM
Run ID	
Last Error	
Last Status	

Diagram KNN opened

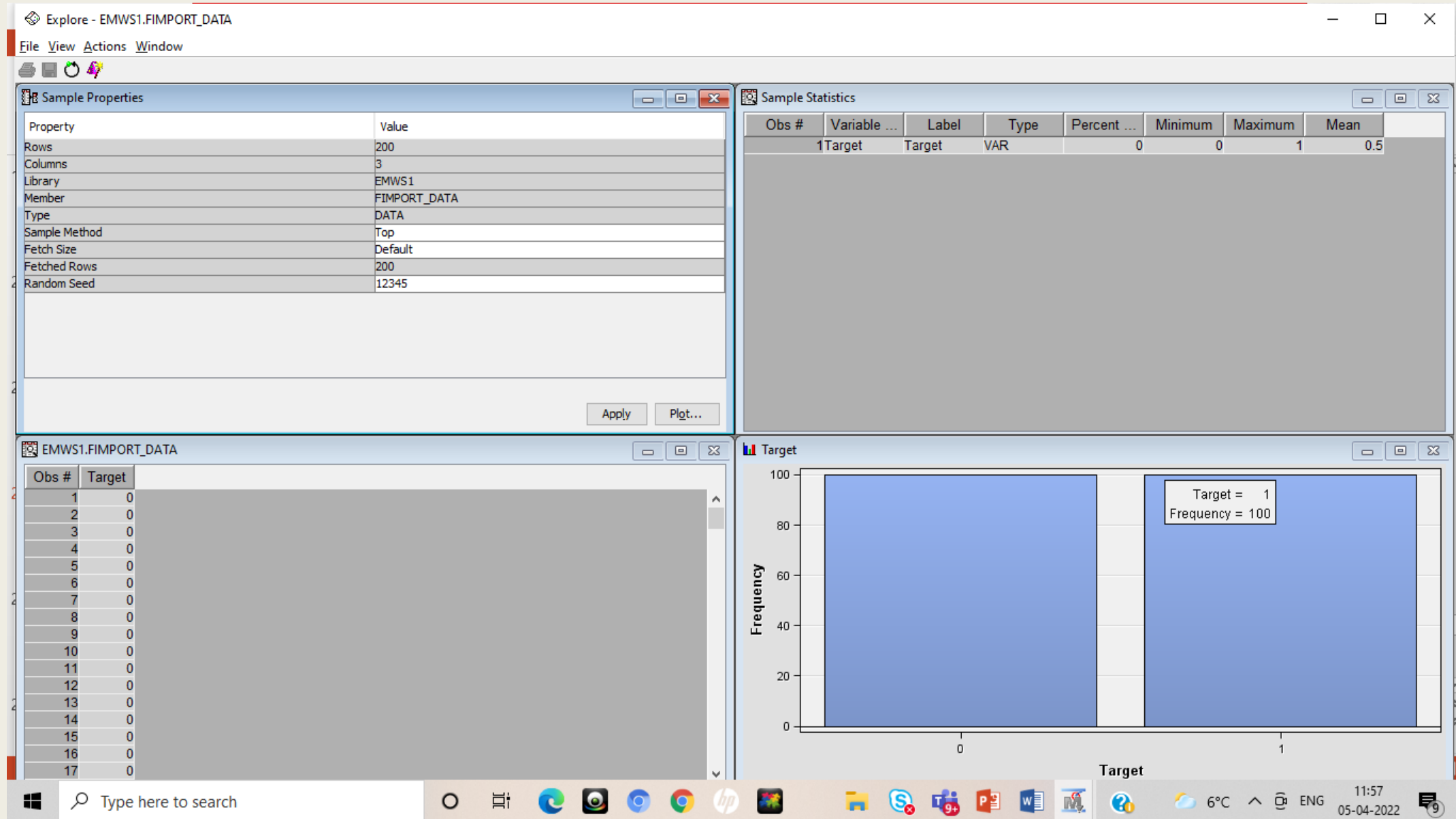
u59397413 as u59397413 Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

Type here to search

5°C 11:52 05-04-2022

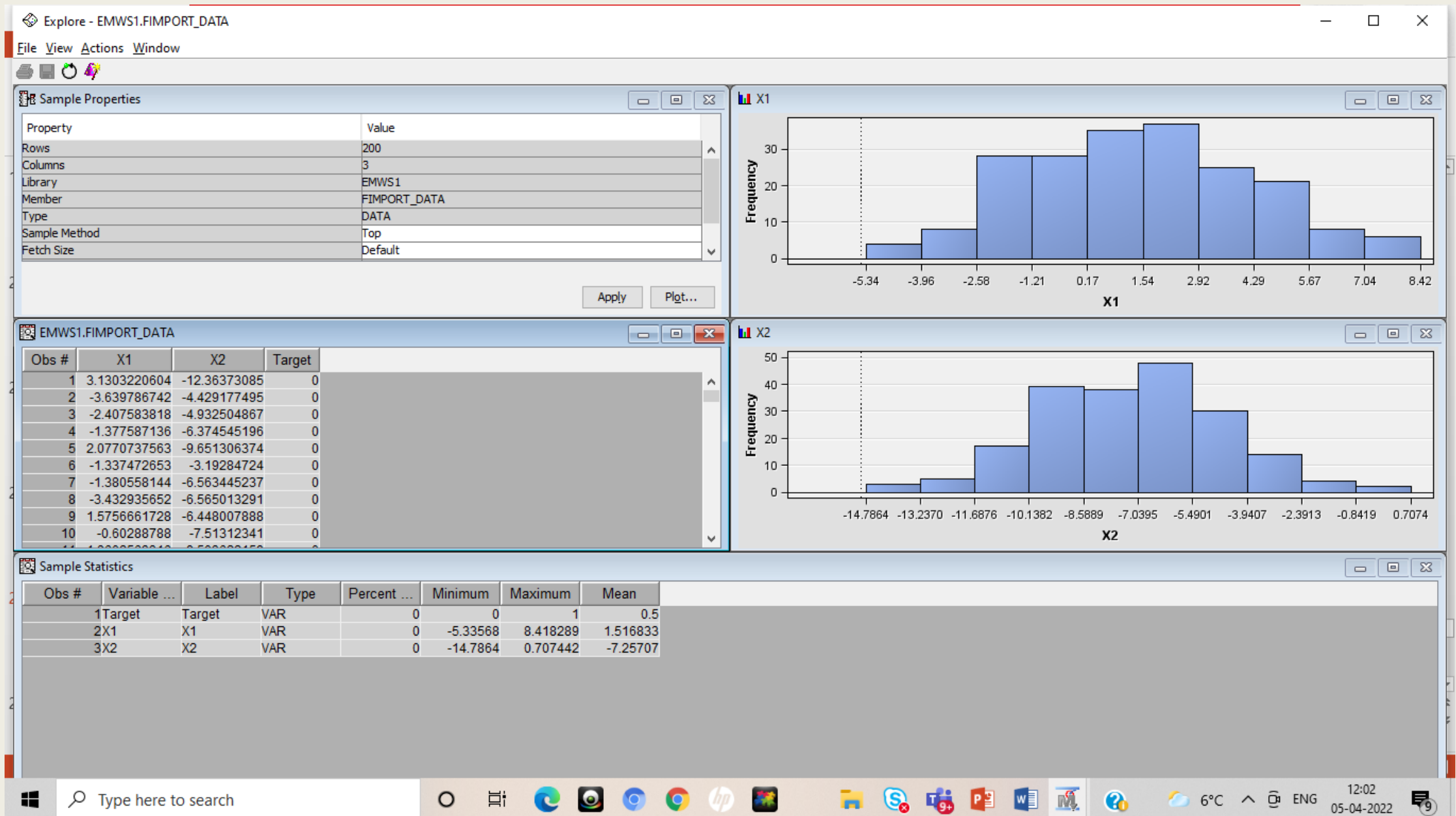
Answer 14 Part 1: How many rows are imported and how many have Target = 1

200 rows were imported out of which 100 have Target = 1



Answer 14 Part 2: Are the distributions of X1 and X2 normal

X1 and X2 are normally distributed



Answer 15: How many data samples are in the training set. What percentage has Target = 1?

There are 158 data samples in the training data set, out which 50 percent have Target = 1

45	Summary Statistics for Class Targets					
46						
47	Data=DATA					
48						
49		Numeric	Formatted	Frequency		
50	Variable	Value	Value	Count	Percent	Label
51						
52	Target	0	0	100	50	Target
53	Target	1	1	100	50	Target
54						
55						
56	Data=TRAIN					
57						
58		Numeric	Formatted	Frequency		
59	Variable	Value	Value	Count	Percent	Label
60						
61	Target	0	0	79	50	Target
62	Target	1	1	79	50	Target
63						
64						
65	Data=VALIDATE					
66						
67		Numeric	Formatted	Frequency		
68	Variable	Value	Value	Count	Percent	Label
69						
70	Target	0	0	21	50	Target
71	Target	1	1	21	50	Target
72						

22	Partition Summary		
23			
24			
25			
26	Type	Data Set	Number of
27			Observations
28	DATA	EMWS1.FIMPORT_train	200
29	TRAIN	EMWS1.Part_TRAIN	158
30	VALIDATE	EMWS1.Part_VALIDATE	42
31			

Answer 18 Part 1: What is the default method under Train?

The default method under Train is RD-TREE

The screenshot displays the SAS Enterprise Miner - Workshop5KNN interface. The left pane shows a project tree with 'Workshop5KNN' containing 'Data Sources', 'Diagrams', and 'Model Packages'. The right pane shows the 'KNN' model configuration. The 'Train' tab is selected, and the 'Method' is set to 'RD-Tree'. The 'Status' tab shows the model was created on 5/4/22 at 4:18 PM.

Property	Value
General	
Node ID	MBR
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Method	RD-Tree
Number of Neighbors	16
Epsilon	0.0
Number of Buckets	8
Weighted	Yes
Create Nodes	No
Create Neighbor Variables	Yes
Status	
Create Time	5/4/22 4:18 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a diagram with three nodes: 'File Import', 'Data Partition', and 'MBR', connected in sequence. The 'Diagram' tab is selected at the bottom.

Diagram KNN opened

u59397413 as u59397413 | Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

12:23
05-04-2022

Answer 18 Part 2: What is the default value for k (number of neighbors)?

The default value of k is 16

The screenshot displays the SAS Enterprise Miner - Workshop5KNN interface. The left pane shows the project structure with 'Workshop5KNN' containing 'Data Sources', 'Diagrams', and 'Model Packages'. The 'KNN' model is selected. The right pane shows the 'KNN' model configuration window with a workflow diagram: 'File Import' -> 'Data Partition' -> 'MBR'.

Property Value

Property	Value
General	
Node ID	MBR
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Method	Scan
Number of Neighbors	16
Epsilon	0.0
Number of Buckets	8
Weighted	No
Create Nodes	No
Create Neighbor Variables	Yes
Status	
Create Time	5/4/22 4:18 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Weighted

Diagram KNN opened

u59397413 as u59397413 Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

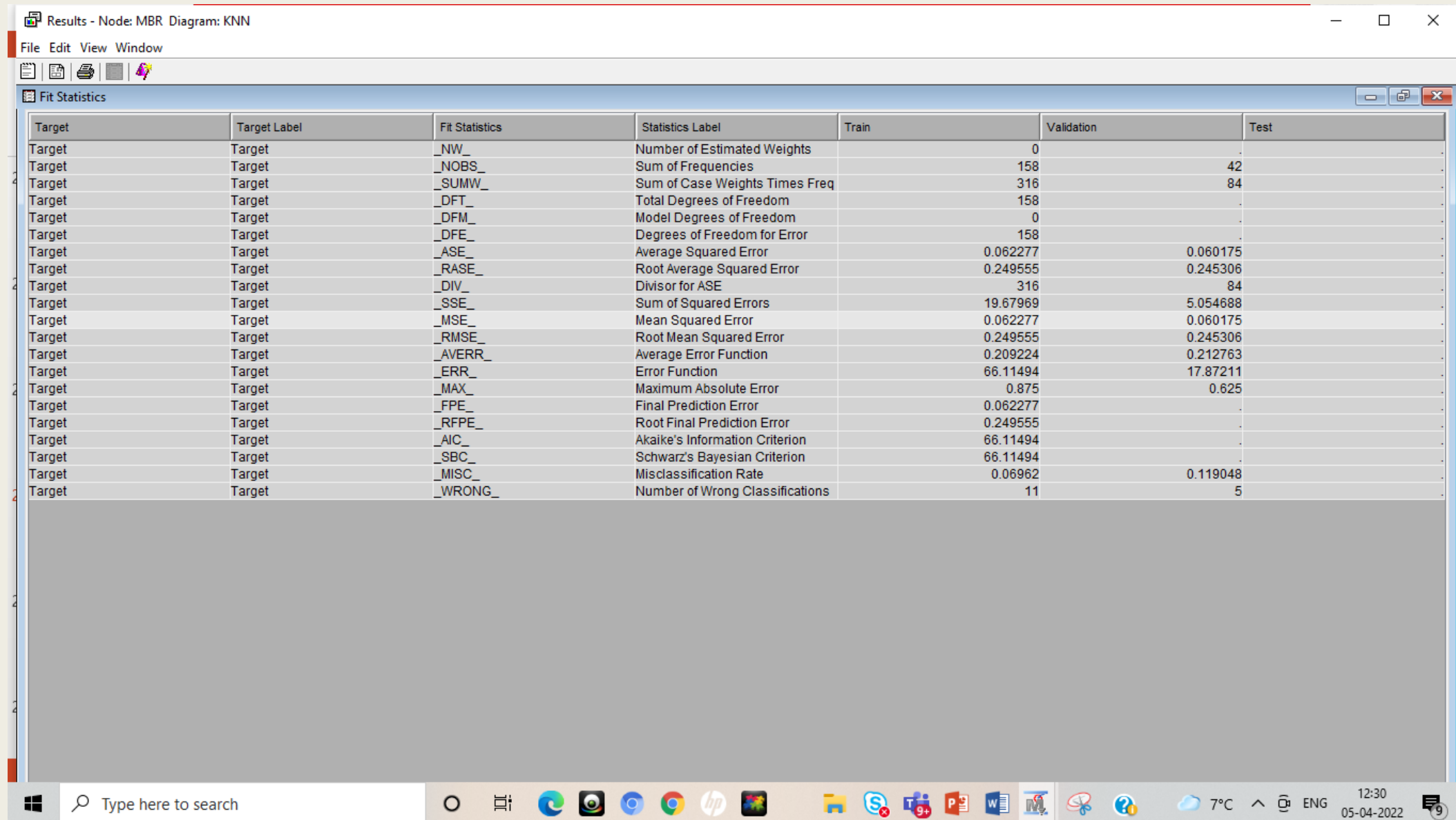
Type here to search

6°C 12:26 05-04-2022

Answer 19 Part 1: What is the MSE for training versus validation data set?

The MSE is as follows:

Train dataset = 0.062277 and Validation dataset = 0.060175



Results - Node: MBR Diagram: KNN

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target	Target	_NW_	Number of Estimated Weights	0		
Target	Target	_NOBS_	Sum of Frequencies	158		42
Target	Target	_SUMW_	Sum of Case Weights Times Freq	316		84
Target	Target	_DFT_	Total Degrees of Freedom	158		
Target	Target	_DFM_	Model Degrees of Freedom	0		
Target	Target	_DFE_	Degrees of Freedom for Error	158		
Target	Target	_ASE_	Average Squared Error	0.062277	0.060175	
Target	Target	_RASE_	Root Average Squared Error	0.249555	0.245306	
Target	Target	_DIV_	Divisor for ASE	316		84
Target	Target	_SSE_	Sum of Squared Errors	19.67969	5.054688	
Target	Target	_MSE_	Mean Squared Error	0.062277	0.060175	
Target	Target	_RMSE_	Root Mean Squared Error	0.249555	0.245306	
Target	Target	_AVERR_	Average Error Function	0.209224	0.212763	
Target	Target	_ERR_	Error Function	66.11494	17.87211	
Target	Target	_MAX_	Maximum Absolute Error	0.875	0.625	
Target	Target	_FPE_	Final Prediction Error	0.062277		
Target	Target	_RFPE_	Root Final Prediction Error	0.249555		
Target	Target	_AIC_	Akaike's Information Criterion	66.11494		
Target	Target	_SBC_	Schwarz's Bayesian Criterion	66.11494		
Target	Target	_MISC_	Misclassification Rate	0.06962	0.119048	
Target	Target	_WRONG_	Number of Wrong Classifications	11		5

Type here to search

12:30 05-04-2022

Answer 19 Part 2: Find the number of FN, FP, etc. for the validation set and paste here

```
.17
.18 Event Classification Table
.19
.20 Data Role=TRAIN Target=Target Target Label=Target
.21
.22     False      True      False      True
.23 Negative Negative Positive Positive
.24
.25      8        76         3        71
.26
.27
.28 Data Role=VALIDATE Target=Target Target Label=Target
.29
.30     False      True      False      True
.31 Negative Negative Positive Positive
.32
.33      3        19         2        18
.34
```

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 18/(18+19) = 0.486$$

$$\text{Precision (P)} = TP/(TP+FP) = 18/(18+2) = 0.90$$

$$F1 = 2P.R/(P+R) = (2*(0.486*0.90))/(0.486+0.90) = 0.631$$

The Precision value is 0.90, which signifies that 90.0% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 63.1%.

Answer 19 Part 3: Run Duration of the model

The run duration of the model is : 0 Hr. 0 Min. 4.91 Sec.

.. Property		Value
General		
Node ID		MBR
Imported Data		
Exported Data		
Notes		
Train		
Variables		
Method		Scan
Number of Neighbors		16
Epsilon		0.0
Number of Buckets		8
Weighted		No
Create Nodes		No
Create Neighbor Variables		Yes
Status		
Create Time		5/4/22 4:18 PM
Run ID		19a14696-6af2-fe45-9fc2-4ad60a3ce9e4
Last Error		
Last Status		Complete
Last Run Time		5/4/22 4:27 PM
Run Duration		0 Hr. 0 Min. 4.91 Sec.
Grid Host		
User-Added Node		No

Answer 20 Part 1: What is the warning and what is the MSE value this time?

The MSE is as follows:

Train dataset = 0 and Validation dataset = 0.119048

Results - Node: MBR Diagram: KNN

File Edit View Window

Warning

- 1 The computed Average Squared Error for Target is 0.
- 2 This indicates a possible target duplication issue.
- 3 Please review the list of inputs used in the model.
- 4

The warning is as shown on the left

Results - Node: MBR Diagram: KNN

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target	Target	_NW_	Number of Estimated Weights	0	.	.
Target	Target	_NOBS_	Sum of Frequencies	158	42	.
Target	Target	_SUMW_	Sum of Case Weights Times Freq	316	84	.
Target	Target	_DFT_	Total Degrees of Freedom	158	.	.
Target	Target	_DFM_	Model Degrees of Freedom	0	.	.
Target	Target	_DFE_	Degrees of Freedom for Error	158	.	.
Target	Target	_ASE_	Average Squared Error	0	0.119048	.
Target	Target	_RASE_	Root Average Squared Error	0	0.345033	.
Target	Target	_DIV_	Divisor for ASE	316	84	.
Target	Target	_SSE_	Sum of Squared Errors	0	10	.
Target	Target	_MSE_	Mean Squared Error	0	0.119048	.
Target	Target	_RMSE_	Root Mean Squared Error	0	0.345033	.
Target	Target	_AVERR_	Average Error Function	0	0.822352	.
Target	Target	_ERR_	Error Function	0	69.07755	.
Target	Target	_MAX_	Maximum Absolute Error	0	1	.
Target	Target	_FPE_	Final Prediction Error	0	.	.
Target	Target	_RFPE_	Root Final Prediction Error	0	.	.
Target	Target	_AIC_	Akaike's Information Criterion	.	.	.
Target	Target	_SBC_	Schwarz's Bayesian Criterion	.	.	.
Target	Target	_MISC_	Misclassification Rate	0	0.119048	.
Target	Target	_WRONG_	Number of Wrong Classifications	0	5	.

Answer 20 Part 2: What is the number of FN, FP, etc. for validation set here. How did the results change?

```
116 Event Classification Table
117
118 Data Role=TRAIN Target=Target Target Label=Target
119
120     False      True      False      True
121 Negative Negative Positive Positive
122
123      .         79         .         79
124
125
126 Data Role=VALIDATE Target=Target Target Label=Target
127
128     False      True      False      True
129 Negative Negative Positive Positive
130
131      1         17         4         20
132
```

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 17/(17+1) = 0.944$$

$$\text{Precision (P)} = TP/(TP+FP) = 17/(17+4) = 0.8095$$

$$F1 = 2P.R/(P+R) = (2*(0.955*0.8095))/(0.8095+0.955) = 0.8664$$

The Precision value is 0.80, which signifies that 80.95% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 86.64% which is the highest as compared to all the models.

Answer 20 Part 3: Run Duration of the model

The run duration of the model is : 0 Hr. 0 Min. 3.56 Sec.

.. Property	Value
General	
Node ID	MBR
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Method	Scan
Number of Neighbors	1
Epsilon	0.0
Number of Buckets	8
Weighted	No
Create Nodes	No
Create Neighbor Variables	Yes
Status	
Create Time	5/4/22 4:18 PM
Run ID	dcf96570-112b-1143-b718-ed6a23e372b8
Last Error	
Last Status	Complete
Last Run Time	5/4/22 4:37 PM
Run Duration	0 Hr. 0 Min. 3.56 Sec.
Grid Host	
User-Added Node	No

Answer 21: What are Eigen Values and What percentage of the energy is explained by the first principal component?

The DMNEURL Procedure

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.64015952	1.28031904	0.8201	0.8201
2	0.35984048		0.1799	1.0000

```
53
54  *-----*
55  Summary of Exported Principal Components
56  *-----*
57
58  Remark:The number of inputs is used as the maximum number of principal components
59
60  Total number of input variables: 2
61  Maximum number cutoff of principal components: 2
62  Cumulative proportional eigenvalue cutoff: 0.99
63  Proportional eigenvalue increment cutoff: 0.001
64  Number of the selected principal components: 2
65  Total variation explained by the selected principal components: 1
66
```

Answer 22: Event Classification Table after Principal Component and k=16

```
117
118 Event Classification Table
119
120 Data Role=TRAIN Target=Target Target Label=Target
121
122     False     True     False     True
123 Negative Negative Positive Positive
124
125      8       75      4       71
126
127
128 Data Role=VALIDATE Target=Target Target Label=Target
129
130     False     True     False     True
131 Negative Negative Positive Positive
132
133      1       20      1       20
134
```

Please find the Calculations:

$$\text{Recall (R)} = TP/(TP+FN) = 20/(20+1) = 0.952$$

$$\text{Precision (P)} = TP/(TP+FP) = 20/(20+1) = 0.952$$

$$F1 = 2P.R/(P+R) = (2*(0.952*0.952))/(0.952+0.952) = 0.952$$

The Precision value is 0.80, which signifies that 95.20% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 95.20% which is better than the previous model.

Answer 23: KNN Diagram and the improvements in results



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

