

**APPLIED ANALYTIC MODELING  
BA 706 FALL 2022  
GROUP PROJECT**

**A FULL DATA MINING AND PREDICTIVE MODELING PROJECT AND REPORT ON  
'EMPLOYEE TURNOVER DATASET'  
( Gotten from kaggle.com )**

**SUBMITTED BY  
GROUP 9**

**IBIYENGHA TOBIN- 301256083  
EMMANUELLA ADEGBENJO-301233296  
IBUKUNOLUWA OLUKOKO-301259629**

## INTRODUCTION

Employee attrition happens when employees resign from a company at a rate higher than the company employs new hires. This can be a concern for a company as the cost implications from brain drain and reputational risk can become a challenge if this phenomenon is not checked.

We are assuming the problem statement is the high turnover of employees in different industries and our mission is to discover the reason for this high turnover rate and give recommendations on how to reduce the risk of losing good hands in the industries by reducing employee turnover.

Our data set, Employee Turnover (Kaggle.com) contains 16 variables and over 1000 observations. There are no missing values in this dataset. We will be predicting the variables that impact an employee's decision to resign from their place of employment.

The data was collected using questionnaires completed by the staff who quit. The employee database was also used to collate historical and present generic information about all employees.

There are several factors that could affect an employee's decision to leave their place of employment. Below is the dictionary showing the description of the variables we will be drawing our conclusions from in this project:

NAME	MODEL ROLE	MEASUREMENT LEVEL	DESCRIPTION
STAG	INPUT	INTERVAL	Experience(time) 0.39 - 179
EVENT	TARGET	BINARY	Employee turnover (0-NO or 1-YES)
GENDER	INPUT	NOMINAL	Employee's gender (Female/Male)
AGE	INPUT	INTERVAL	Employee's Age
INDUSTRY	INPUT	NOMINAL	Employee's Industry
PROFESSION	INPUT	NOMINAL	Employee's profession
TRAFFIC	REJECTED	NOMINAL	From what pipeline did the employee come to the company (empjs, advert, rabrecNErab, youufs,referral,recNErab)
COACH	REJECTED	NOMINAL	Presence of training on probation (YES or NO)
HEAD_GENDER	INPUT	NOMINAL	Supervisor's gender (Female/Male)
GREYWAGE	INPUT	NOMINAL	Salary (White/Grey)
WAY	INPUT	NOMINAL	Employee's way of transportation
EXTRAVERSION	INPUT	INTERVAL	Extraversion score (1- 10)
INDEPEND	INPUT	INTERVAL	Independence score (1-10)
SELFCONTROL	INPUT	INTERVAL	Self-control score (1-10)
ANXIETY	INPUT	INTERVAL	Anxiety score (1-10)
NOVATOR	INPUT	INTERVAL	Innovator score (1-10)

## Description of the Traffic and Greywage variables

TRAFFIC	<i>This describes the means by which the employee learned of the vacancy in the company and how they applied. For example, direct recommendation by a friend who is an employee or through a job site to mention a few. As we have chosen to reject the variable, we will not dwell much on defining the terms.</i>
GREYWAGE	<b>Grey-wage:</b> in Russia or Ukraine means that the employer pays just a tiny bit amount of salary above the white-wage <b>White wage:</b> minimum wage

## DATA SETUP/EXPLORATION

We first imported the **employee dataset** and then made '**event**' our target variable as it is a binary variable that indicates whether an employee resigned from the company or not.

We rejected '**traffic**' and '**coach**' because we do not think that the pipeline in which an employee comes into the company or the presence of training during probation affects an employee's decision to quit. This is our business assumption because we believe whether the person came through referral or not that would not impact on the person's decision to quit.

The screenshot shows the SAS Enterprise Miner interface with the 'Variables - FIMPORT' dialog open. The dialog lists variables with their properties:

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
age	Input	Interval	No	No	.	.	.
seniority	Input	Interval	No	No	.	.	.
coach	Rejected	Nominal	No	No	.	.	.
event	Target	Binary	No	No	.	.	.
extraversion	Input	Interval	No	No	.	.	.
openness	Input	Nominal	No	No	.	.	.
greywage	Input	Nominal	No	No	.	.	.
head	header	Input	Nominal	No	.	.	.
independ	Input	Interval	No	No	.	.	.
industry	Input	Nominal	No	No	.	.	.
novator	Input	Interval	No	No	.	.	.
profession	Input	Nominal	No	No	.	.	.
selfcontrol	Input	Interval	No	No	.	.	.
stage	Input	Interval	No	No	.	.	.
traffic	Rejected	Nominal	No	No	.	.	.
way	Input	Nominal	No	No	.	.	.

The 'Properties' panel on the left shows settings like 'Import File' (H:\BA706F22\BA706DEMO.csv), 'Maximum Rows' (1000000), and 'File Type' (csv). The status bar at the bottom right shows 'Connected to ClassApps-31' and the date '12/12/2022'.

A data partition node was connected to the dataset with the below allocations:

Data Partition	Allocation	# of observations
Training	50%	564
Validation	50%	565
Test	0	0

The screenshot shows the SAS Enterprise Miner interface. On the left, the project tree displays a folder named 'BA706DEMO' containing 'Data Sources', 'Diagrams', and 'Predictive Analytics'. A 'Employee' diagram is selected. On the right, the 'Results - Node: Data Partition Diagram: Employee' window is open, showing the 'Output' tab. The output pane displays the following text:

```

1 * -----
2 User: 301259629
3 Date: December 06, 2022
4 Time: 15:16:28
5 -----
6 * Training Output
7 -----
8 -----
9 -----
10 -----
11 -----
12 Variable Summary
13 -----
14 Role Measurement Frequency Count
15 -----
16 INPUT INTERVAL 7
17 INPUT NOMINAL 6
18 REJECTED NOMINAL 2
20 TARGET BINARY 1
21 -----
22 -----
23 -----
24 -----
25 Partition Summary
26 -----
27 TYPE Data Set Number of Observations
28 -----
29 DATA EMPS1.FINP007_train 1129
30 TRAIN EMPS1.Part_TRAIN 564
32 VALIDATE EMPS1.Part_VALIDATE 565
33 -----
34 -----
35 -----
36 * Score Output
37 -----
38 -----
39 -----
40 -----
41 -----
42 -----
43 -----
44 -----
45 -----
46 -----
47 -----
48 -----
49 -----
50 -----
51 -----
52 -----
53 -----
54 -----
55 -----
56 -----
57 -----
58 -----
59 -----
60 -----
61 -----
62 -----
63 -----
64 -----
65 -----
66 -----
67 -----
68 -----
69 -----
70 -----
71 -----
72 -----
73 -----
74 -----
75 -----
76 -----
77 -----
78 -----
79 -----
80 -----
81 -----
82 -----
83 -----
84 -----
85 -----
86 -----
87 -----
88 -----
89 -----
90 -----
91 -----
92 -----
93 -----
94 -----
95 -----
96 -----
97 -----
98 -----
99 -----
100% 301259629 as 301259629 Connected to ClassApps-31

```

Below the results window, the status bar shows 'Diagram Employee opened' and the system clock '4:36 PM 12/12/2022'.

## DECISION TREES

The first set of models we used are decision trees. The trees are classified based on the method used , the number of branches, and the assessment measure. The following tree models were built:

1. Maximal Trees( Two and Three Branch)
2. Classification Trees( Two and Three Branch)
3. Probability Trees(Two and Three Branch)

### TWO-BRANCH MAXIMAL TREE

The decision tree node was attached to the data partition node with a maximum of two branches indicated, ‘largest’ used as the method and ‘decision’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Data Sources
- Business
- Employees
- GAME
- Loan
- Predictive Analytics
- Model Students
- Model Packages

Property Value

- Panel Size: Large
- Panel Search: No
- User Decisions: No
- Use Folds: No
- Iterations: 20000
- Size Sample: 20000
- Subtrees: 1
- Method: Largest
- Number of Leaves: 7
- Assessment Measure: Decision
- Assessment Fraction: 0.25
- Cross Validation: No
- Number of Folds: 10
- Number of Repeats: 1
- Seed: 12345
- Observation Based Impute
- Observation Based Impute
- Number Single Var Imputes: 0
- Number of Subsets: 10
- Number of Repeats: 1
- Seed: 12345
- General Properties

Diagram Log

Type here to search

SAS I.T. 2022-12-0... Balance f... Enterprise... Enterprise... Enterprise... Results ... ENG 2:43 PM US 14/12/2022

Score Rankings Overlay: event

Leaf Statistics

Fit Statistics

Tree

Output

Tremap

Diagram Log

Type here to search

SAS I.T. 2022-12-0... Balance f... Enterprise... Enterprise... Enterprise... Results ... ENG 2:43 PM US 14/12/2022

Results - Node: Max Decision Tree Diagram: Employee

File Edit View Window

Tree

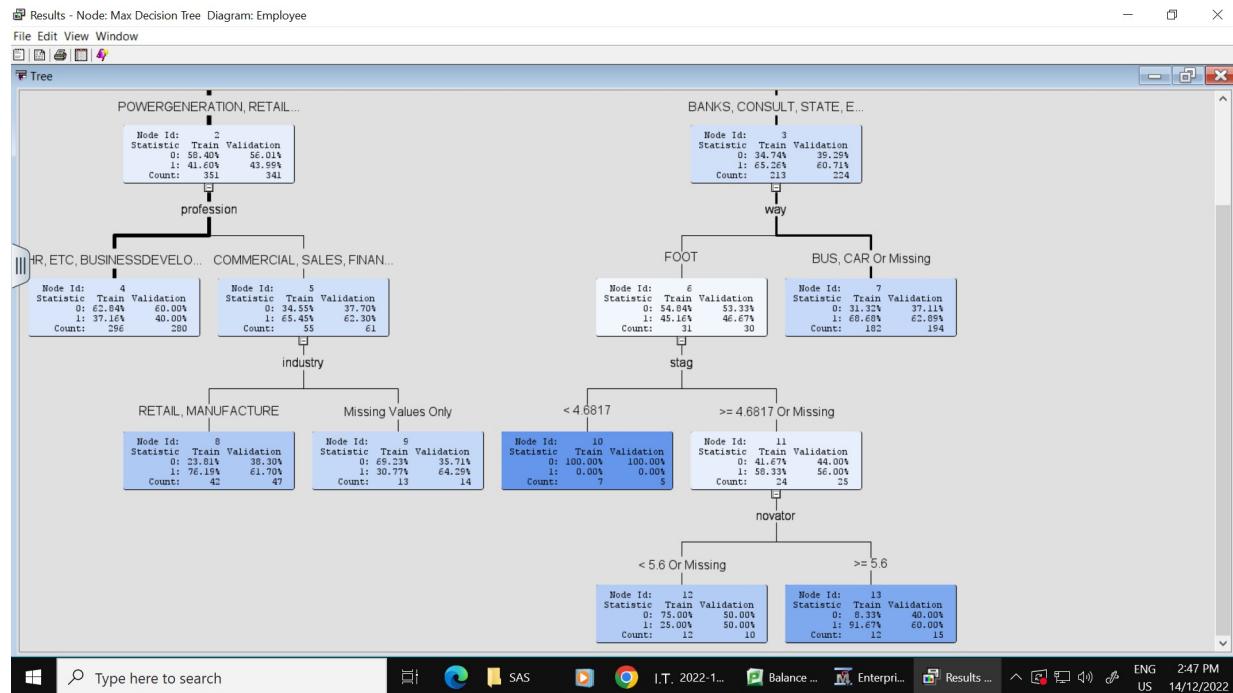
```

graph TD
    Node1[Node Id: 1 Statistic: Train Validation O: 45.47% D: 55.40% I: 41.60% C: 351] -- industry --> Node2[Node Id: 2 Statistic: Train Validation O: 56.01% D: 58.40% I: 43.99% C: 341]
    Node1 -- industry --> Node3[Node Id: 3 Statistic: Train Validation O: 35.29% D: 34.74% I: 65.26% C: 213]
    Node2 -- profession --> Node4[Node Id: 4 Statistic: Train Validation O: 60.00% D: 58.40% I: 37.18% C: 266]
    Node2 -- profession --> Node5[Node Id: 5 Statistic: Train Validation O: 37.30% D: 65.45% I: 62.30% C: 55]
    Node3 -- way --> Node6[Node Id: 6 Statistic: Train Validation O: 55.21% D: 45.16% I: 46.87% C: 31]
    Node3 -- way --> Node7[Node Id: 7 Statistic: Train Validation O: 27.06% D: 26.89% I: 68.68% C: 182]
    Node4 -- industry --> Node8[Node Id: 8 Statistic: Train Validation O: 38.00% D: 37.81% I: 76.19% C: 42]
    Node4 -- industry --> Node9[Node Id: 9 Statistic: Train Validation O: 38.00% D: 35.71% I: 60.77% C: 47]
    Node5 -- industry --> Node10[Node Id: 10 Statistic: Train Validation O: 100.00% D: 100.00% I: 0.00% C: 7]
    Node6 -- stag --> Node11[Node Id: 11 Statistic: Train Validation O: 44.00% D: 31.33% I: 68.33% C: 24]
    Node6 -- stag --> Node12[Node Id: 12 Statistic: Train Validation O: 44.00% D: 31.33% I: 68.33% C: 25]
    Node7 -- <= 4 6817 --> Node13[Node Id: 13 Statistic: Train Validation O: 40.00% D: 35.71% I: 64.00% C: 14]
    Node7 -- >= 4 6817 --> Node14[Node Id: 14 Statistic: Train Validation O: 40.00% D: 35.71% I: 64.00% C: 13]
  
```

Diagram Log

Type here to search

SAS I.T. 2022-12-0... Balance f... Enterprise... Enterprise... Enterprise... Results ... ENG 2:47 PM US 14/12/2022



Results - Node: Max Decision Tree Diagram: Employee

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	NOBS		Sum of Frequencies	564	565	.
event	MSE		Misclassification Rate	0.328014	0.39392	.
event	MAX		Maximum Absolute Error	0.916667	0.916667	.
event	SSE		Sum of Squared Errors	243.6498	276.9288	.
event	ASE		Average Squared Error	0.216002	0.245057	.
event	RASE		Root Mean Squared Error	0.462776	0.495045	.
event	DIV		Divisor for ASE	1128	1130	.
event	DFT		Total Degrees of Freedom	564		.

Type here to search

I.T. 2022-1... SAS Balance ... Enterprise... Results ... ENG US 2:47 PM 14/12/2022

```

Results - Node: Max Decision Tree Diagram: Employee
File Edit View Window
Output
57 -----
58
59
60
61 Variable Importance
62
63
64 Number of
65 Variable Splitting
66 Name Label Rules Importance Validation Ratio of
67
68 industry 2 1.0000 1.0000 1.0000
69 profession 1 0.6261 0.8761 1.3993
70 innovator 1 0.5306 0.0600 0.0000
71 stag 1 0.4413 0.6450 1.4616
72 way 1 0.3934 0.3931 1.0045
73
74
75
76 Tree Leaf Report
77
78 Training Validation Validation
79 Node Observations Percent 1 Observations Percent 1
80 Id Depth
81
82 4 2 296 0.37 280 0.40
83 7 2 182 0.69 194 0.65
84 8 3 42 0.76 47 0.62
85 9 3 13 0.31 14 0.64
86 12 4 12 0.25 10 0.50
87 13 4 12 0.92 15 0.60
88 10 3 7 0.00 5 0.00
89
90
91
92
93 Fit Statistics
94
95 Target=event Target Label=' '
96
97 Fit
98 Statistics Statistics Label Train Validation
99

```

The tree had seven leaves with five variable splits. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their jobs. 60% of employees with innovator score of greater than 5.6 who have worked greater than 4.7 months in the Bank, Consult, and State industries are likely to quit. 62% of Commercial, Sales, and Finance professionals in the power generation industry are likely to quit.

The Validation Average Square Error for the two-branch Maximal Tree is 0.24507.

## THREE-BRANCH MAXIMAL TREE

The decision tree node was attached to the data partition node with a maximum of three branches indicated, ‘largest’ used as the method and ‘decision’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Data Sources
- Diagrams
  - Employee
  - GAME
  - Loan
  - Predictive Analytics
  - Student
- Model Packages

Results - Node: Max Decision Tree 3 maxB Diagram: Employee

Leaf Statistics

Leaf Node	Training Percent	Validation Percent
1	~0.85	~0.75
2	~0.80	~0.70
3	~0.75	~0.65
4	~0.70	~0.60
5	~0.65	~0.55
6	~0.60	~0.50
7	~0.55	~0.45
8	~0.50	~0.40
9	~0.45	~0.35

Tree

Score Rankings Overlay: event

Cumulative Lift

Fit Statistics

Target	Target Label	Fit Statistics	Statistics	Train	Validation	Test
event	NOBS	Sum of Fr...	564	565		
event	MISC	Misclassif...	0.324468	0.40177		
event	MAX	Maximum ...	0.857143	0.857143		

Treemap

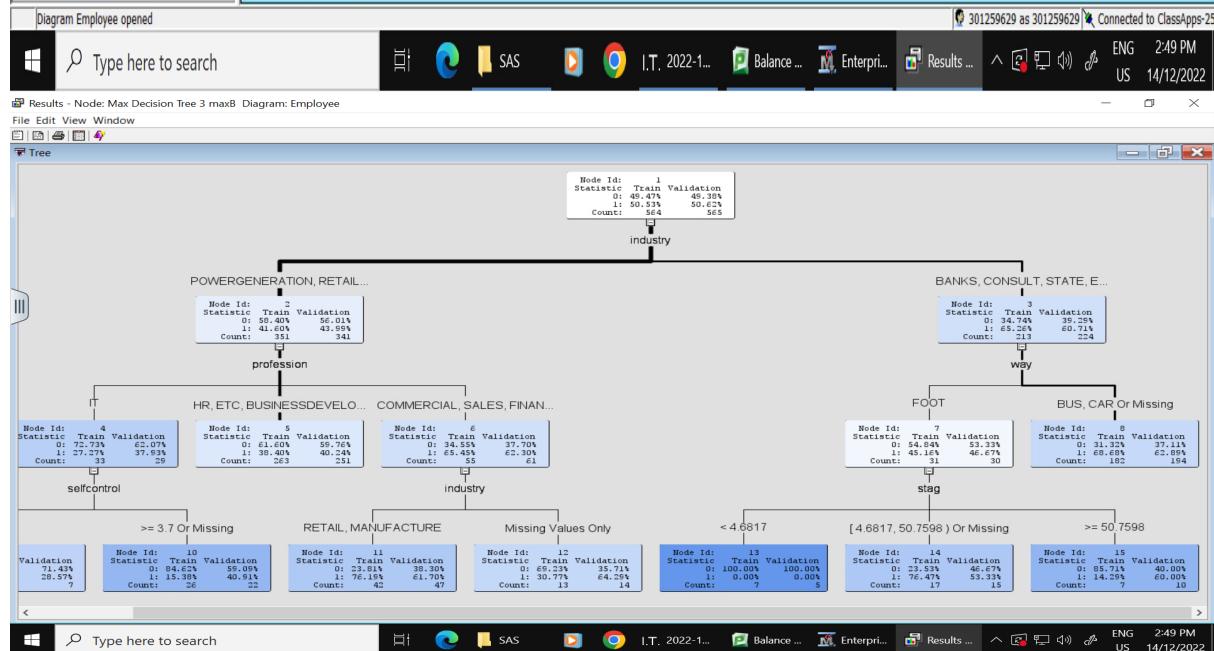
Output

```

1 -----
2 User: 301239629
3 Date: December 14, 2022
4 Time: 11:52:31
5 -----

```

Diagram Log



Results - Node: Max Decision Tree 3 maxB Diagram: Employee

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		N OBS	Sum of Frequencies	564	565	
event		MISC	Misclassification Rate	0.324468	0.40177	
event		MAX	Maximum Absolute Error	0.857143	0.857143	
event		SSE	Sum of Squared Errors	240.9574	283.4896	
event		A SE	Average Squared Error	0.213816	0.250876	
event		RASE	Root Average Squared Error	0.462185	0.500875	
event		DIV	Divisor for A SE	1128	1130	
event		DFT	Total Degrees of Freedom	564		

Type here to search

SAS I.T. 2022-1... Balance... Enterprise... Results... ENG US 2:49 PM 14/12/2022

Results - Node: Max Decision Tree 3 maxB Diagram: Employee

File Edit View Window

Output

```

61 Variable Importance
62
63
64 Number of
65 Variable Splitting Ratio of
66 Name Label Rules Importance Validation Validation to Training
67
68 industry 2 1.0000 1.0000 1.0000
69 profession 1 0.6560 0.8436 1.3859
70 stac 1 0.6302 0.0000 0.0000
71 selfcontrol 1 0.4277 0.0000 0.0000
72 way 1 0.3934 0.3951 1.0045

```

Tree Leaf Report

```

76
77
78 Training Validation Validation
79 Node Id Depth Observations Percent 1 Observations Percent 1
80
81
82 5 2 263 0.38 251 0.40
83 8 2 182 0.69 194 0.63
84 11 3 42 0.76 47 0.62
85 10 3 26 0.15 22 0.41
86 14 3 17 0.76 15 0.53
87 12 3 13 0.31 14 0.64
88 9 3 7 0.71 7 0.29
89 13 3 7 0.00 5 0.00
90 15 3 7 0.14 10 0.60

```

Fit Statistics

```

95 Target=event Target Label=' '
96
97 Fit
100 Statistics Statistics Label Train Validation
101 _N OBS_ Sum of Frequencies 564.00 565.00
102 _MISC_ Misclassification Rate 0.32 0.40

```

Type here to search

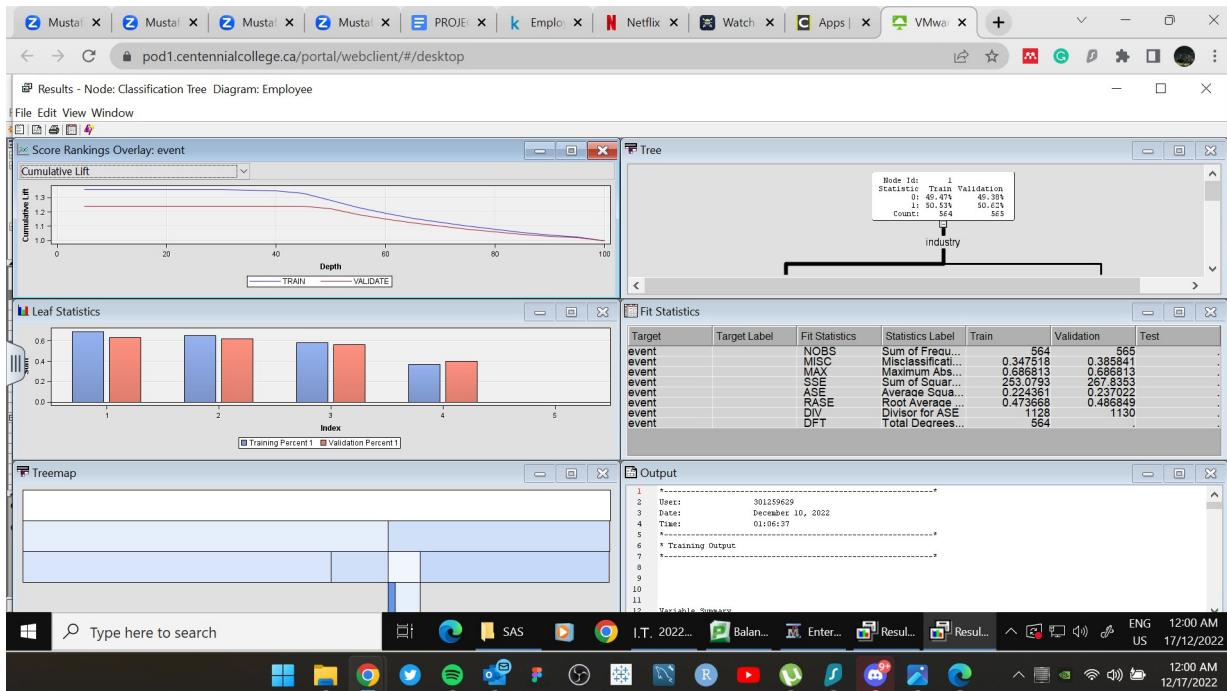
SAS I.T. 2022-1... Balance... Enterprise... Results... ENG US 2:50 PM 14/12/2022

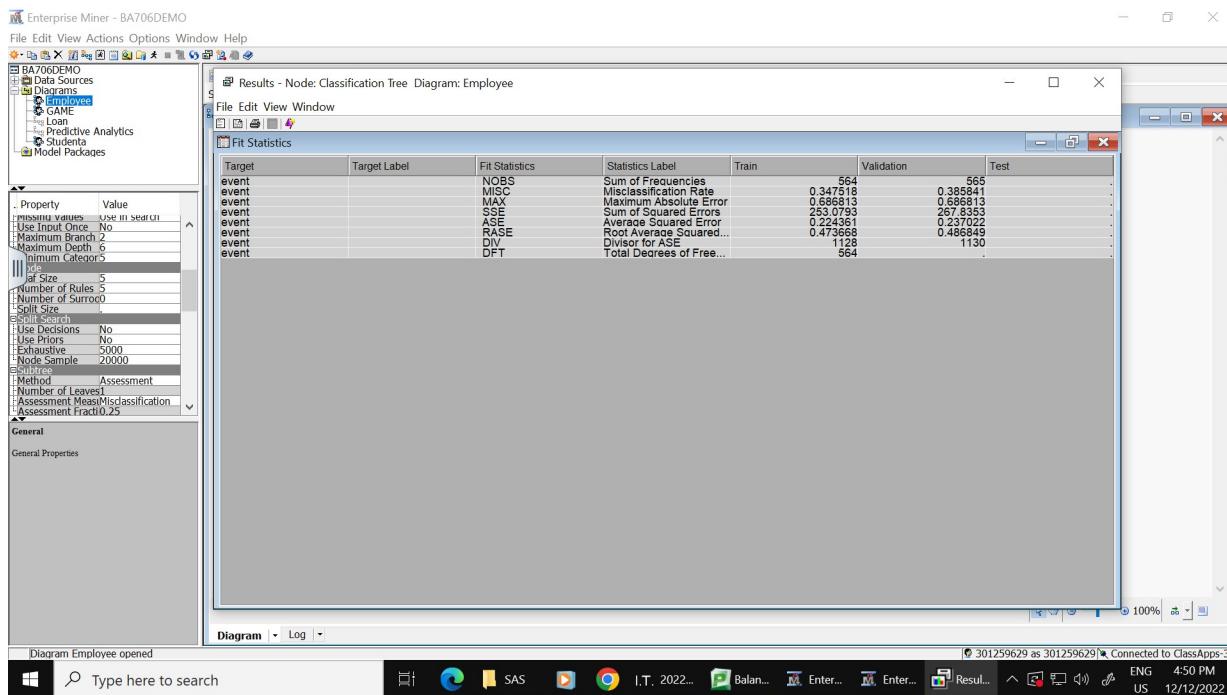
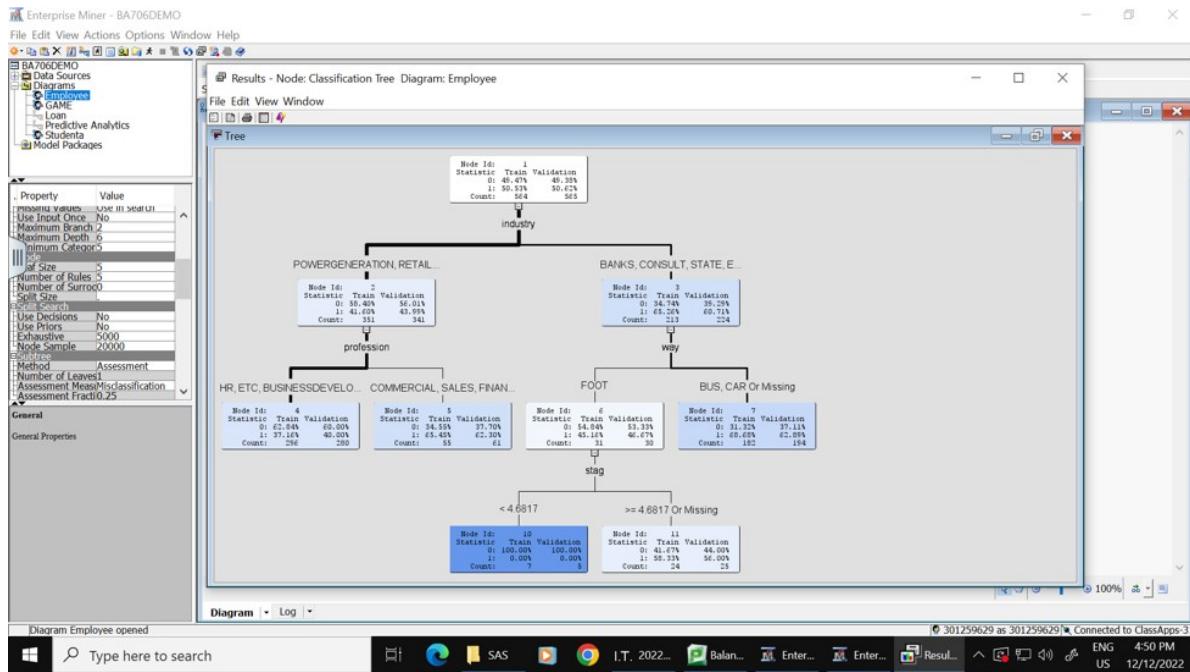
The tree had nine leaves with five variable splits. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their jobs. 60% of employees who have worked greater than 51 months in the Bank, Consult, and State industries and walk to the office are likely to quit. 62.3% of Commercial, Sales, and Finance professionals in the power generation industry are likely to quit.

The Validation Average Square Error for the three-branch Maximal Tree is 0.250876. We observed that the Two Branch Maximal Tree is a better model than the Three Branch Maximal tree, as this has a higher Validation Average Square Error.

## TWO-BRANCH CLASSIFICATION TREE

The decision tree node was attached to the data partition node with a maximum of two branches indicated, ‘assessment’ was used for the method and ‘misclassification’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.





The screenshot shows the SAS Enterprise Miner interface. On the left, a tree diagram labeled 'Employee' is visible. In the center, a window titled 'Results - Node: Classification Tree Diagram: Employee' displays the output of the classification tree. The output includes several tables and a tree leaf report.

```

42 PREDICTED F_event Predicted: event=1
43 RESIDUAL R_event Residual: event=1
44 PREDICTED F_event Predicted: event=0
45 RESIDUAL R_event Residual: event=0
46 PRED F_event From: event
47 INTO I_event Into: event
48
49
50 *-----*
51 * Root Output
52 *-----*
53
54
55 *-----*
56 * Report Output
57 *-----*
58
59
60
61 Variable Importance
62
63
64 Variable Importance
65 Number of Splitting Rules
66 Name Label Importance Validation Importance Ratio of Validation to Training Importance
67
68 industry 1 1.0000 1.0000 1.0000
69 profession 1 0.7073 0.8761 1.2388
70 staq 1 0.4984 0.6450 1.2939
71 war 1 0.4443 0.3951 0.8892
72
73
74 Tree Leaf Report
75
76
77 Node Depth Training Observations Percent 1 Validation Observations Percent 1
78 ID
79 1 2 296 0.37 280 0.40
80
81 4 2 162 0.59 154 0.63
82 3 2 162 0.41 151 0.43
83 5 1 76 0.21 71 0.23

```

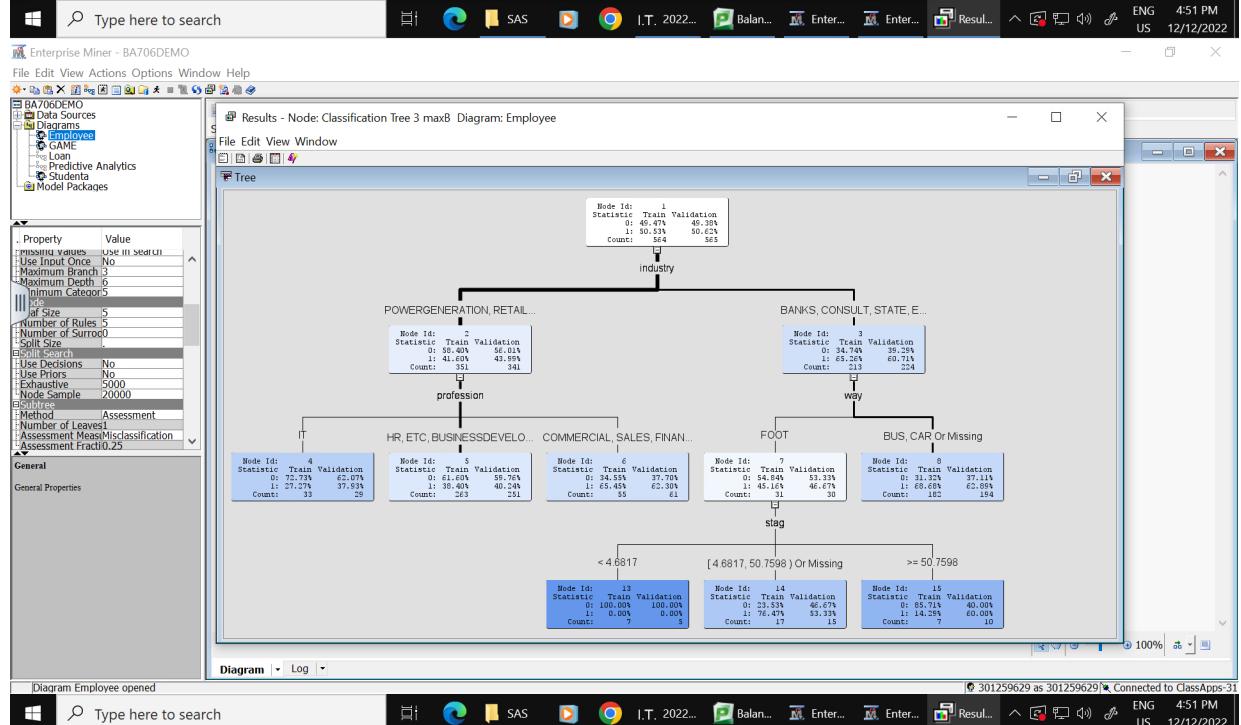
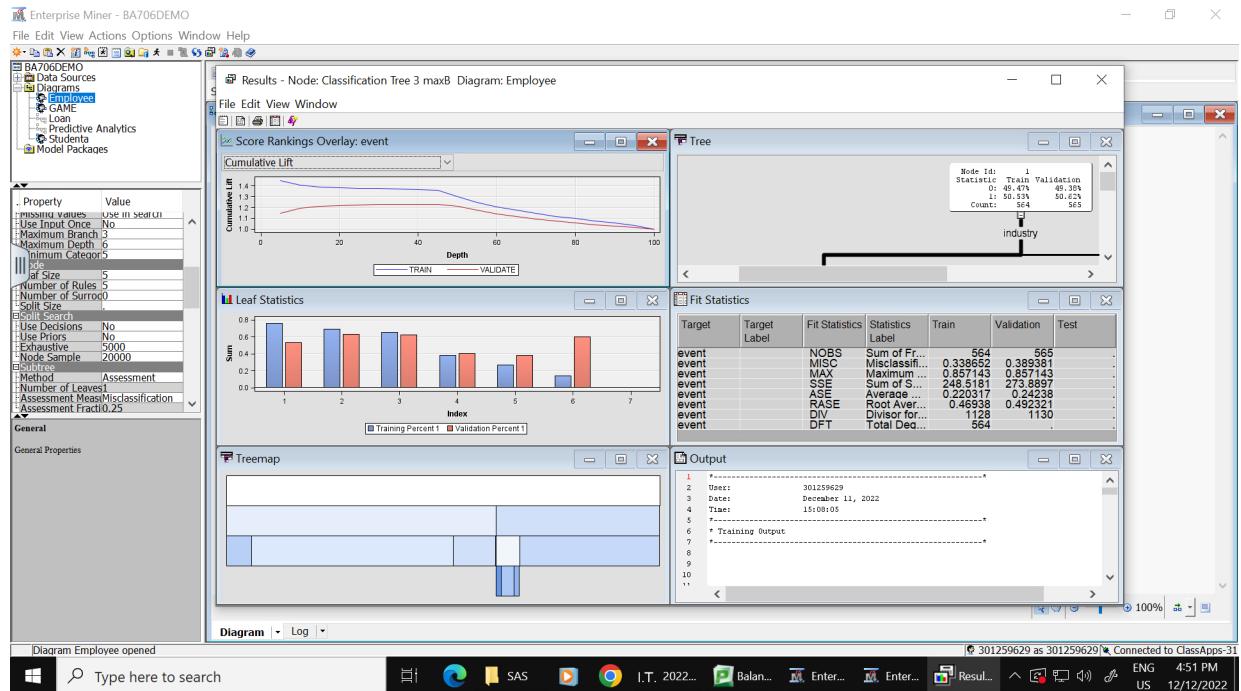
Diagram Employee opened

The tree has five leaves with four variable splits. This reduction is a result of pruning using the misclassification measure. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees in Banks, Consult and State industries who go to work in vehicles are likely to quit their jobs. 56% of employees who have worked greater than 4.7 months in the Bank, Consult, and State industries and walk to the office are likely to quit. 62.3% of Commercial, Sales, and Finance professionals in the power generation industry are also likely to quit.

The Validation Average Square Error for the two-branch classification Tree is 0.237022.

## THREE-BRANCH CLASSIFICATION TREE

The decision tree node was attached to the data partition node with a maximum of three branches indicated, ‘assessment’ used as the method and ‘misclassification’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.



The screenshot shows the SAS Enterprise Miner interface with a classification tree results window open. The window displays the following output:

```

Results - Node: Classification Tree 3 maxB. Diagram: Employee
File Edit View Window Help
SAS 301259629 as 301259629 Connected to ClassApps-31
Diagram Employee opened
Type here to search 4:52 PM 12/12/2022
File Edit View Window
Output
44 PREDICTED  P_event0 Predicted: event=0
45 RESIDUAL   P_event0 Residual: event=0
46 FROM      P_event From: event
47 INTO      I_event Into: event
48
49 *
50 * Score Output
51 *
52 *
53 *
54 *
55 * Report Output
56 *
57 *
58 *
59 *
60 Variable Importance
61 Variable Number of Splitting Validation Ratio of
62 Name    Label Rules Importance Importance to Training
63           Importance
64 industry      1 1.0000 1.0000 1.0000
65 profession   1 0.7410 0.8436 1.1384
66 stag         1 0.7119 0.0000 0.0000
67 way          1 0.4443 0.3951 0.8892
68
69
70
71
72
73
74
75 Tree Leaf Report
76
77 Node Training Validation Validation
78 Id Depth Observations Percent Observations Percent
79
80
81 5   2   263  0.58  251  0.40
82 6   2   182  0.49  194  0.43
83 6   2   55   0.65  41   0.62
84 4   2   33   0.27  29   0.38
85 1,0 5   17   0.76  14   0.43

```

The tree has seven leaves with four variable splits. The increased number of branches also increased the number of leaves when compared to the two branch model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their jobs. 60% of employees who have worked greater than 51 months in the Bank, Consult, and State industries and walk to the office are likely to quit. 62.3% of Commercial, Sales, and Finance professionals in the power generation industry are likely to quit.

The Validation Average Square Error for the three-branch classification Tree is 0.24238. The Two Branch Classification Tree is a better model than the Three Branch Classification tree, as it has a lower Validation Average Square Error.

## TWO BRANCH PROBABILITY TREE

The decision tree node was attached to the data partition node with a maximum branch of two, ‘assessment’ used as the method and ‘average squared error’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.

The screenshot displays the Enterprise Miner software interface with the title "Enterprise Miner - BA706DEMO". The menu bar includes File, Edit, View, Actions, Options, Window, Help. The left sidebar shows a tree structure with nodes: Data Sources, Diagrams, Employee, Predictive Analytics, Loan, Students, and Model Packages. A property grid on the left lists parameters like Property, Value, and Status. Several windows are open: 1) "Results - Node: Probability tree Diagram: Employee" showing a "Score Rankings Overlay: event" plot with Cumulative Lift vs Depth (TRAIN in blue, VALIDATE in red). 2) "Tree" window showing a decision tree structure with Node Id 1, Statistics, Train Validation, and Count details. 3) "Leaf Statistics" window showing a bar chart of NOBS vs Index (1-5) for Training and Validation Percentages. 4) "Fit Statistics" window listing Target, Target Label, Fit Statistics, Statistics, Train, Validation, and Test values for various metrics like NOBS, MISC, MAX, SAE, RASE, DIV, and DFT. 5) "Trellimap" window showing a treemap visualization. 6) "Output" window displaying log entries. The bottom status bar shows "Diagram Employee opened" and "301259629 as 301259629 Connected to ClassApps-31".

The screenshot shows the Enterprise Miner interface with the following details:

- File Edit View Actions Options Window Help** menu bar.
- Data Sources**, **Diagrams**, **Predictive Analytics**, **Student**, and **Model Packages** in the left navigation pane.
- Properties** panel on the left with various settings like **Use Input Once**, **Maximum Branch**, **Maximum Depth**, and **Minimum Category**.
- Diagram Employee opened** at the bottom left.
- Diagram** and **Log** buttons at the bottom center.
- Diagram Employee opened** at the bottom right.

The main area displays a **Probability tree** diagram for the **Employee** dataset:

- Root Node (Node Id: 1)**: Statistic: Train Validation O: 45.47% 49.38% I: 50.13% 54.21% Count: 364. The node is labeled **industry**.
- Left Branch (Node Id: 2)**: Statistic: Train Validation O: 50.40% 56.01% I: 48.01% 43.34% Count: 351. The node is labeled **POWERGENERATION, RETAIL...**.
- Right Branch (Node Id: 3)**: Statistic: Train Validation O: 34.74% 39.29% I: 65.13% 60.71% Count: 213. The node is labeled **BANKS, CONSULT, STATE, E...**.
- Left Node (Node Id: 4)**: Statistic: Train Validation O: 62.84% 60.00% I: 37.16% 39.00% Count: 256. The node is labeled **HR, ETC, BUSINESSEVELOP...**.
- Right Node (Node Id: 5)**: Statistic: Train Validation O: 34.55% 37.70% I: 65.45% 62.19% Count: 155. The node is labeled **COMMERCIAL, SALES, FINAN...**.
- Left Node (Node Id: 6)**: Statistic: Train Validation O: 54.84% 53.33% I: 45.16% 46.67% Count: 31. The node is labeled **FOOT**.
- Right Node (Node Id: 7)**: Statistic: Train Validation O: 31.32% 37.11% I: 68.67% 62.88% Count: 182. The node is labeled **BUS, CAR Or Missing**.
- Left Node (Node Id: 8)**: Statistic: Train Validation O: 100.00% 100.00% I: 0.00% 0.00% Count: 7. The node is labeled **stag**.
- Left Leaf Node (Node Id: 10)**: Statistic: Train Validation O: 100.00% 100.00% I: 0.00% 0.00% Count: 5.
- Right Leaf Node (Node Id: 11)**: Statistic: Train Validation O: 41.67% 44.00% I: 58.33% 56.00% Count: 24.

Enterprise Miner - BA706DEMO

File Edit View Actions Window Help

Data Sources Diagrams Employee Loan Predictive Analytics Model Packages

Properties

- Pruning values: Use in Search
- Use Input Once: No
- Maximum Branch: 2
- Maximum Depth: 6
- Minimum Category: 5
- of Size: 5
- Number of Rules: 5
- Number of Surro: 0
- Split Size: L
- Split Search: Split Search
- Use Decisions: No
- Use Priors: No
- Exhaustive: 5000
- Node Sample: 20000
- Method: Assessment
- Number of Leaves: 1
- Assessment Meas: Average Square Error
- Assessment Fract: 0.25

General Properties

Diagram Employee opened

Results - Node: Probability tree Diagram: Employee

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		NOBS	Sum of Frequencies	564	565	.
event		MSE	Misclassification Rate	0.347818	0.385111	0.686813
event		MAX	Maximum Absolute Error	0.686813	0.686813	0.686813
event		SSE	Sum of Squared Errors	253.0793	267.8353	.
event		ASE	Average Squared Error	0.224361	0.237022	0.247669
event		DSE	Root Mean Squared...	0.473669	0.486848	.
event		DIV	Divisor for ASE	1128	1130	.
event		DFT	Total Degrees of Free...	564	.	.

Diagram | Log | 301259629 as 301259629 Connected to ClassApps-31 ENG 4:53 PM US 12/12/2022

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

Data Sources Diagrams Employee Loan Predictive Analytics Model Packages

Properties

- Pruning values: Use in Search
- Use Input Once: No
- Maximum Branch: 2
- Maximum Depth: 6
- Minimum Category: 5
- of Size: 5
- Number of Rules: 5
- Number of Surro: 0
- Split Size: L
- Split Search: Split Search
- Use Decisions: No
- Use Priors: No
- Exhaustive: 5000
- Node Sample: 20000
- Method: Assessment
- Number of Leaves: 1
- Assessment Meas: Average Square Error
- Assessment Fract: 0.25

General Properties

Diagram Employee opened

Results - Node: Probability tree Diagram: Employee

Output

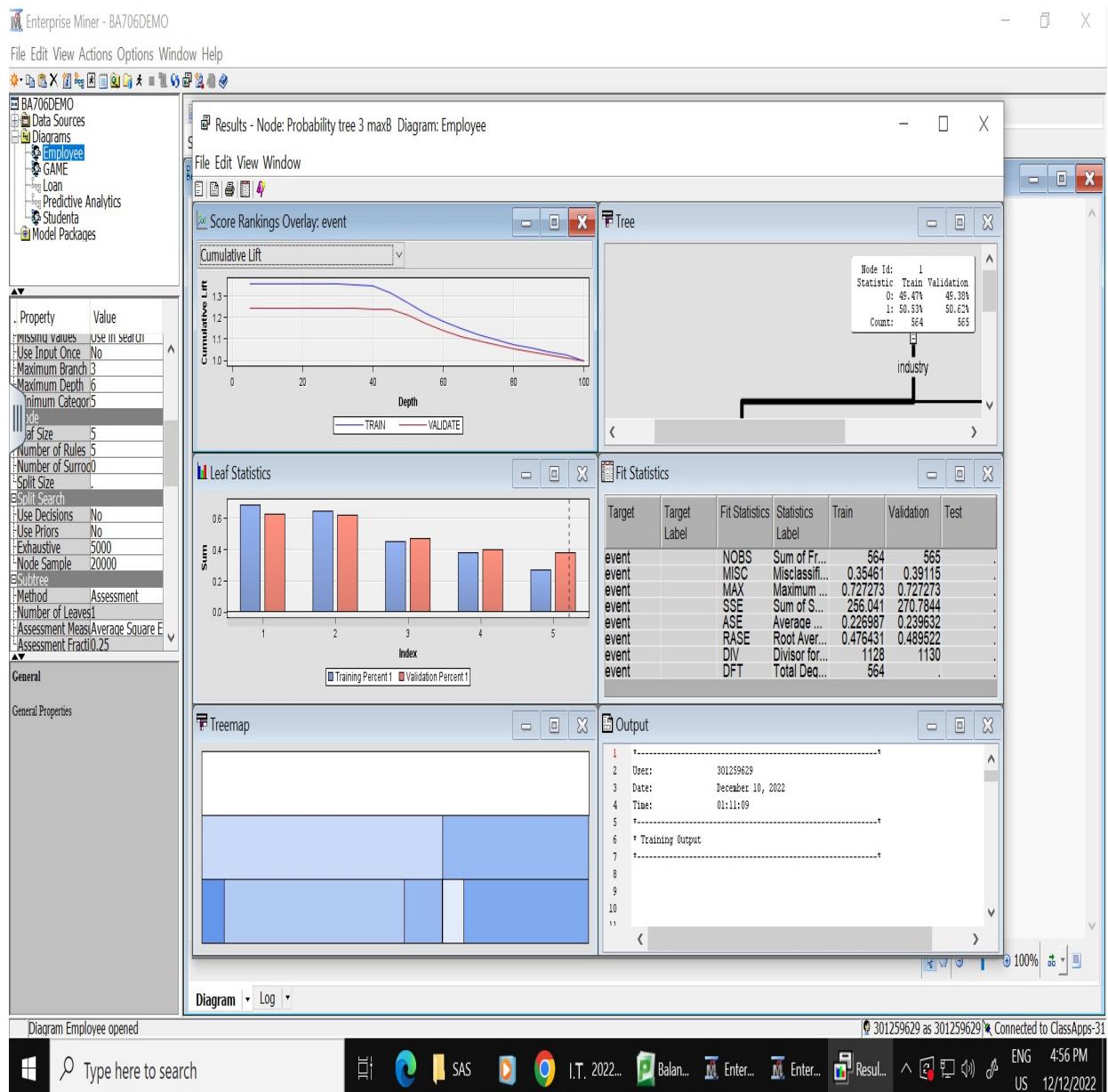
```

49 *
50 *-----+
51 * Score Output
52 *-----+
53 *
54 *
55 *
56 *-----+
57 * Report Output
58 *-----+
59 *
60 *
61 Variable Importance
62 *
63 *
64   Number of
65   Variable   Splitting
66   Name       Rules    Validation
67   Label      Importance Importance
68   Ratio of
69   Validation
70   to Training
71   Importance
72   Importance
73 *
74 Tree Leaf Report
75 *
76   Training
77   Node Depth Observations Percent Validation Validation
78   Id      Observations 1 Observations Percent 1
79   1       2       296  0.37  280  0.40
80   2       2       182  0.69  194  0.63
81   3       2       55   0.65  61   0.62
82   4       3       24   0.58  25   0.56
83   5       3       11   0.00  5    0.00
84   6       3       7    0.00
85   7       3       1
86   8       3       1
87   9       3       1
88   10      3       1
89   11      3       1
90   12      3       1
91   13      3       1
92   14      3       1
93   15      3       1
94   16      3       1
95   17      3       1
96   18      3       1
97   19      3       1
98   20      3       1
99   21      3       1
100  22      3       1
101  23      3       1
102  24      3       1
103  25      3       1
104  26      3       1
105  27      3       1
106  28      3       1
107  29      3       1
108  30      3       1
109  31      3       1
110  32      3       1
111  33      3       1
112  34      3       1
113  35      3       1
114  36      3       1
115  37      3       1
116  38      3       1
117  39      3       1
118  40      3       1
119  41      3       1
120  42      3       1
121  43      3       1
122  44      3       1
123  45      3       1
124  46      3       1
125  47      3       1
126  48      3       1
127  49      3       1
128  50      3       1
129  51      3       1
130  52      3       1
131  53      3       1
132  54      3       1
133  55      3       1
134  56      3       1
135  57      3       1
136  58      3       1
137  59      3       1
138  60      3       1
139  61      3       1
140  62      3       1
141  63      3       1
142  64      3       1
143  65      3       1
144  66      3       1
145  67      3       1
146  68      3       1
147  69      3       1
148  70      3       1
149  71      3       1
150  72      3       1
151  73      3       1
152  74      3       1
153  75      3       1
154  76      3       1
155  77      3       1
156  78      3       1
157  79      3       1
158  80      3       1
159  81      3       1
160  82      3       1
161  83      3       1
162  84      3       1
163  85      3       1
164  86      3       1
165  87      3       1
166  88      3       1
167  89      3       1
168  90      3       1
169  91      3       1
170  92      3       1
171  93      3       1
172  94      3       1
173  95      3       1
174  96      3       1
175  97      3       1
176  98      3       1
177  99      3       1
178  100     3       1
179  101     3       1
180  102     3       1
181  103     3       1
182  104     3       1
183  105     3       1
184  106     3       1
185  107     3       1
186  108     3       1
187  109     3       1
188  110     3       1
189  111     3       1
190  112     3       1
191  113     3       1
192  114     3       1
193  115     3       1
194  116     3       1
195  117     3       1
196  118     3       1
197  119     3       1
198  120     3       1
199  121     3       1
200  122     3       1
201  123     3       1
202  124     3       1
203  125     3       1
204  126     3       1
205  127     3       1
206  128     3       1
207  129     3       1
208  130     3       1
209  131     3       1
210  132     3       1
211  133     3       1
212  134     3       1
213  135     3       1
214  136     3       1
215  137     3       1
216  138     3       1
217  139     3       1
218  140     3       1
219  141     3       1
220  142     3       1
221  143     3       1
222  144     3       1
223  145     3       1
224  146     3       1
225  147     3       1
226  148     3       1
227  149     3       1
228  150     3       1
229  151     3       1
230  152     3       1
231  153     3       1
232  154     3       1
233  155     3       1
234  156     3       1
235  157     3       1
236  158     3       1
237  159     3       1
238  160     3       1
239  161     3       1
240  162     3       1
241  163     3       1
242  164     3       1
243  165     3       1
244  166     3       1
245  167     3       1
246  168     3       1
247  169     3       1
248  170     3       1
249  171     3       1
250  172     3       1
251  173     3       1
252  174     3       1
253  175     3       1
254  176     3       1
255  177     3       1
256  178     3       1
257  179     3       1
258  180     3       1
259  181     3       1
260  182     3       1
261  183     3       1
262  184     3       1
263  185     3       1
264  186     3       1
265  187     3       1
266  188     3       1
267  189     3       1
268  190     3       1
269  191     3       1
270  192     3       1
271  193     3       1
272  194     3       1
273  195     3       1
274  196     3       1
275  197     3       1
276  198     3       1
277  199     3       1
278  200     3       1
279  201     3       1
280  202     3       1
281  203     3       1
282  204     3       1
283  205     3       1
284  206     3       1
285  207     3       1
286  208     3       1
287  209     3       1
288  210     3       1
289  211     3       1
290  212     3       1
291  213     3       1
292  214     3       1
293  215     3       1
294  216     3       1
295  217     3       1
296  218     3       1
297  219     3       1
298  220     3       1
299  221     3       1
300  222     3       1
301  223     3       1
302  224     3       1
303  225     3       1
304  226     3       1
305  227     3       1
306  228     3       1
307  229     3       1
308  230     3       1
309  231     3       1
310  232     3       1
311  233     3       1
312  234     3       1
313  235     3       1
314  236     3       1
315  237     3       1
316  238     3       1
317  239     3       1
318  240     3       1
319  241     3       1
320  242     3       1
321  243     3       1
322  244     3       1
323  245     3       1
324  246     3       1
325  247     3       1
326  248     3       1
327  249     3       1
328  250     3       1
329  251     3       1
330  252     3       1
331  253     3       1
332  254     3       1
333  255     3       1
334  256     3       1
335  257     3       1
336  258     3       1
337  259     3       1
338  260     3       1
339  261     3       1
340  262     3       1
341  263     3       1
342  264     3       1
343  265     3       1
344  266     3       1
345  267     3       1
346  268     3       1
347  269     3       1
348  270     3       1
349  271     3       1
350  272     3       1
351  273     3       1
352  274     3       1
353  275     3       1
354  276     3       1
355  277     3       1
356  278     3       1
357  279     3       1
358  280     3       1
359  281     3       1
360  282     3       1
361  283     3       1
362  284     3       1
363  285     3       1
364  286     3       1
365  287     3       1
366  288     3       1
367  289     3       1
368  290     3       1
369  291     3       1
370  292     3       1
371  293     3       1
372  294     3       1
373  295     3       1
374  296     3       1
375  297     3       1
376  298     3       1
377  299     3       1
378  300     3       1
379  301     3       1
380  302     3       1
381  303     3       1
382  304     3       1
383  305     3       1
384  306     3       1
385  307     3       1
386  308     3       1
387  309     3       1
388  310     3       1
389  311     3       1
390  312     3       1
391  313     3       1
392  314     3       1
393  315     3       1
394  316     3       1
395  317     3       1
396  318     3       1
397  319     3       1
398  320     3       1
399  321     3       1
400  322     3       1
401  323     3       1
402  324     3       1
403  325     3       1
404  326     3       1
405  327     3       1
406  328     3       1
407  329     3       1
408  330     3       1
409  331     3       1
410  332     3       1
411  333     3       1
412  334     3       1
413  335     3       1
414  336     3       1
415  337     3       1
416  338     3       1
417  339     3       1
418  340     3       1
419  341     3       1
420  342     3       1
421  343     3       1
422  344     3       1
423  345     3       1
424  346     3       1
425  347     3       1
426  348     3       1
427  349     3       1
428  350     3       1
429  351     3       1
430  352     3       1
431  353     3       1
432  354     3       1
433  355     3       1
434  356     3       1
435  357     3       1
436  358     3       1
437  359     3       1
438  360     3       1
439  361     3       1
440  362     3       1
441  363     3       1
442  364     3       1
443  365     3       1
444  366     3       1
445  367     3       1
446  368     3       1
447  369     3       1
448  370     3       1
449  371     3       1
450  372     3       1
451  373     3       1
452  374     3       1
453  375     3       1
454  376     3       1
455  377     3       1
456  378     3       1
457  379     3       1
458  380     3       1
459  381     3       1
460  382     3       1
461  383     3       1
462  384     3       1
463  385     3       1
464  386     3       1
465  387     3       1
466  388     3       1
467  389     3       1
468  390     3       1
469  391     3       1
470  392     3       1
471  393     3       1
472  394     3       1
473  395     3       1
474  396     3       1
475  397     3       1
476  398     3       1
477  399     3       1
478  400     3       1
479  401     3       1
480  402     3       1
481  403     3       1
482  404     3       1
483  405     3       1
484  406     3       1
485  407     3       1
486  408     3       1
487  409     3       1
488  410     3       1
489  411     3       1
490  412     3       1
491  413     3       1
492  414     3       1
493  415     3       1
494  416     3       1
495  417     3       1
496  418     3       1
497  419     3       1
498  420     3       1
499  421     3       1
500  422     3       1
501  423     3       1
502  424     3       1
503  425     3       1
504  426     3       1
505  427     3       1
506  428     3       1
507  429     3       1
508  430     3       1
509  431     3       1
510  432     3       1
511  433     3       1
512  434     3       1
513  435     3       1
514  436     3       1
515  437     3       1
516  438     3       1
517  439     3       1
518  440     3       1
519  441     3       1
520  442     3       1
521  443     3       1
522  444     3       1
523  445     3       1
524  446     3       1
525  447     3       1
526  448     3       1
527  449     3       1
528  450     3       1
529  451     3       1
530  452     3       1
531  453     3       1
532  454     3       1
533  455     3       1
534  456     3       1
535  457     3       1
536  458     3       1
537  459     3       1
538  460     3       1
539  461     3       1
540  462     3       1
541  463     3       1
542  464     3       1
543  465     3       1
544  466     3       1
545  467     3       1
546  468     3       1
547  469     3       1
548  470     3       1
549  471     3       1
550  472     3       1
551  473     3       1
552  474     3       1
553  475     3       1
554  476     3       1
555  477     3       1
556  478     3       1
557  479     3       1
558  480     3       1
559  481     3       1
560  482     3       1
561  483     3       1
562  484     3       1
563  485     3       1
564  486     3       1
565  487     3       1
566  488     3       1
567  489     3       1
568  490     3       1
569  491     3       1
570  492     3       1
571  493     3       1
572  494     3       1
573  495     3       1
574  496     3       1
575  497     3       1
576  498     3       1
577  499     3       1
578  500     3       1
579  501     3       1
580  502     3       1
581  503     3       1
582  504     3       1
583  505     3       1
584  506     3       1
585  507     3       1
586  508     3       1
587  509     3       1
588  510     3       1
589  511     3       1
590  512     3       1
591  513     3       1
592  514     3       1
593  515     3       1
594  516     3       1
595  517     3       1
596  518     3       1
597  519     3       1
598  520     3       1
599  521     3       1
600  522     3       1
601  523     3       1
602  524     3       1
603  525     3       1
604  526     3       1
605  527     3       1
606  528     3       1
607  529     3       1
608  530     3       1
609  531     3       1
610  532     3       1
611  533     3       1
612  534     3       1
613  535     3       1
614  536     3       1
615  537     3       1
616  538     3       1
617  539     3       1
618  540     3       1
619  541     3       1
620  542     3       1
621  543     3       1
622  544     3       1
623  545     3       1
624  546     3       1
625  547     3       1
626  548     3       1
627  549     3       1
628  550     3       1
629  551     3       1
630  552     3       1
631  553     3       1
632  554     3       1
633  555     3       1
634  556     3       1
635  557     3       1
636  558     3       1
637  559     3       1
638  560     3       1
639  561     3       1
640  562     3       1
641  563     3       1
642  564     3       1
643  565     3       1
644  566     3       1
645  567     3       1
646  568     3       1
647  569     3       1
648  570     3       1
649  571     3       1
650  572     3       1
651  573     3       1
652  574     3       1
653  575     3       1
654  576     3       1
655  577     3       1
656  578     3       1
657  579     3       1
658  580     3       1
659  581     3       1
660  582     3       1
661  583     3       1
662  584     3       1
663  585     3       1
664  586     3       1
665  587     3       1
666  588     3       1
667  589     3       1
668  590     3       1
669  591     3       1
670  592     3       1
671  593     3       1
672  594     3       1
673  595     3       1
674  596     3       1
675  597     3       1
676  598     3       1
677  599     3       1
678  600     3       1
679  601     3       1
680  602     3       1
681  603     3       1
682  604     3       1
683  605     3       1
684  606     3       1
685  607     3       1
686  608     3       1
687  609     3       1
688  610     3       1
689  611     3       1
690  612     3       1
691  613     3       1
692  614     3       1
693  615     3       1
694  616     3       1
695  617     3       1
696  618     3       1
697  619     3       1
698  620     3       1
699  621     3       1
700  622     3       1
701  623     3       1
702  624     3       1
703  625     3       1
704  626     3       1
705  627     3       1
706  628     3       1
707  629     3       1
708  630     3       1
709  631     3       1
710  632     3       1
711  633     3       1
712  634     3       1
713  635     3       1
714  636     3       1
715  637     3       1
716  638     3       1
717  639     3       1
718  640     3       1
719  641     3       1
720  642     3       1
721  643     3       1
722  644     3       1
723  645     3       1
724  646     3       1
725  647     3       1
726  648     3       1
727  649     3       1
728  650     3       1
729  651     3       1
730  652     3       1
731  653     3       1
732  654     3       1
733  655     3       1
734  656     3       1
735  657     3       1
736  658     3       1
737  659     3       1
738  660     3       1
739  661     3       1
740  662     3       1
741  663     3       1
742  664     3       1
743  665     3       1
744  666     3       1
745  667     3       1
746  668     3       1
747  669     3       1
748  670     3       1
749  671     3       1
7
```

The Validation Average Square Error for the two-branch classification Tree is 0.237022. This has a similar validation ASE as the two-branch classification tree model.

### **THREE BRANCH PROBABILITY TREE**

The decision tree node was attached to the data partition node with a maximum branch of three, ‘assessment’ used as the method and ‘average squared error’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.



The screenshot shows the Enterprise Miner software interface with the title "Enterprise Miner - BA706DEMO". The menu bar includes File, Edit, View, Actions, Options, Window, Help. The left sidebar displays the project structure under "BA706DEMO" with nodes like Data Sources, Diagrams, and Predictive Analytics. A property grid on the left shows settings such as "Number of Ruled 5", "Number of SortOrder 0", and "Assessment Fraction 0.25". The main window displays a "Tree" diagram titled "Results - Node: Probability tree 3 maxB: Diagram: Employee". The tree structure is as follows:

```

graph TD
    Node1[Node Id: 1  
Statistic: Train Validation  
0: 45.21% 45.14%  
1: 54.78% 54.85%  
Count: 564 565] -- industry --> Node2[Node Id: 2  
Statistic: Train Validation  
0: 38.40% 56.01%  
1: 41.59% 43.98%  
Count: 351 341]
    Node1 -- industry --> Node3[Node Id: 3  
Statistic: Train Validation  
0: 34.74% 35.29%  
1: 65.25% 64.71%  
Count: 213 224]
    Node2 -- profession --> Node4[Node Id: 4  
Statistic: Train Validation  
0: 72.73% 62.07%  
1: 27.27% 37.92%  
Count: 33 29]
    Node2 -- profession --> Node5[Node Id: 5  
Statistic: Train Validation  
0: 61.03% 55.76%  
1: 38.96% 44.24%  
Count: 263 251]
    Node3 -- way --> Node6[Node Id: 6  
Statistic: Train Validation  
0: 34.55% 37.70%  
1: 65.45% 62.29%  
Count: 55 61]
    Node3 -- way --> Node7[Node Id: 7  
Statistic: Train Validation  
0: 54.84% 53.33%  
1: 45.15% 46.67%  
Count: 31 30]
    Node4 -- IT --> Node8[Node Id: 8  
Statistic: Train Validation  
0: 31.32% 37.11%  
1: 68.68% 62.89%  
Count: 182 194]
    Node5 -- HR, ETC, BUSINESSDEVELO... --> Node8
    Node6 -- COMMERCIAL, SALES, FINA... --> Node8
    Node7 -- FOOT --> Node8
    Node8 -- BUS, CAR Or Missing --> Node9[Node Id: 9  
Statistic: Train Validation  
0: 31.32% 37.11%  
1: 68.68% 62.89%  
Count: 182 194]
  
```

The screenshot shows the Enterprise Miner software interface for the BA706DEMO project. The main window displays the results of a "Probability tree 3 maxB" model on the "Employee" diagram. The results table includes columns for Target, Target Label, Fit Statistics, Statistics Label, Train, Validation, and Test. The properties panel on the left contains sections for General, Model Properties, and various algorithm-specific parameters like Tree Size, Split Search, and Subtree.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	NOBS	Sum of Frequencies	564	565		
event	MISC	Misclassification Rate	0.35461	0.39115		
event	MAX	Maximum Average Error	0.72793	0.77673		
event	SSE	Sum of Squared Errors	256.041	270.7844		
event	ASE	Average Squared Error	0.226987	0.239632		
event	RASE	Root Average Squared...	0.476431	0.489522		
event	DIV	Divisor for ASE	1128	1130		
	DFT	Total Degrees of Free...	564			

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Sources
- Diagrams
- Employee**
- GAME
- ML
- Predictive Analytics
- Student
- Model Packages

Property Value

- maxnum values Use in Start UI
- Use Input Once No
- Maximum Branch 3
- Maximum Depth 6
- maximum Categoricals 5
- leaf Size 5
- Number of Rules 5
- Number of Surrogate 0
- Split Size
- Split Search
- Use Decisions No
- Use Stag No
- Exhaustive 5000
- Node Sample 20000
- Method Assessment
- Number of Leaves 1
- Assessment MeasAverage Square Error
- Assessment Fract0.25

General

General Properties

Results - Node: Probability tree 3 maxB: Diagram: Employee

File Edit View Window

Output

```

47 INTO I_event INTO: event
48
49
50 *-----*
51 * Score Output
52 *-----*
53
54
55 *-----*
56 * Report Output
57 *-----*
58
59
60
61 Variable Importance
62
63
64 Number of
65 Variable Splitting
66 Name Label Rules Importance Validation
67 Ratio of
68 industry 1 1.0000 1.0000 1.0000
69 profession 1 0.7410 0.8436 1.1384
70 way 1 0.4443 0.3951 0.8092
71
72
73 Tree Leaf Report
74
75
76 Training Validation Validation
77 Node Percent Observations Percent 1 Observations Percent 1
78 Id Depth Observations
79
80 5 2 263 0.38 231 0.40
81 6 2 182 0.69 194 0.63
82 6 2 55 0.65 61 0.62
83 4 2 33 0.27 29 0.38
84 7 2 31 0.45 30 0.47
85
86
87
88

```

Diagram Log 301259629 as 301259629 Connected to ClassApps-31

Type here to search

The tree has five leaves with three variable splits. This model has the least amount of significant variables when compared to all the other tree models. The stag variable which shows the amount of time an employee has worked was not considered significant enough to be included. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their jobs. 62.3% of Commercial, Sales, and Finance professionals in the power generation industry are likely to quit.

The Validation Average Square Error for the three-branch classification Tree is 0.239632. The Two Branch Probability Tree gives a higher validation assessment than this tree.

## EXPLORING OUR DATA

After we were done with analyzing our decision trees we decided to do a state explore to see if we have any missing values in the dataset. As decision trees are forgiving with respect to data, they are not affected by missing values, this step was not necessary before the tree models. However, for the next set of models, regressions and neural networks, the models are unable to run if there are values missing.

Data Role	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	event 0	stad	30.225...	0	279	0.4928...	155.36...	38.761...	35.248...	1.3232...	1.3131...	INPUT	stad	0.0596...	0.0584...	1
TRAIN	event 0	selfcon...	23.556...	0	285	0.3942...	160.05...	34.54...	30.173...	1.5672...	2.4039...	INPUT	selfcon...	0.0541...	0.0584...	2
TRAIN	event 0	selfcon...	5.7	0	279	1	10.57878...	1.9780...	0.0035...	-0.72717...	0.0295...	INPUT	selfcon...	0.0302...	0.0295...	1
TRAIN	event 1	selfcon...	5.7	0	285	1	10.545193...	1.90375...	0.1322...	-0.32953...	0.0295...	INPUT	selfcon...	-0.02958...	0.0295...	2
TRAIN	event 0	independ...	5.5	0	279	1	9.5232053...	1.7014...	-0.0266...	-0.1605...	0.0227...	INPUT	independ...	0.02327...	0.0227...	1
TRAIN	event 1	independ...	5.5	0	265	1	10.5276...	1.7014...	-0.0385...	-0.1605...	0.0227...	INPUT	independ...	0.02327...	0.0227...	2
TRAIN	event 0	anxiety	5.6	0	279	1.7	10.579319...	1.7066...	0.0877...	-0.38356...	0.0227...	INPUT	anxiety	0.0168...	0.0164...	1
TRAIN	event 1	anxiety	5.6	0	285	1.7	9.456035...	1.7591...	0.2170...	-0.71277...	0.0164...	INPUT	anxiety	-0.01647...	0.0164...	2
TRAIN	event 0	novator	5.8	0	279	1	10.59414...	1.8340...	-0.24588...	-0.29084...	0.0153...	INPUT	novator	0.0153...	0.0153...	1
TRAIN	event 1	novator	5.8	0	285	1	10.59414...	1.8340...	-0.24588...	-0.29084...	0.0153...	INPUT	novator	0.0153...	0.0153...	2
TRAIN	event 0	extrave...	5.4	0	279	1	10.55720...	1.9076...	-0.07998...	-0.17857...	0.00464...	INPUT	extrave...	-0.00474...	0.00464...	1
TRAIN	event 1	extrave...	5.4	0	285	1	10.56245...	1.8409...	0.0590...	-0.52159...	0.00464...	INPUT	extrave...	0.00464...	0.00464...	2
TRAIN	event 0	ace	30	0	279	19	54.31230...	6.4133...	0.6127...	-0.00539...	0.00301...	INPUT	ace	0.0030...	0.00301...	1
TRAIN	event 1	ace	30	0	285	18	54.31041...	7.0952...	0.48201...	-0.41283...	0.00301...	INPUT	ace	-0.00301...	0.00301...	2

## CAP AND FLOOR

We noticed there were no missing values, hence no impute node required in the diagram. However, there was a skewness in the dataset so we resolved it with a cap and floor by connecting a replacement node. The result shows that the variables ‘age’ and ‘stag’ were affected after running the cap and floor. Age was capped at 51.42497 and floored at 10.84524 while stag was capped and floored at 137.8856 and 64.7267 respectively.

**Total Replacement Counts**

Variable	Label	Role	Train	Validation
age	age	INPUT	3	4
anxiety	anxiety	INPUT	0	0
extraversion	extraversion	INPUT	0	0
independ	independ	INPUT	0	0
novator	novator	INPUT	0	0
selfcontrol	selfcontrol	INPUT	0	0
stag	stag	INPUT	12	11

**Interval Variables**

Variable	Replace Variable	Limits Method	Lower limit	Upper Limit	Label	Replacement Method	Lower Replacement Value	Upper Replacement Value
age	REP age	STDDEV	10.84524	51.42497 age		COMPUTED	10.84524	51.42497
anxiety	REP anxiety	STDDEV	0.494046	10.9063 anxiety		COMPUTED	0.494046	10.9063
extraversion	REP extraversion	STDDEV	-0.0689	11.9185 extraversion		COMPUTED	-0.0689	11.9185
independ	REP independ	STDDEV	0.322176	10.5069 independ		COMPUTED	0.322176	10.5069
novator	REP novator	STDDEV	0.242424	11.46041 novator		COMPUTED	0.242424	11.46041
selfcontrol	REP selfcontrol	STDDEV	-0.2212	11.4573 selfcontrol		COMPUTED	-0.2212	11.4573
stag	REP stag	STDDEV	-64.7267	137.8856 stag		COMPUTED	-64.7267	137.8856

A stat explore node was connected to the replacement node to see the effect of the cap and floor on the skewed values.

Mustai | Mustai | Mustai | Mustai | PROJE | Empl... | Netflix | Watch | Apps | VMwa | +

pod1.centennialcollege.ca/portal/webclient/#/desktop

Results - Node: StatExplore (2) Diagram: Employee

File Edit View Window

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	event	0	REP stag	30.22587	0	279	0.492813	137.8856	38.53872	34.57223	1.240112	0.943682	INPUT	Replacem...	0.059168	0.057922	1
TRAIN	event	1	REP stag	23.5647	0	285	0.394251	137.8856	34.27829	31.59778	1.474722	1.907757	INPUT	Replacem...	-0.05792	0.057922	2
TRAIN	event	0	REP selfcontrol	5.7	0	279	1	10	5.787814	1.978098	0.003512	-0.72717	INPUT	Replacem...	0.030211	0.029575	1
TRAIN	event	1	REP selfcontrol	5.7	0	285	1	10	5.45193	1.90375	0.132246	-0.32953	INPUT	Replacem...	-0.02958	0.029575	2
TRAIN	event	0	REP independ	5.5	0	279	1	9.8	5.28853	1.701421	-0.01926	-0.52105	INPUT	Replacem...	-0.02327	0.022782	1
TRAIN	event	1	REP independ	5.5	0	285	1	10	5.537895	1.887434	0.038587	-0.16023	INPUT	Replacem...	0.022782	0.022782	2
TRAIN	event	0	REP anxiety	5.6	0	279	1.7	10	5.79319	1.706628	0.087777	-0.38356	INPUT	Replacem...	0.016824	0.016469	1
TRAIN	event	1	REP anxiety	5.6	0	285	1.7	9.4	5.603509	1.759158	0.217084	-0.71277	INPUT	Replacem...	-0.01647	0.016469	2
TRAIN	event	0	REP novator	6	0	279	1	10	5.759498	1.904304	-0.23808	-0.51157	INPUT	Replacem...	-0.01571	0.015378	1
TRAIN	event	1	REP novator	6	0	285	1	10	5.941404	1.834001	-0.24588	-0.29084	INPUT	Replacem...	0.015378	0.015378	2
TRAIN	event	0	REP extraversion	5.4	0	279	1	10	5.572043	1.907635	-0.07998	-0.17537	INPUT	Replacem...	-0.00474	0.00464	1
TRAIN	event	1	REP extraversion	5.4	0	285	1	10	5.624561	1.840954	0.059095	-0.52159	INPUT	Replacem...	0.00464	0.00464	2
TRAIN	event	0	REP ade	30	0	279	19	51.42497	31.2216	6.382259	0.576385	-0.16957	INPUT	Replacem...	0.003105	0.003039	1
TRAIN	event	1	REP ade	30	0	285	18	51.42497	31.03035	7.065542	0.449732	-0.54091	INPUT	Replacem...	-0.00304	0.003039	2

Type here to search

SAS I.T. 2022... Balan... Enter... Resul... Resul... ENG 12:33 AM US 17/12/2022

12:33 AM 12/17/2022

After exploring cap and floor, we realized that the data was still skewed on the Rep stag( which was imputed by the cap and floor to fix the initial skewness of the data). So we used a log transform to try and fix the skewness.

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Data Sources
- Diagrams
- Employee**
- Log
- Predictive Analytics
- Model Packages

Variables - Trans

Columns:  Label  Mining  Basic  Statistics

Name	Method	Number of Bins	Role	Level
REP age	Default	4	Input	Interval
REP anxiety	Default	4	Input	Interval
REP extraversion	Default	4	Input	Interval
REP indepen	Default	4	Input	Interval
REP novator	Default	4	Input	Interval
REP selfcontrol	Default	4	Input	Interval
REP staa	Log	4	Input	Interval
age	Default	4	Rejected	Interval
anxiety	Default	4	Rejected	Interval
coach	Default	4	Rejected	Nominal
event	Default	4	Target	Binary
extraversion	Default	4	Rejected	Interval
independ	Default	4	Input	Nominal
novatoe	Default	4	Rejected	Interval
prevwage	Default	4	Input	Nominal
heavy gender	Default	4	Input	Nominal
independ	Default	4	Rejected	Interval
industry	Default	4	Input	Nominal
independ	Default	4	Rejected	Interval
profession	Default	4	Input	Nominal
selfcontrol	Default	4	Rejected	Interval
staa	Default	4	Rejected	Interval
stop	Default	4	Rejected	Interval
traffic	Default	4	Rejected	Nominal
way	Default	4	Input	Nominal

Apply Reset

Explore... Update Path OK Cancel

Diagram Employee opened

301259629 as 301259629 Connected to ClassApps-31

Type here to search

Windows Taskbar: SAS, Chrome, I.T. 2022-1..., Balance..., Enterpr..., Enterpr..., ENG US 12/12/2022

Next was to stat explore the data and we noticed the skewness was sorted as shown below:

pod1.centennialcollege.ca/portal/webclient/#/desktop

Results - Node: StatExplore (3) Diagram: Employee

File Edit View Window

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	event	0	REP selfcontrol	5.7	0	279	1	10	5.767814	1.978098	0.003512	-0.72717	INPUT	Replace...	0.030211	0.028575	1
TRAIN	event	1	REP selfcontrol	5.7	0	285	1	10	6.45193	1.90375	0.132246	-0.32953	INPUT	Replace...	-0.02958	0.028575	2
TRAIN	event	0	REP independ	5.5	0	279	1	9.8	5.28853	1.701421	-0.01926	-0.52105	INPUT	Replace...	-0.02327	0.022782	1
TRAIN	event	1	REP independ	5.5	0	285	1	10	5.53785	1.687434	0.038587	-0.16023	INPUT	Replace...	0.022782	0.022782	2
TRAIN	event	0	REP novatoe	5.5	0	279	1.7	10	5.53785	1.687434	0.038587	-0.16023	INPUT	Replace...	0.022782	0.022782	1
TRAIN	event	1	REP anxiety	5.6	0	285	1.7	9.4	5.603509	1.759158	0.217084	-0.71277	INPUT	Replace...	-0.016469	0.016469	2
TRAIN	event	0	REP novatoe	6	0	279	1	10	5.759498	1.904304	-0.23808	-0.51157	INPUT	Replace...	-0.01571	0.015378	1
TRAIN	event	1	REP novatoe	6	0	285	1	10	5.759498	1.904304	-0.23808	-0.51157	INPUT	Replace...	0.015378	0.015378	2
TRAIN	event	0	LOG REP staa	3.4001247	0	279	0.400662	4.93365	3.220393	1.068795	-0.24959	-0.47026	INPUT	Transform...	0.012582	0.012582	1
TRAIN	event	1	LOG REP staa	3.200975	0	285	0.332357	4.93365	3.140051	0.997076	-0.41894	-0.24068	INPUT	Transform...	-0.01258	0.012582	2
TRAIN	event	0	REP extraversion	5.4	0	279	1	10	5.572043	1.907635	-0.07998	-0.17537	INPUT	Replace...	-0.00474	0.00464	1
TRAIN	event	1	REP extraversion	5.4	0	285	1	10	5.572043	1.907635	-0.07998	-0.17537	INPUT	Replace...	0.00464	0.00464	2
TRAIN	event	0	REP ace	30	0	279	19	51.42497	31.2216	6.382259	0.876385	0.8957	INPUT	Replace...	0.003105	0.003039	1
TRAIN	event	1	REP ace	30	0	285	18	51.42497	31.03035	7.065542	0.449732	-0.54091	INPUT	Replace...	-0.00304	0.003039	2

Type here to search

Windows Taskbar: SAS, Chrome, I.T. 2022..., Balance..., Enterpr..., Enterpr..., ENG US 12/12/2022

12:35 AM 17/12/2022

## REPLACEMENT

In logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and increases the stability and significance of the coefficients.

Therefore we made the Industry variables **HoReCa** and **etc** ‘unknown’ because we do not know what they are.

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
head_gender	m		299C		m	.
head_gender	f		265C		f	.
head_gender	_UNKNOWN_	DEFAULT_	C			.
industry	Retail		143C		Retail	.
industry	manufacture		76C		manufacture	.
industry	IT		63C		IT	.
industry	Banks		57C		Banks	.
industry	etc	Unknown	49C		etc	.
industry	Consult		37C		Consult	.
industry	State		23C		State	.
industry	PowerGeneration		19C		PowerGeneration	.
industry	transport		19C		transport	.
industry	Building		17C		Building	.
industry	Telecom		16C		Telecom	.
industry	Mining		14C		Mining	.
industry	Pharma		10C		Pharma	.
industry	RealEstate		8C		RealEstate	.
industry	HoReCa	Unknown	7C		HoReCa	.
industry	Agriculture		6C		Agriculture	.
industry	_UNKNOWN_	DEFAULT_	C			.
profession	HR		377C		HR	.

And we combined all the industry variables except for HR, IT and Sales and named it 'Other'. We also combined the Profession variables BUS and Car and named it 'Motor'.

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
industry	UNKNOWN_	DEFAULT_		C		
profession	HR		377C	HR		
profession	IT		40C	IT		
profession	Sales		31C	Sales		
profession	etc	Other	22C	etc		
profession	Commercial	Other	14C	Commercial		
profession	Marketing	Other	14C	Marketing		
profession	Consult	Other	13C	Consult		
profession	BusinessDevelopment	Other	12C	BusinessDevelopment		
profession	Finance	Other	8C	Finance		
profession	Teaching	Other	8C	Teaching		
profession	manage	Other	8C	manage		
profession	Engineer	Other	7C	Engineer		
profession	Accounting	Other	4C	Accounting		
profession	Law	Other	4C	Law		
profession	PR	Other	2C	PR		
profession	UNKNOWN_	DEFAULT_	C			
way	bus	Motor	341C	bus		
way	car	Motor	157C	car		
way	foot		66C	foot		
way	UNKNOWN_	DEFAULT_	C			

**Results - Node: Replacement**

Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value	Label
industry	etc	C	etc	.	Unknown	
industry	HoReCa	C	HoReCa	.	Unknown	
profession	etc	C	etc	.	Other	
profession	Commercial	C	Commercial	.	Other	
profession	Marketing	C	Marketing	.	Other	
profession	Consult	C	Consult	.	Other	
profession	BusinessDevelopment	C	BusinessDevelopment	.	Other	
profession	Finance	C	Finance	.	Other	
profession	Teaching	C	Teaching	.	Other	
profession	Management	C	Management	.	Other	
profession	Engineer	C	Engineer	.	Other	
profession	Accounting	C	Accounting	.	Other	
profession	Law	C	Law	.	Other	
profession	IT	C	IT	.	Other	
way	bus	C	bus	.	Motor	
way	car	C	car	.	Motor	

**Total Replacement Counts**

Variable	Role	Label	Train	Validation
industry	INPUT	56	49	49
profession	INPUT	116	116	116
way	INPUT	498	514	514

Dummy variables, which are useful in regression analysis, use the values 0 or 1 to represent the absence or existence of any categorical effect that would be anticipated to shift the outcome. So, now we can begin building our regression models.

# REGRESSION

As stated earlier, there were no missing values so we did not need to connect the impute node.

Therefore, we are going into regression. We first created :

## FULL REGRESSION

And in other to optimize complexity, we also created:

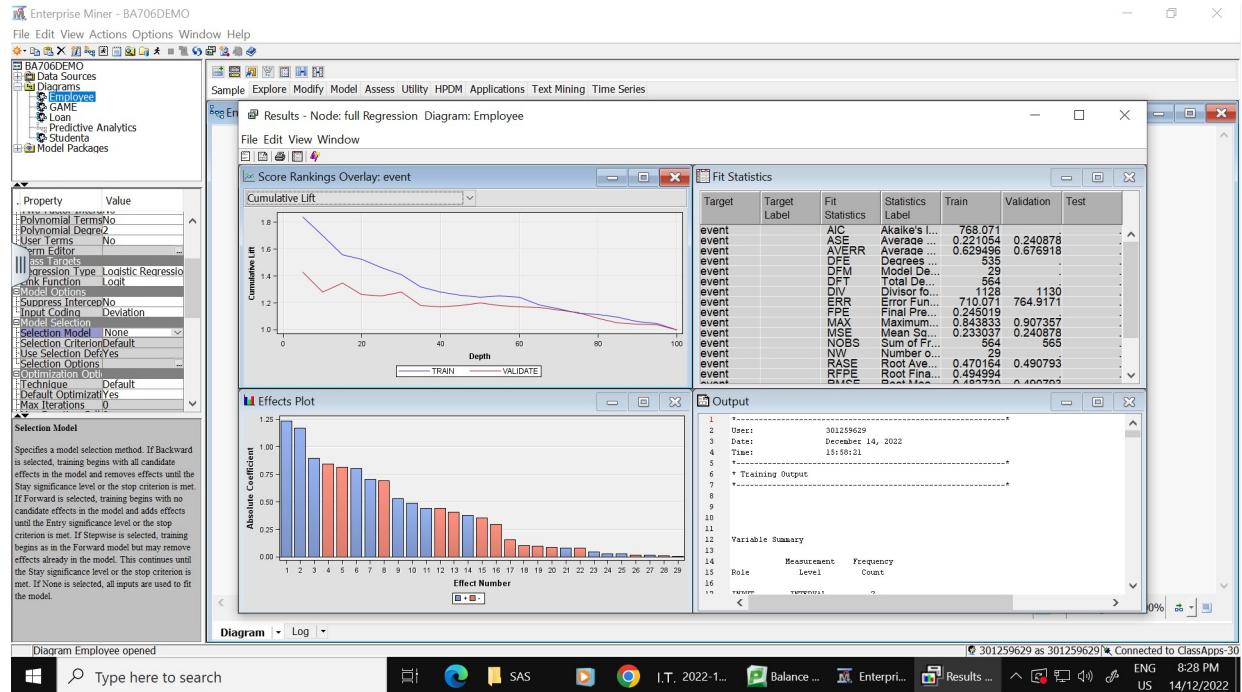
## FORWARD REGRESSION

## STEPWISE REGRESSION

## BACKWARD REGRESSION

## FULL REGRESSION

The full regression used 28 variables to run the model. The average squared error is 0.240878



Two screenshots of the Enterprise Miner software interface showing regression results for the "Employee" dataset.

**Screenshot 1 (Top): Fit Statistics**

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		AIC	Akaike's Information Cr...	768.071		
event		ASE	Average Square Error	0.221054	0.240878	
event		AVERR	Average Error Function	0.625936	0.576918	
event		DFE	Degrees of Freedom f...	535		
event		DFM	Model Degrees of Free...	29		
event		DIT	Total Degrees of Free...	564		
event		DIV	Divisor for ASE	1128	1130	
event		ERR	Error Function	710.071	764.9171	
event		FINAL	Final Prediction Error	0.470119		
event		MAX	Maximum Absolute Error	0.643833	0.907357	
event		MSE	Mean Square Error	0.233037	0.240878	
event		NBGS	Sum of Frequencies	564	565	
event		NW	Number of Wrong W...	29		
event		RASE	Root Average Sum of...	0.470164	0.490793	
event		RFPE	Root Final Prediction...	0.494994		
event		RMSE	Root Mean Squared E...	0.482739	0.490793	
event		SBC	Sum of Case Weights Cr...	893.76		
event		SSE	Sum of Squared Errors	249.3464	272.1922	
event		SUMW	Sum of Case Weights ...	1128	1130	
event		MISC	Misclassification Rate	0.379433	0.39469	

**Screenshot 2 (Bottom): Odds Ratio Estimates**

Effect	Point Estimate	
LOG_PEP_stag	0.907	
REP_apr	1.003	
REP_industry	0.926	
REP_extraversion	0.919	
REP_independ	1.012	
REP_industry	Agriculture vs transport	6.303
REP_industry	Banks vs transport	4.113
REP_industry	Building vs transport	2.446
REP_industry	Consult vs transport	4.479
REP_industry	IT vs transport	1.169
REP_industry	Mining vs transport	3.859
REP_industry	Pharm vs transport	3.004
REP_industry	Food/retail vs transport	1.327
REP_industry	Realestate vs transport	1.136
REP_industry	Retail vs transport	1.857
REP_industry	Steel vs transport	2.721
REP_industry	Telecom vs transport	2.370
REP_industry	Unknown vs transport	2.712
REP_industry	Manufacture vs transport	1.704
REP_profession	HR vs Sales	1.045
REP_profession	IT vs Sales	0.749
REP_profession	Other vs Sales	0.671
REP_profession	Other vs Sales	2.039
REP_selfcontrol	0.905	
REP_way	Motor vs foot	2.457
REP_way	f vs m	1.171
greywhite	grey vs white	0.972
head_gender	f vs m	0.985

## Interpretation of Full Regression

Variables	point Estimate	When compared with the Transport industry
-----------	----------------	---

REP_industry	Building vs transport	8.546	Employees in the Building industry are 8.5 times more likely to quit.
REP_industry	Agric vs transport	6.503	Employees in the Agriculture Industry are 6.5 times more likely to quit.
REP_industry	pharma vs transport	5.904	Employees in the Pharmacy industry are 5.9 times more likely to quit.
REP_industry	Consult vs transport	4.478	Employees in the Consult industry are 4.5 times more likely to quit.
REP_industry	Banks vs transport	4.113	Employees in the Bank industry are 4.1 times more likely to quit.
REP_industry	Mining vs transport	3.858	Employees in the Mining industry are 3.9 times more likely to quit.
REP_industry	State vs transport	2.721	Employees in the State industry are 2.7 times more likely to quit.
REP_industry	Unknown vs transport	2.712	Employees in the Unknown industry are 2.7 times more likely to quit.
REP_industry	Telecom vs transport	2.270	Employees in the Telecom industry are 2.3 times more likely to quit.
REP_industry	Retail vs	1.857	Employees in the Retail industry are 85.7% more likely to

	transport		quit.
REP_industry	Manufacture vs transport	1.704	Employees in the Manufacture industry are 70.4% more likely to quit.
REP_industry	PowerGen vs transport	1.327	Employees in the PowerGen industry are 32.7% more likely to quit.
REP_industry	IT vs transport	1.169	Employees in the IT industry are 16.9% more likely to quit.
REP_industry	RealEstate vs transport	1.136	Employees in the Real Estate industry are 13.6% more likely to quit.
			<b>When compared with Sales Profession</b>
REP_profession	Other vs Sales	2.039	Employees in other professions besides HR and IT are more likely to quit
REP_profession	HR vs Sales	0.749	Employees in HR are 25.1% less likely to quit.
REP_profession	IT vs Sales	0.671	Employees in IT are 32.9% less likely to quit.
			<b>Other Variables</b>

REP_way	Motor vs foot	2.667	Employees who come by Motor are 2.7 times more likely quit than those who come by foot
gender	f vs m	1.171	Female employees are 17.1% more likely to quit than male employees
REP_novator		1.045	For each 1 score the odds of quitting changed by a factor 1.045, a 4.5% increase
ependen		1.012	For each 1 score the odds of quitting changed by a factor of 1.012, a 1.2% increase
REP_age		1.003	For every year added in age, employees are 0.3% more quit
head_gender	f vs m	0.985	Female supervisor is 2% less likely to quit than a male supervisor
greywage	grey vs white	0.972	Employees who earn above minimum wage are 2.8% less likely to quit than those who earn minimum wage
REP_anxiety		0.926	For each 1 score the odds of quitting changed by a factor of 0.926, a 7.4% decrease
raversion		0.918	For each 1 score the odds of quitting changed by a factor of 0.918, a 8.2% decrease
LOG_REP_stag		0.907	For each 1 score the odds of quitting changed by a factor of 0.907, a 9.3% decrease

REP_selfcontrol	0.905	For each 1 score the odds of quitting changed by a factor of 0.905, a 9.5% decrease
-----------------	-------	---

## FORWARD REGRESSION

The forward regression used 18 variables to build the model and the outcome is an ASE of 0.237256 which is a better result than the full regression.

The gender of the employee or supervisor, self control score , anxiety score, all other scores, wage level, and experience were not considered significant in building this model.

Apps | Center X VMware Horizon X Lesson 8 - 22F X Mustafa Ahme X ChatGPT X Employee - io X Untitled - Cola X + v - X

pod1.centennialcollege.ca/portal/webclient/#/desktop

Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Data Sources
- Diagrams
- Employee**
- GAME
- Loan
- Predictive Analytics
- Student
- Model Packages

. Property Value

- :Polynomial TermsNo
- :Polynomial Degre2
- User Terms No
- Term Editor
- Regression Type Logistic Regression
- Link Function Logit
- Model Options
- :Suppress InterceptNo
- :Input Coding Deviation
- Model Selection
- :Selection Model Forward
- :Selection Criterion Validation Error
- :Use Selection DefYes
- :Selection Options
- Optimization Opt
- :Technique Default
- :Default OptimizatYes
- Max Iterations 0

General Properties

Diagram Employee opened

Score Rankings Overlay: event

Cumulative Lift

Depth	Cumulative Lift (TRAIN)	Cumulative Lift (VALIDATE)
0	1.80	1.60
10	1.55	1.45
20	1.45	1.35
30	1.40	1.30
40	1.35	1.25
50	1.30	1.20
60	1.25	1.15
70	1.20	1.10
80	1.15	1.05
90	1.10	1.00
100	1.05	0.95

Fit Statistics

Target	Target Label	Fit Statistics	Statistics	Train	Validation	Test
event		AIC	Akaike's I...	756.7345		
event		ASE	Average ...	0.224052	0.237256	
event		AVERR	Average ...	0.637176	0.667418	
event		DFE	Degrees ...	545		
event		DFM	Model De...	19		
event		DFT	Total De...	564		
event		DIV	Divisor fo...	1128	1130	
event		ERR	Error Fun...	718.7345	754.1829	
event		FPE	Final Pre...	0.239674		
event		MAX	Maximum...	0.861805	0.877769	
event		MSE	Mean Sq...	0.231863	0.237256	
event		NOBS	Sum of Fr...	564	565	
event		NW	Number o...	19		
event		RASE	Root Ave...	0.473342	0.48709	

Effects Plot

Effect Number	Absolute Coefficient (Blue)	Absolute Coefficient (Red)
1	1.00	
2	0.90	
3		0.80
4		0.70
5		0.65
6	0.60	
7	0.55	
8	0.50	
9	0.45	
10		0.40
11		0.35
12	0.30	
13		0.30
14	0.25	
15		0.20
16	0.15	
17		0.10
18		0.05
19		0.05

Output

```

1 -----
2 User: 301239629
3 Date: December 14, 2022
4 Time: 15:59:23
5 -----
6 * Training Output
7 -----
8
9
10
11
12 Variable Summary
13
14 Role Measurement Frequency Count
15 Level

```

Type here to search

SAS I.T. 2022-1... Balance... Enterprise... Results... ENG 8:36 PM US 14/12/2022

Lesson 8.R Data.csv Show all X

8:36 PM 12/14/2022

Results - Node: forward reg. Diagram: Employee

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event			Akaike's Information Cr...	756.7345		
event			Average Square Error	0.224602	0.237256	
event			Average Function	0.637176	0.657418	
event			Degrees of Freedom f...	545		
event			Model Degrees of Free...	19		
event			DIT	100.000000		
event			DIV	564		
event			Error Function	1128	1130	
event			Error Function	718.7345	754.1829	
event			Final Function Error	0.239671		
event			MAX	0.661005	0.877769	
event			MSE	0.231863	0.237256	
event			Mean Square Error	564	565	
event			NBDS			
event			RASE	Root Average Sum of ...	0.473342	0.48709
event			RFPE	Root Final Prediction ...	0.489565	
event			RMSE	Root Mean Squared Erro...	0.481522	0.48709
event			SBC	Sum of Case Weights Cr...	630.005	
event			SSE	Sum of Squared Errors	250.031	268.0996
event			SUMW	Sum of Case Weights ...	1128	1130
event			MISC	Misclassification Rate	0.37234	0.410619

Results - Node: forward reg. Diagram: Employee

Effect	Point Estimate
REP_industry Agriculture vs transport	6.769
REP_industry Banks vs transport	4.600
REP_industry Building vs transport	7.635
REP_industry Consult vs transport	4.408
REP_industry IT vs transport	1.198
REP_industry Manufacturing vs transport	2.437
REP_industry Pharma vs transport	5.934
REP_industry PowerGeneration vs transport	1.292
REP_industry RealEstate vs transport	1.291
REP_industry Retail vs transport	1.250
REP_industry State vs transport	3.063
REP_industry Teleco vs transport	2.356
REP_industry Unknown vs transport	2.806
REP_industry manufacture vs transport	1.628
REP_profession Admin vs Sales	0.115
REP_profession IT vs Sales	0.563
REP_profession Other vs Sales	1.762
REP_way Motor vs foot	2.531

## Interpretation of Forward Regression

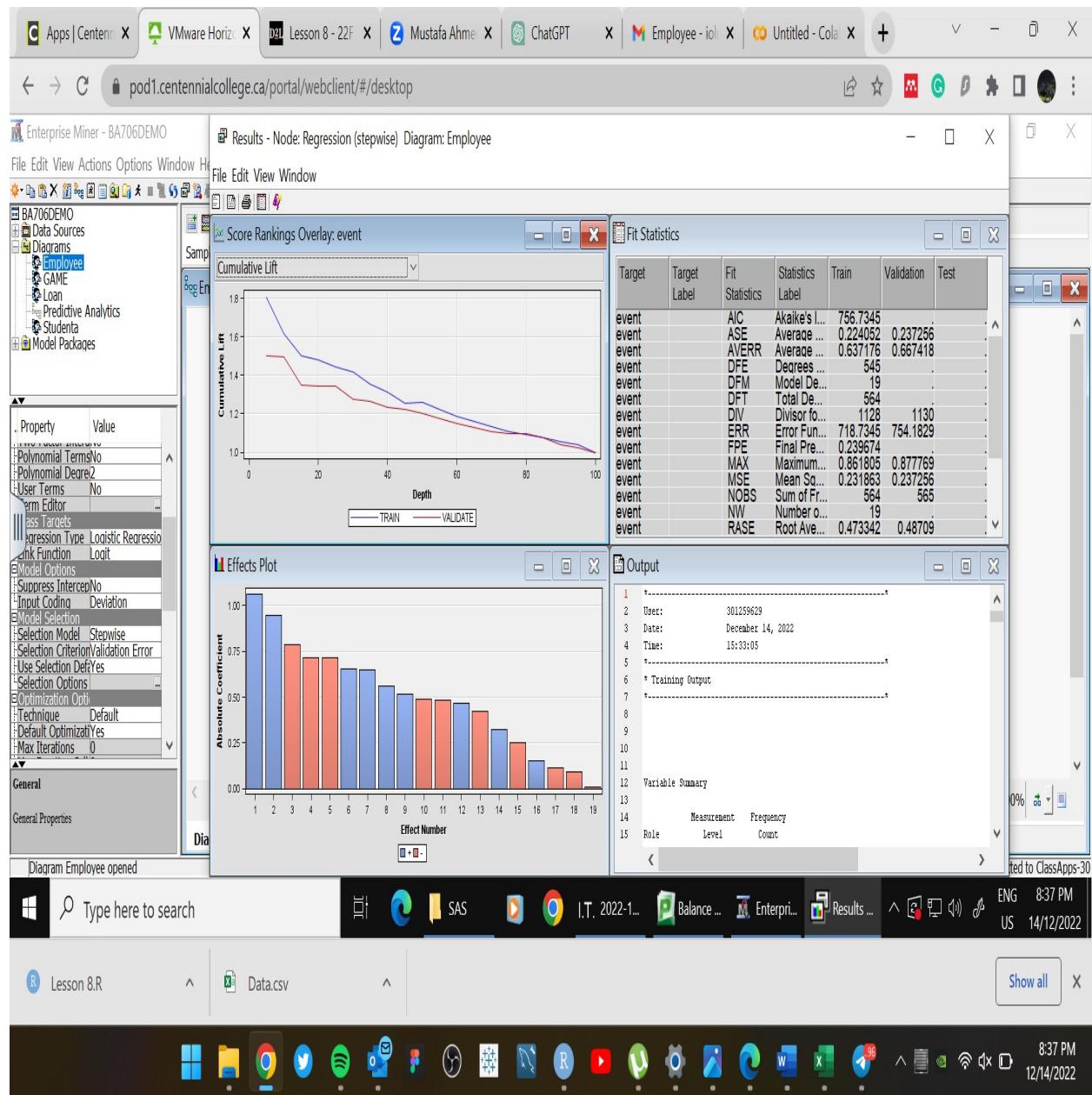
	PoinT Estimate	When compared with the Transport industry
REP_industry	Building vs transport	7.635 Employees in the Building industry are 7.6 times more likely to quit.
REP_industry	Agric vs transport	6.769 Employees in the Agriculture industry are 6.8 times more likely to quit.
REP_industry	Pharma vs transport	5.034 Employees in the Pharmaceutical industry are 5 times more likely to quit.
REP_industry	Banks vs transport	4.600 Employees in the Banking industry are 4.6 times more likely to quit.
REP_industry	Consult vs transport	4.408 Employees in the Consulting industry are 4.4 times more likely to quit.
REP_industry	Mining vs transport	3.637 Employees in the Mining industry are 3.6 times more likely to quit.
REP_industry	State vs transport	3.063 Employees in the State industry are 3 times more likely to quit.

REP_industry	Unknown vs transport	2.606	Employees in the all other industries not listed are 2.6 times more likely to quit.
REP_industry	Telecom vs transport	2.356	Employees in the Telecom industry are 2.4% more likely to quit.
REP_industry	Retail vs transport	1.723	Employees in the Retail industry are 72.3% more likely to quit.
REP_industry	Manufacture vs transport	1.628	Employees in the Manufacture industry are 62.8% more likely to quit.
REP_industry	PowerGen vs sport	1.292	Employees in the Power generating industry are 29.2% times more likely to quit.
REP_industry	RealEstate vs transport	1.291	Employees in the Real Estate industry are 29.1% times more likely to quit.
REP_industry	IT vs transport	0.880	Employees in the IT industry are 12% less likely to quit.
			When compared with Sales Profession

REP_profession	HR vs Sales	1.762	Employees in professions other than HR and IT are 76% more likely to quit
REP_profession	HR vs Sales	0.715	Employees in HR are 28.5% less likely to quit
REP_profession	IT vs Sales	0.563	Employees in IT are 43.7% less likely to quit
Other Variables			
REP_way	Motor vs foot	2.531	Employees who go to the office with a vehicle are 2.5 times more likely to quit than those who walk to the office.

## STEPWISE REGRESSION

The stepwise model also used 18 variables and had the same outcome as the forward regression with an ASE of 0.237256.



Two screenshots of the Enterprise Miner software interface showing stepwise regression results for the "Employee" dataset.

**Screenshot 1: Fit Statistics**

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		AIC	Akaike's Information Cr...	756.7345		
event		ASE	Average Square Error	0.224602	0.237256	
event		AMERR	Average Model Function	0.637176	0.657418	
event		DFE	Degrees of Freedom f...	545		
event		DFM	Model Degrees of Free...	19		
event		DIT	Total Degrees of Free...	564		
event		DIV	Divisor for ASE	1128	1130	
event		ERR	Error Function	718.7345	754.1829	
event		FINAL	Final Iteration Error	0.239675		
event		MAX	Maximum Absolute Error	0.581005	0.877769	
event		MSE	Mean Square Error	0.231863	0.237256	
event		NBDS	Sum of Frequencies	564	565	
event		NW	Number of Weights	19		
event		RASE	Root Average Sum of...	0.473342	0.48709	
event		RFPE	Root Final Prediction...	0.489565		
event		RMSE	Root Mean Squared Er...	0.481522	0.48709	
event		SBD	Sum of Deviation Cr...	639.056		
event		SSE	Sum of Squared Errors	252.031	268.0996	
event		SUMW	Sum of Case Weights ...	1128	1130	
event		MISC	Misclassification Rate	0.37234	0.410619	

**Screenshot 2: Odds Ratio Estimates**

Effect	Point Estimate
REP_industry Agriculture vs transport	6.769
REP_industry Banks vs transport	4.800
REP_industry Building vs transport	7.635
REP_industry Construction vs transport	4.400
REP_industry IT vs transport	1.198
REP_industry Mining vs transport	3.637
REP_industry Pharma vs transport	5.034
REP_industry Power vs transport	1.390
REP_industry RealEstate vs transport	1.391
REP_industry Retail vs transport	1.723
REP_industry State vs transport	3.063
REP_industry Telecom vs transport	2.356
REP_industry Unknown vs transport	2.406
REP_profession manufacturer vs transport	1.629
REP_profession HR vs Sales	0.715
REP_profession HR vs Sales	0.563
REP_profession Other vs Sales	1.762
REP_way Motor vs foot	2.531

## Interpretation of Stepwise Regression

Variables		Point Estimate	When compared to Transport Industry
Rep_Industry	Agriculture vs transport	6.769	Employees in the

			Agriculture industry are 6.8 times more likely to quit
Rep_Industry	Banks vs transport	4.600	Employees in the Bank industry are 4.6 times more likely to quit
Rep_Industry	Building vs transport	7.635	Employees in the Building industry are 7.6 times more likely to quit
Rep_Industry	Consult vs transport	4.408	Employees in the Consult industry are 4.4 times more likely to quit
Rep_Industry	IT vs transport	1.198	Employees in the IT industry are 19.8% more likely to quit
Rep_Industry	Mining vs transport	3.637	Employees in the Mining industry are 3.6 times more likely to quit
Rep_Industry	Pharma vs transport	5.034	Employees in the Pharma industry are 5 times more likely to quit
Rep_Industry	Power Generation vs transport	1.292	Employees in the Power Generation industry are 29.2% more likely to quit
Rep_Industry	Real Estate vs transport	1.291	Employees in the Real Estate industry are 29.1% more likely to quit
Rep_Industry	Retail vs transport	1.723	Employees in the Retail Industry are 72.3% more likely to quit
Rep_Industry	State vs transport	3.063	Employees in the State industry are 3.1 times more likely to quit

Rep_Industry	Telecom vs transport	2.356	Employees in the Telecom industry are 2.4 times more likely to quit
Rep_Industry	Unknown vs transport	2.606	Employees in the Unknown industry are 2.6 times more likely to quit
Rep_Industry	Manufacture vs transport	1.628	Employees in the Manufacture industry are 62.8% more likely to quit
			<b>When compared with Sales Profession</b>
Rep_Profession	HR vs Sales	0.715	Employees in HR are 28.5% less likely to quit
Rep_Profession	IT vs Sales	0.563	Employees in IT are 43.7% less likely to quit
Rep_Profession	Other vs Sales	1.762	Employees in Other Professions are 76.2% more likely to quit
			<b>Other Variable</b>
Rep_Way	Motor vs Foot	2.531	Employees who come to work by Motor are 2.5 times more likely to quit than those who come by Foot

## BACKWARD REGRESSION

The backward regression model also used 18 variables and had the same outcome as the forward regression and stepwise regression with an ASE of 0.237256.

The screenshot shows the Enterprise Miner interface with the following details:

- Title Bar:** Apps Center, VMware Horizon, Lesson 8 - 22F, Mustafa Ahme, ChatGPT, Employee - iol, Untitled - Cola.
- Address Bar:** pod1.centennialcollege.ca/portal/webclient/#/desktop
- Left Panel (Data Sources):** BA706DEMO, Data Sources, Diagrams, Employee, GAME, Loan, Predictive Analytics, Student, Model Packages.
- Properties Panel:** General Properties, Diagram Employee opened.
- Score Rankings Overlay: event** window:
  - Cumulative Lift Plot:** Shows Cumulative Lift vs Depth (0 to 100). The plot has two lines: TRAIN (blue) and VALIDATE (red). The lift starts at approximately 1.6 and decreases as depth increases.
  - Fit Statistics Table:**

Target	Target Label	Fit Statistics	Train	Validation	Test
event	Target	AIC	756.7345		
event		ASE	0.224052	0.237256	
event		AVERGE	Average ...	0.637176	0.667418
event		DFE	Degrees ...	545	
event		DFM	Model De...	19	
event		DFT	Total De...	564	
event		DIV	Divisor fo...	1128	1130
event		ERR	Error Fun...	718.7345	754.1829
event		FPE	Final Pre...	0.239674	
event		MAX	Maximum...	0.861805	0.877769
event		MSE	Mean Sq...	0.231863	0.237256
event		NOBS	Sum of Fr...	564	565
event		NW	Number o...	19	
event		RASE	Root Ave...	0.473342	0.48709
- Effects Plot:** Shows Absolute Coefficient vs Effect Number (1 to 19). Blue bars represent positive coefficients, and red bars represent negative coefficients.
- Output Window:** Displays log information:

```
1 -----
2 User: 301259629
3 Date: December 14, 2022
4 Time: 15:59:55
5 -----
6 * Training Output
7 -----
8 -----
9 -----
10 -----
11 -----
12 Variable Summary
13 -----
14 Role Measurement Frequency
15 Level Count
```
- Taskbar:** Type here to search, Windows Start, Data.csv, I.T. 2022-1..., Balance..., Enterprise Miner, Results..., Show all.
- System Tray:** 8:39 PM, 14/12/2022.

File Edit View Actions Options Window Help

Results - Node: Regression (back) Diagram: Employee

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		AIC	Akaike's Information Cr...	756.7345		
event		ASE	Average Square Error	0.224052	0.237256	
event		AERR	Average Error Function	0.637176	0.657418	
event		DFE	Degrees of Freedom f...	545		
event		DFM	Model Degrees of Free...	19		
event		DIT	Total Degrees of Free...	564		
event		DIV	Divisor for ASE	1128		
event		ERR	Error Function	718.7345	754.1829	
event		FSE	Final Solution Error	0.239675		
event		MAX	Maximum Absolute Error	0.591005	0.877769	
event		MSE	Mean Square Error	0.231863	0.237256	
event		NBS	Sum of Frequencies	564	565	
event		NW	Number of Wrong W...	19		
event		RASE	Root Average Sum of...	0.473342	0.48709	
event		RFPE	Root Final Prediction ...	0.489565		
event		RMSE	Root Mean Squared E...	0.481522	0.48709	
event		SDC	Sum of Deviation Err...	639.055		
event		SSE	Sum of Squared Errors	252.031	268.0996	
event		SUMW	Sum of Case Weights ...	1128	1130	
event		MISC	Misclassification Rate	0.37234	0.410619	

Properties Value

- Polynomial Terms No
- Polynomial Degree 2
- User Terms No
- Term Order
- lass Targets
- Regression Type Logistic Regression
- Link Function Logit
- Model Options
- Suppress Intercept No
- Input Coding Deviation
- Output Selection
- Selection Model Backward
- Selection Criterion Validation Error
- Use Selection Def Yes
- Cost Function
- Optimization Options
- Technique Default
- Default Optimizat Yes
- Max Iterations 0

General Properties

Diagram Employee opened

Type here to search

File Edit View Window

Results - Node: Regression (back) Diagram: Employee

File Edit View Window

Output

Effect	Point Estimate
1535 REP_industry IT	1 -0.7874 0.3239 5.91 0.0151 0.455
1536 REP_industry Mining	1 0.3232 0.5615 0.33 0.5649 1.392
1537 REP_industry Pharms	1 0.6482 0.6634 0.95 0.3285 1.912
1538 REP_industry PowerGeneration	1 -0.1022 0.2622 2.27 0.1616 0.401
1539 REP_industry RetailTrade	1 -0.7123 0.7216 0.97 0.8346 0.461
1540 REP_industry Retailware	1 -0.4241 0.2148 3.90 0.0484 0.654
1541 REP_industry State	1 0.1514 0.4534 0.11 0.7384 1.164
1542 REP_industry Telecom	1 -0.1108 0.5946 0.05 0.8261 0.895
1543 REP_industry Unknown	1 -0.0257 0.2537 0.07 0.7327 0.590
1544 REP_industry manufacture	1 -0.4808 0.2629 3.34 0.0674 0.618
1545 REP_profession HR	1 -0.2497 0.1697 2.16 0.1412 0.779
1546 REP_profession IT	1 -0.4809 0.3206 2.33 0.1268 0.613
1547 REP_profession Other	1 0.6525 0.2176 8.99 0.0027 1.920
1548 REP_way Motor	1 0.4643 0.1495 9.63 0.0019 1.591
1549	
1550	Odds Ratio Estimates
1551	
1552	
1553	
1554	
1555	
1556	
1557	
1558	
1559	
1560	
1561	
1562	
1563	
1564	
1565	
1566	
1567	
1568	
1569	
1570	
1571	
1572	
1573	
1574	
1575	
1576	*
1577	* Score Output
1578	*
1579	*
1580	*
1581	* Report Output
1582	*
1583	*
1584	*
1585	*
1586	*
1587	*
1588	*
1589	*
1590	*
1591	*
1592	*
1593	*
1594	*
1595	*
1596	*
1597	*
1598	*
1599	*
1600	*
1601	*
1602	*
1603	*
1604	*
1605	*
1606	*
1607	*
1608	*
1609	*
1610	*
1611	*
1612	*
1613	*
1614	*
1615	*
1616	*
1617	*
1618	*
1619	*
1620	*
1621	*
1622	*
1623	*
1624	*
1625	*
1626	*
1627	*
1628	*
1629	*
1630	*
1631	*
1632	*
1633	*
1634	*
1635	*
1636	*
1637	*
1638	*
1639	*
1640	*
1641	*
1642	*
1643	*
1644	*
1645	*
1646	*
1647	*
1648	*
1649	*
1650	*
1651	*
1652	*
1653	*
1654	*
1655	*
1656	*
1657	*
1658	*
1659	*
1660	*
1661	*
1662	*
1663	*
1664	*
1665	*
1666	*
1667	*
1668	*
1669	*
1670	*
1671	*
1672	*
1673	*
1674	*
1675	*
1676	*
1677	*
1678	*
1679	*
1680	*
1681	*
1682	*
1683	*
1684	*
1685	*
1686	*
1687	*
1688	*
1689	*
1690	*
1691	*
1692	*
1693	*
1694	*
1695	*
1696	*
1697	*
1698	*
1699	*
1700	*
1701	*
1702	*
1703	*
1704	*
1705	*
1706	*
1707	*
1708	*
1709	*
1710	*
1711	*
1712	*
1713	*
1714	*
1715	*
1716	*
1717	*
1718	*
1719	*
1720	*
1721	*
1722	*
1723	*
1724	*
1725	*
1726	*
1727	*
1728	*
1729	*
1730	*
1731	*
1732	*
1733	*
1734	*
1735	*
1736	*
1737	*
1738	*
1739	*
1740	*
1741	*
1742	*
1743	*
1744	*
1745	*
1746	*
1747	*
1748	*
1749	*
1750	*
1751	*
1752	*
1753	*
1754	*
1755	*
1756	*
1757	*
1758	*
1759	*
1760	*
1761	*
1762	*
1763	*
1764	*
1765	*
1766	*
1767	*
1768	*
1769	*
1770	*
1771	*
1772	*
1773	*
1774	*
1775	*
1776	*
1777	*
1778	*
1779	*
1780	*
1781	*
1782	*
1783	*
1784	*
1785	*
1786	*
1787	*
1788	*
1789	*
1790	*
1791	*
1792	*
1793	*
1794	*
1795	*
1796	*
1797	*
1798	*
1799	*
1800	*
1801	*
1802	*
1803	*
1804	*
1805	*
1806	*
1807	*
1808	*
1809	*
1810	*
1811	*
1812	*
1813	*
1814	*
1815	*
1816	*
1817	*
1818	*
1819	*
1820	*
1821	*
1822	*
1823	*
1824	*
1825	*
1826	*
1827	*
1828	*
1829	*
1830	*
1831	*
1832	*
1833	*
1834	*
1835	*
1836	*
1837	*
1838	*
1839	*
1840	*
1841	*
1842	*
1843	*
1844	*
1845	*
1846	*
1847	*
1848	*
1849	*
1850	*
1851	*
1852	*
1853	*
1854	*
1855	*
1856	*
1857	*
1858	*
1859	*
1860	*
1861	*
1862	*
1863	*
1864	*
1865	*
1866	*
1867	*
1868	*
1869	*
1870	*
1871	*
1872	*
1873	*
1874	*
1875	*
1876	*
1877	*
1878	*
1879	*
1880	*
1881	*
1882	*
1883	*
1884	*
1885	*
1886	*
1887	*
1888	*
1889	*
1890	*
1891	*
1892	*
1893	*
1894	*
1895	*
1896	*
1897	*
1898	*
1899	*
1900	*
1901	*
1902	*
1903	*
1904	*
1905	*
1906	*
1907	*
1908	*
1909	*
1910	*
1911	*
1912	*
1913	*
1914	*
1915	*
1916	*
1917	*
1918	*
1919	*
1920	*
1921	*
1922	*
1923	*
1924	*
1925	*
1926	*
1927	*
1928	*
1929	*
1930	*
1931	*
1932	*
1933	*
1934	*
1935	*
1936	*
1937	*
1938	*
1939	*
1940	*
1941	*
1942	*
1943	*
1944	*
1945	*
1946	*
1947	*
1948	*
1949	*
1950	*
1951	*
1952	*
1953	*
1954	*
1955	*
1956	*
1957	*
1958	*
1959	*
1960	*
1961	*
1962	*
1963	*
1964	*
1965	*
1966	*
1967	*
1968	*
1969	*
1970	*
1971	*
1972	*
1973	*
1974	*
1975	*
1976	*
1977	*
1978	*
1979	*
1980	*
1981	*
1982	*
1983	*
1984	*
1985	*
1986	*
1987	*
1988	*
1989	*
1990	*
1991	*
1992	*
1993	*
1994	*
1995	*
1996	*
1997	*
1998	*
1999	*
2000	*
2001	*
2002	*
2003	*
2004	*
2005	*
2006	*
2007	*
2008	*
2009	*
2010	*
2011	*
2012	*
2013	*
2014	*
2015	*
2016	*
2017	*
2018	*
2019	*
2020	*
2021	*
2022	*
2023	*
2024	*
2025	*
2026	*
2027	*
2028	*
2029	*
2030	*
2031	*
2032	*
2033	*
2034	*
2035	*
2036	*
2037	*
2038	*
2039	*
2040	*
2041	*
2042	*
2043	*
2044	*
2045	*
2046	*
2047	*
2048	*
2049	*
2050	*
2051	*
2052	*
2053	*
2054	*
2055	*
2056	*
2057	*
2058	*
2059	*
2060	*
2061	*
2062	*
2063	*
2064	*
2065	*
2066	*
2067	*
2068	*
2069	*
2070	*
2071	*
2072	*
2073	*
2074	*
2075	*
2076	*
2077	*
2078	*
2079	*
2080	*
2081	*
2082	*
2083	*
2084	*
2085	*
2086	*
2087	*

## Interpretation of Backward Regression

Variables		Point Estimate	When compared to Transport Industry
Rep_Industry	Agriculture vs transport	6.769	Employees in the Agriculture industry are 6.8 times more likely to quit
Rep_Industry	Banks vs transport	4.600	Employees in the Bank industry are 4.6 times more likely to quit
Rep_Industry	Building vs transport	7.635	Employees in the Building industry are 7.6 times more likely to quit
Rep_Industry	Consult vs transport	4.408	Employees in the Consult industry are 4.4 times more likely to quit
Rep_Industry	IT vs transport	1.198	Employees in the IT industry are 19.8% more likely to quit
Rep_Industry	Mining vs transport	3.637	Employees in the Mining industry are 3.6 times more likely to quit
Rep_Industry	Pharma vs transport	5.034	Employees in the Pharma industry are 5 times more likely to quit
Rep_Industry	Power Generation vs transport	1.292	Employees in the Power Generation industry are 29.2% more likely to quit
Rep_Industry	Real Estate vs transport	1.291	Employees in the Real Estate industry are 29.1% more likely to quit

Rep_Industry	Retail vs transport	1.723	Employees in the Retail Industry are 72.3% more likely to quit
Rep_Industry	State vs transport	3.063	Employees in the State industry are 3.1 times more likely to quit
Rep_Industry	Telecom vs transport	2.356	Employees in the Telecom industry are 2.4 times more likely to quit
Rep_Industry	Unknown vs transport	2.606	Employees in the Unknown industry are 2.6 times more likely to quit
Rep_Industry	Manufacture vs transport	1.628	Employees in the Manufacture industry are 62.8% more likely to quit
			<b>When compared with Sales Profession</b>
Rep_Profession	HR vs Sales	0.715	Employees in HR are 28.5% less likely to quit
Rep_Profession	IT vs Sales	0.563	Employees in IT are 43.7% less likely to quit
Rep_Profession	Other vs Sales	1.762	Employees in Other Professions are 76.2% more likely to quit
			<b>Other Variable</b>
Rep_Way	Motor vs Foot	2.531	Employees who come to work by Motor are 2.5 times more likely to quit than those who come by Foot

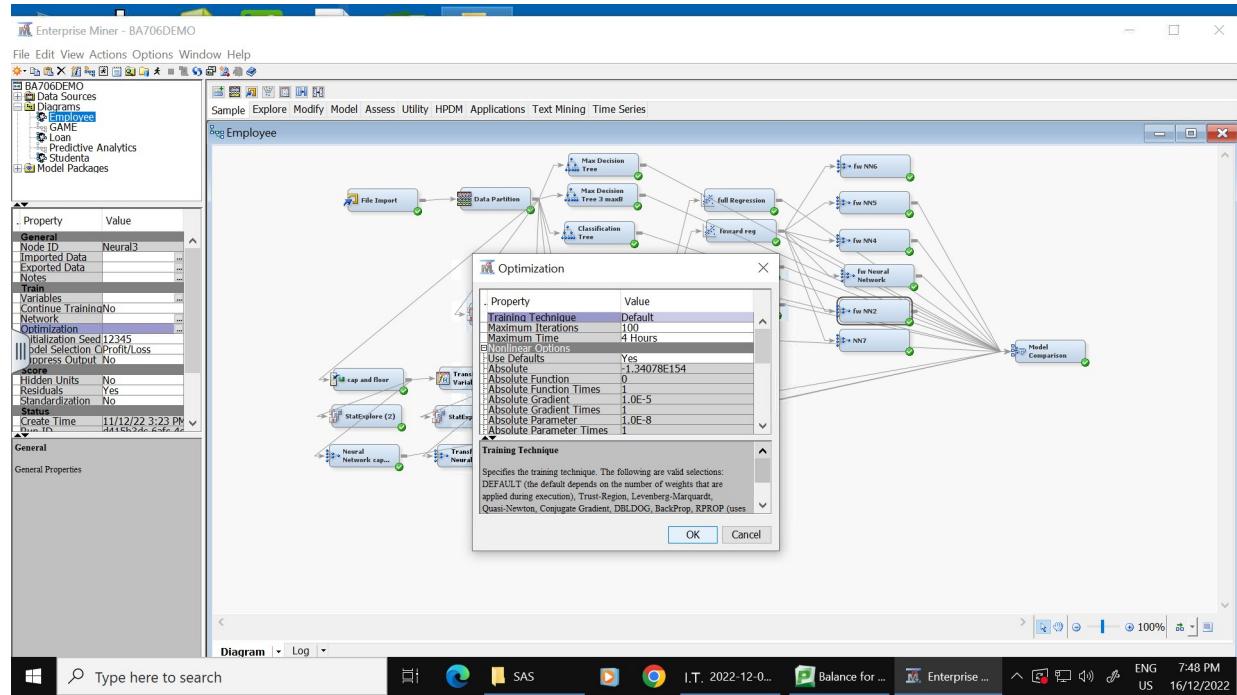
With the full regression having an ASE of 0.240878 and Forward, Stepwise , and Backward regression models all having the same ASE of 0.237256 we have an option of choosing from any of the three with the better average squared error. We chose the Forward regression model because of its additive style of selecting variables.

## NEURAL NETWORK

All neural networks were run using 100 iterations, and preliminary training was turned off. The distinguishing factor for all the neural networks run was the number of hidden units used.

### NEURAL NETWORK WITH 2 HIDDEN UNITS

This neural network node was run using two hidden units. The ASE was 0.234308.



Enterprise Miner - BA706DEMO

File Edit View Actions Options Window Help

BA706DEMO

- Data Sources
- Variables
- Employee**
- GAME
- Predictive Analytics
- Studenta
- Model Packages

Property Value

Exported Data Notes

Variables Unique TrainingNo

Network Optimization

Initialization Seed 12345

Model Selection GProfit/Loss

Suppress Output No

Session Hidden Units No

Residuals Yes

Standardization No

Status Create Time 11/12/22 3:23 PM

Run ID f03d08c8-b50d-4

Last Error

Last Status Complete

General

General Properties

Diagram Log

File Import Data Partitions

StatExplorer

Model Comparison

Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	2
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Standard Deviation	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

Architecture

Specifies which network architecture is used in constructing the network. The following are valid selections: generalized linear model, multilayer perceptron, ordinary radial basis function with equal widths, ordinary radial basis function with unequal widths, normalized radial basis function with equal heights, normalized radial basis function with equal volumes, normalized radial basis function with equal widths, normalized radial basis function with equal widths and heights, normalized radial basis function with unequal widths and heights and a User specified network.

OK Cancel

Type here to search

pod1.centennialcollege.ca/portal/webclient/#/desktop

File Edit View Actions Options Window Help

Enterprise Miner - BA706DEMO

File Edit View Window

Score Rankings Overlay: event

Cumulative Lift

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	DFT	Total Degress	564	.	.	.
event	DFE	Degrees	523	.	.	.
event	DFM	Model De...	41	.	.	.
event	NW	Number o...	41	.	.	.
event	AV	Average ...	792.4392	.	.	.
event	GBC	Schwarz...	970.1701	.	.	.
event	ASE	Averaga...	0.220121	0.234306	.	.
event	MAX	Maximum...	0.850708	0.841782	.	.
event	DV	Degrees	523	170	.	.
event	NOBS	Sum of Fr...	564	565	.	.
event	RASE	Root Ave...	0.469171	0.484054	.	.
event	SSE	Sum of S...	248.2586	264.7718	.	.
event	SMW	FPE	128	1130	.	.
event		Final Pre...	0.254634			

Iteration Plot

Average Square Error

Output

```

1 *-----
2 User: 301239629
3 Date: December 14, 2022
4 File: 17134152
5 *-----
6 * Training Output
7 *
8 *
9 *
10 *
11 *
12 *
13 Variable Summary
14 Role Measurement Frequency Count
15

```

Diagram Employee opened

Type here to search

Lesson 8.R Data.csv

File Edit View Actions Options Window Help

File Import Data Partitions

StatExplorer

Model Comparison

Network

Score Rankings Overlay: event

Cumulative Lift

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	DFT	Total Degress	564	.	.	.
event	DFE	Degrees	523	.	.	.
event	DFM	Model De...	41	.	.	.
event	NW	Number o...	41	.	.	.
event	AV	Average ...	792.4392	.	.	.
event	GBC	Schwarz...	970.1701	.	.	.
event	ASE	Averaga...	0.220121	0.234306	.	.
event	MAX	Maximum...	0.850708	0.841782	.	.
event	DV	Degrees	523	170	.	.
event	NOBS	Sum of Fr...	564	565	.	.
event	RASE	Root Ave...	0.469171	0.484054	.	.
event	SSE	Sum of S...	248.2586	264.7718	.	.
event	SMW	FPE	128	1130	.	.
event		Final Pre...	0.254634			

Iteration Plot

Average Square Error

Output

```

1 *-----
2 User: 301239629
3 Date: December 14, 2022
4 File: 17134152
5 *-----
6 * Training Output
7 *
8 *
9 *
10 *
11 *
12 *
13 Variable Summary
14 Role Measurement Frequency Count
15

```

Show all

8:45 PM 12/14/2022

Screenshot of Enterprise Miner software interface showing results for a neural network model named "fw NN2".

**Results - Node: fw NN2 Diagram: Employee**

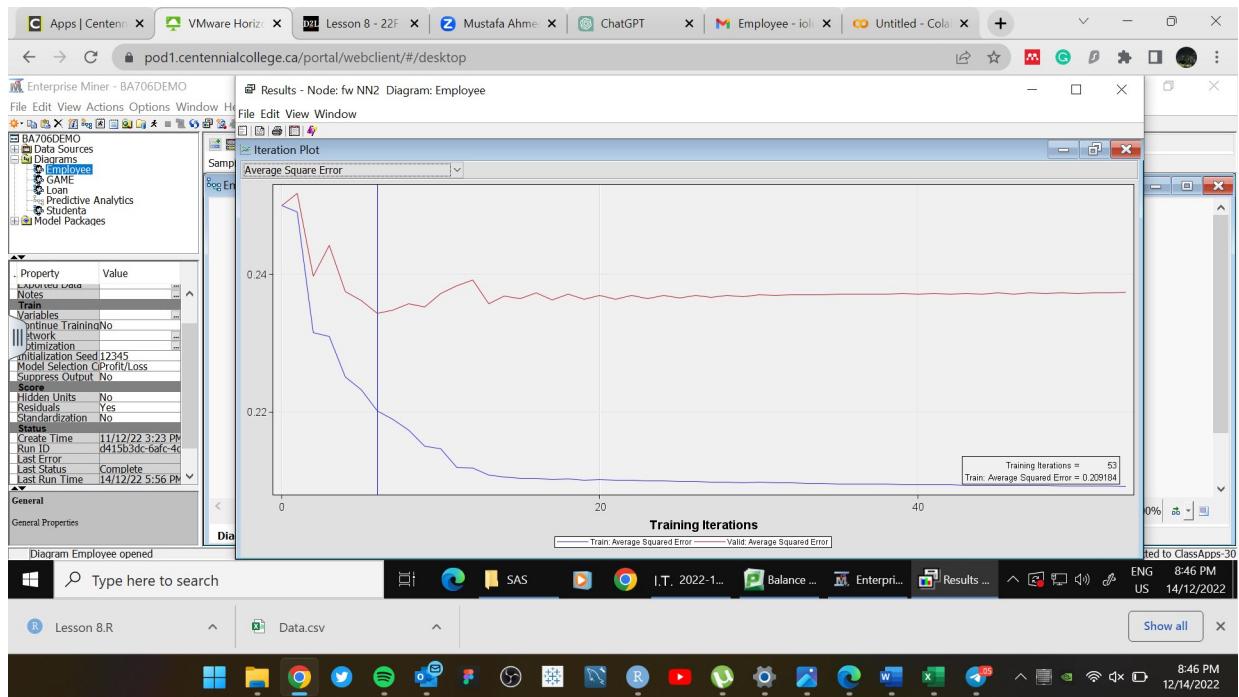
**Fit Statistics**

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event		DFT	Total Degrees of Free...	564	.	.
event		DDE	Degrees of Freedom	523	.	.
event		DFM	Model Degrees of Free...	41	.	.
event		NW	Number of Estimated ...	41	.	.
event		AIC	Akaike's Information Cr...	792.4329	792.4329	792.4329
event		GOC	Gordon's Criterion	792.4329	792.4329	792.4329
event		ASE	Average Squared Error	0.220121	0.234308	0.234308
event		MAX	Maximum Absolute Error	0.850708	0.841782	0.841782
event		DIV	Divisor for ASE	1128	.	.
event		NDBS	Sum of Base Squares	554	555	555
event		RASE	Root Average Squared...	0.469171	0.484054	0.484054
event		SSE	Sum of Squared Errors	248.2969	264.7679	264.7679
event		SUMW	Sum of Case Weights	554	1130	1130
event		FPE	Final Prediction Error	0.254634	.	.
event		MSE	Mean Squared Error	0.237377	0.234308	0.234308
event		RFPE	Root Final Prediction...	0.504612	.	.
event		RMSSE	Root Mean Squared E...	0.484054	0.484054	0.484054
event		AVER	Average Error Function	0.629816	0.661091	0.661091
event		ERR	Error Function	710.4329	747.0324	747.0324
event		MISC	Misclassification Rate	0.359929	0.39469	0.39469
event		WRONG	Number of Wrong Clas...	203	223	223

**Iteration Plot**

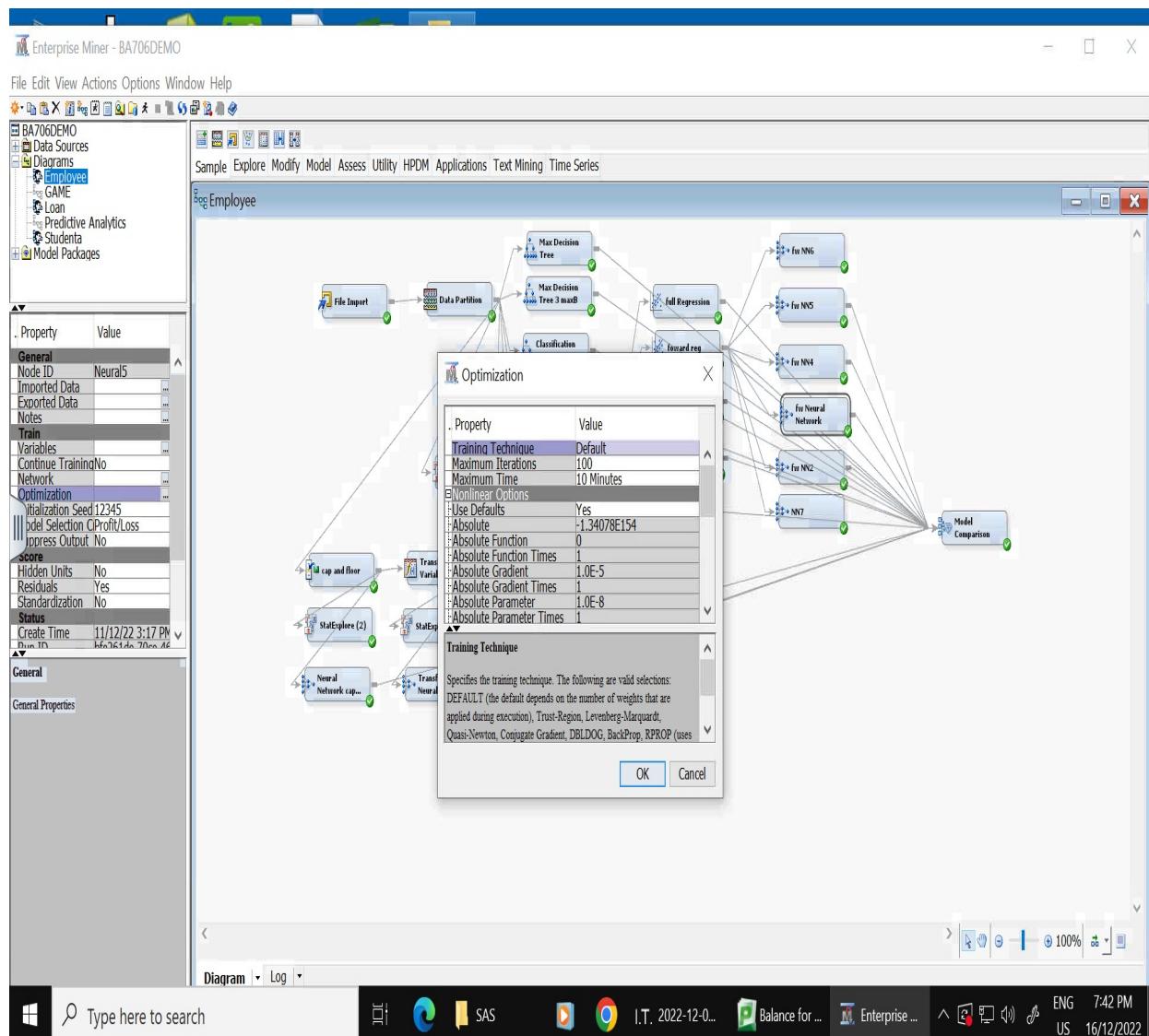
Training Iterations = 53  
Train: Average Squared Error = 0.209184  
Valid: Average Squared Error = 0.245

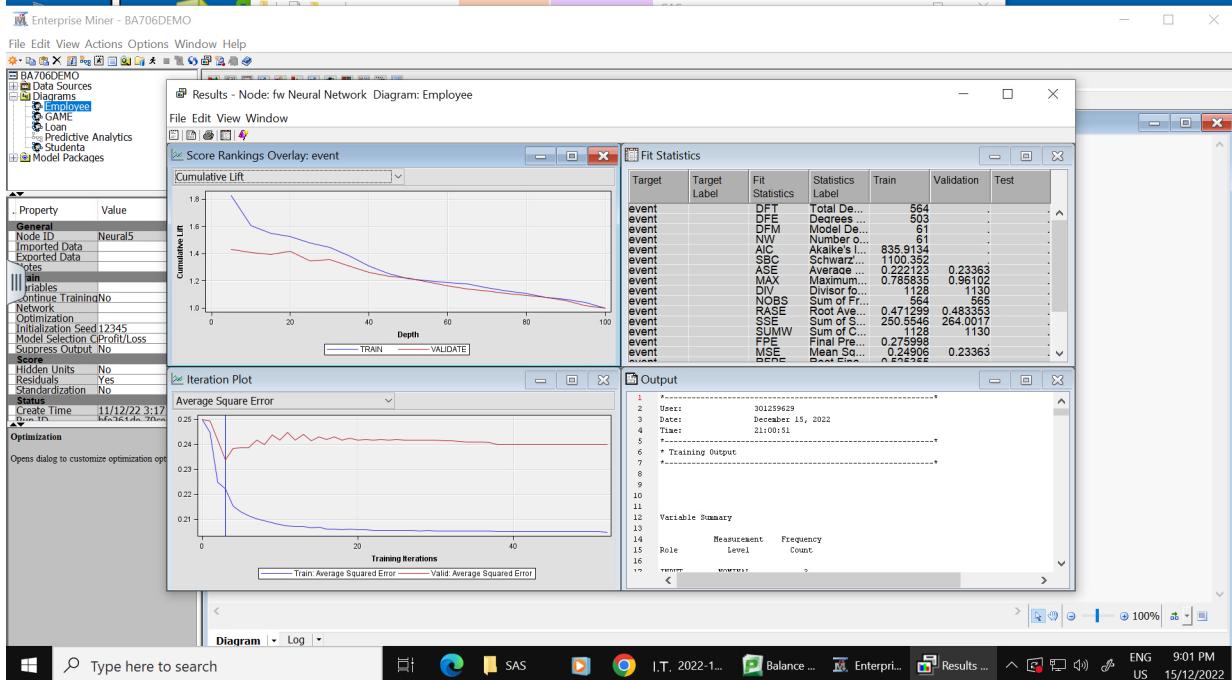
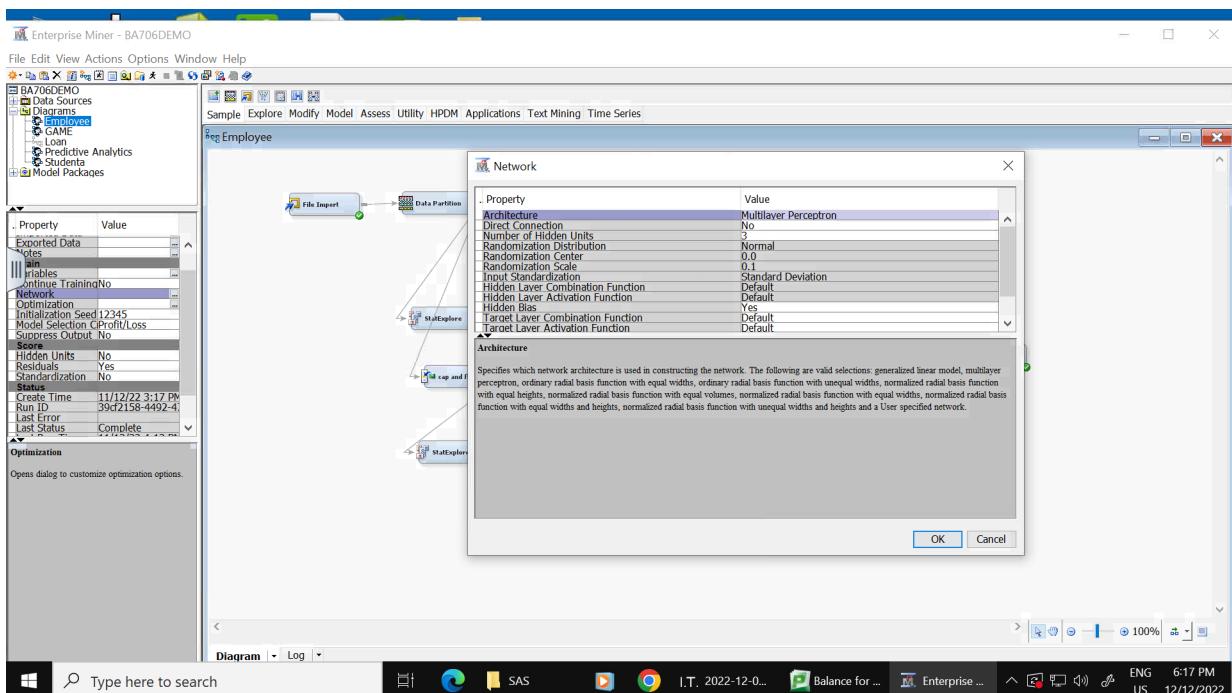
Here, we see that the average squared error of the training and validation data sets start off in opposite directions with the validation getting worse before it changes direction for improvement. The most improvement for the validation ASE was at iteration 6 then a slight increase occurred which is a sign of overfitting. At iteration 53 convergence occurred.



## **NEURAL NETWORK WITH 3 HIDDEN UNITS**

This neural network node was run using three hidden units. The ASE was 0.23363 which is an improvement from the neural network with 2 hidden units.



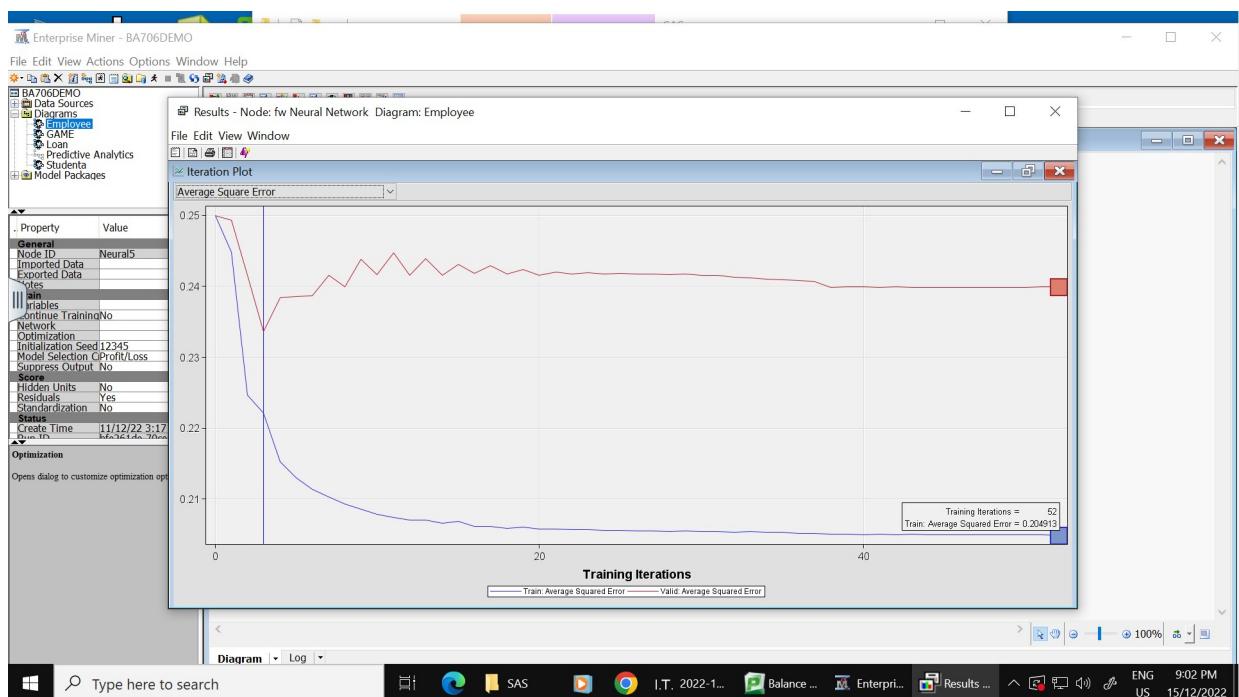


Screenshot of the Enterprise Miner software interface showing the results of a neural network run. The main window displays a table of fit statistics for training, validation, and test datasets.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	DFT	Total Degrees of Free...	564			
event	DIE	Degrees of Freedom	503			
event	DFM	Model Degrees of Free...	61			
event	NW	Number of Estimated ...	61			
event	AIC	Akaike's Information Cr...	835.9134			
event	GOC	Gordon's Criterion	1000.362			
event	ASE	Average Squared Error	0.222123	0.23363		
event	MAX	Maximum Absolute Error	0.785835	0.96102		
event	DIN	Divisor for ASE	1128	1120		
event	NDBS	Sum of Data Squares	554	555		
event	RASE	Root Average Squared...	0.471299	0.483353		
event	SSE	Sum of Squared Errors	250.5546	264.0017		
event	SUMW	Sum of Case Weights	1128	1130		
event	FPE	Final Prediction Error	0.275988			
event	MSE	Mean Squared Error	0.24906	0.23363		
event	RFPE	Root Final Prediction...	0.525355			
event	RMSSE	Root Mean Squared E...	0.498353			
event	AVER	Average Error Function	0.632902	0.660789		
event	ERR	Error Function	713.9134	746.6914		
event	MISC	Misclassification Rate	0.388298	0.39646		
event	WRONG	Number of Wrong Clas...	219	224		

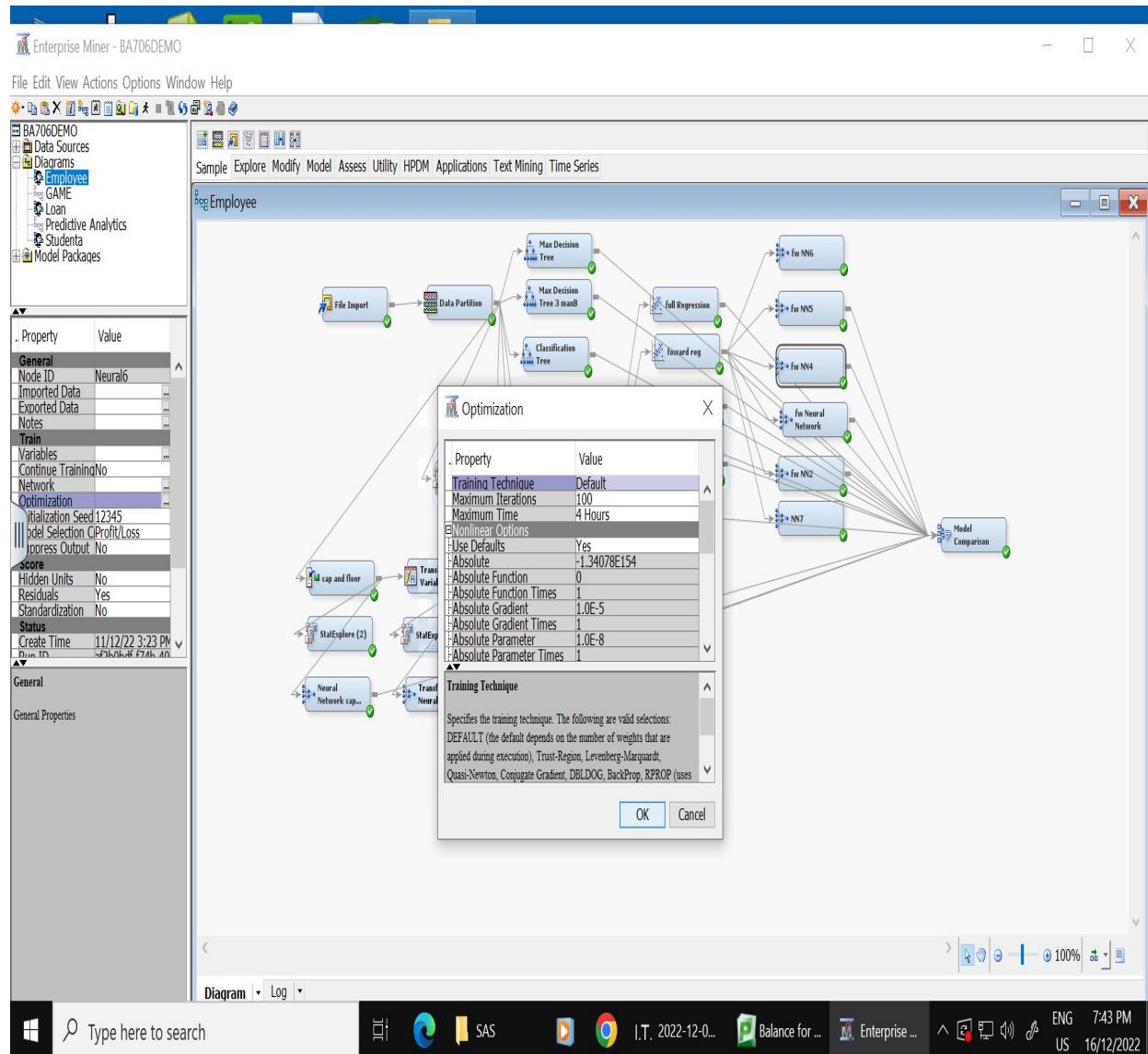
The desktop taskbar at the bottom shows various open applications including R, SAS, and Microsoft Office.

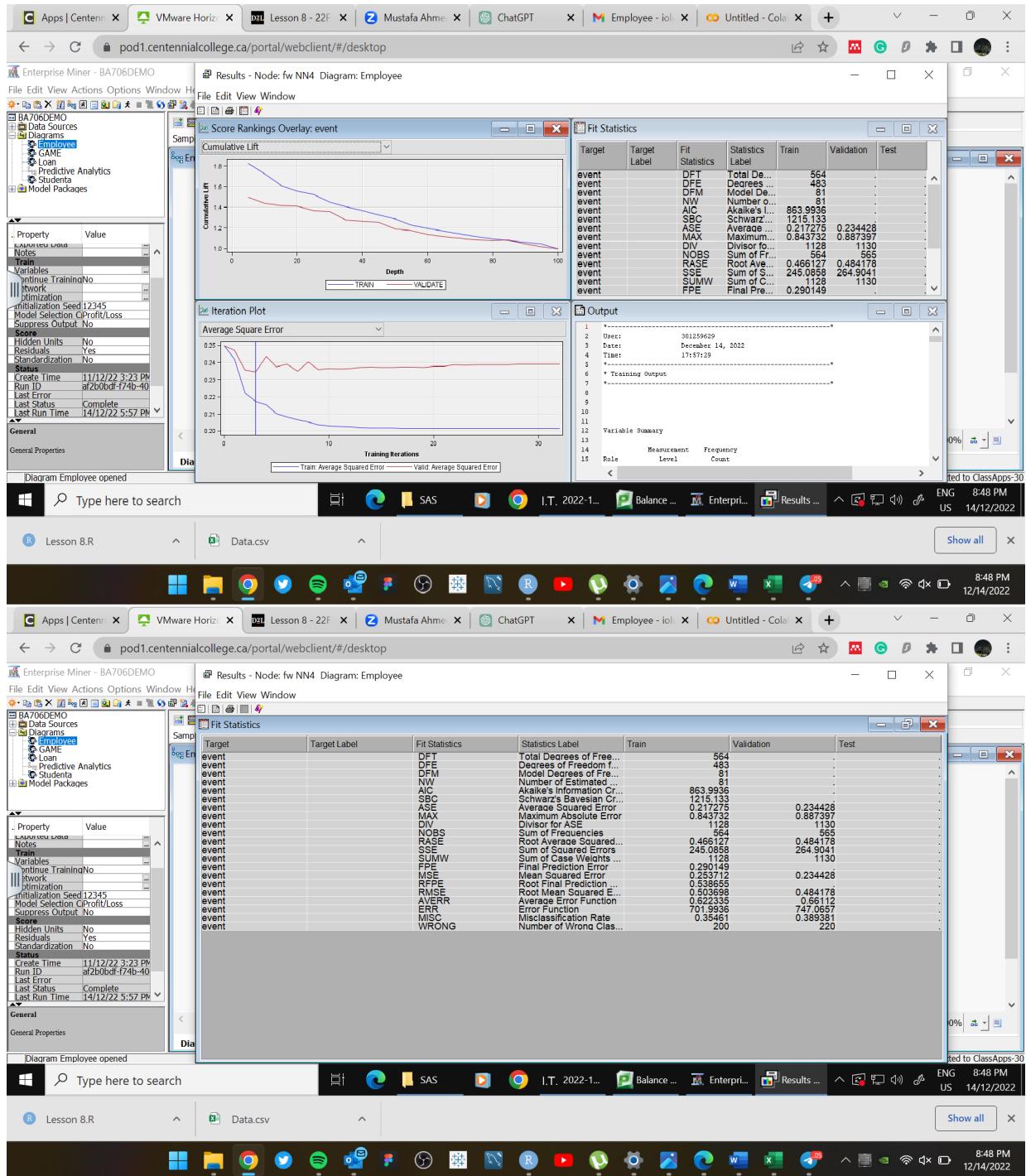
Here, we see that the average squared error of the training and validation data sets start off in the same direction of improvement till iteration 3 where the validation ASE starts to increase as a sign of overfitting. this occurs until the point of convergence at iteration 52.



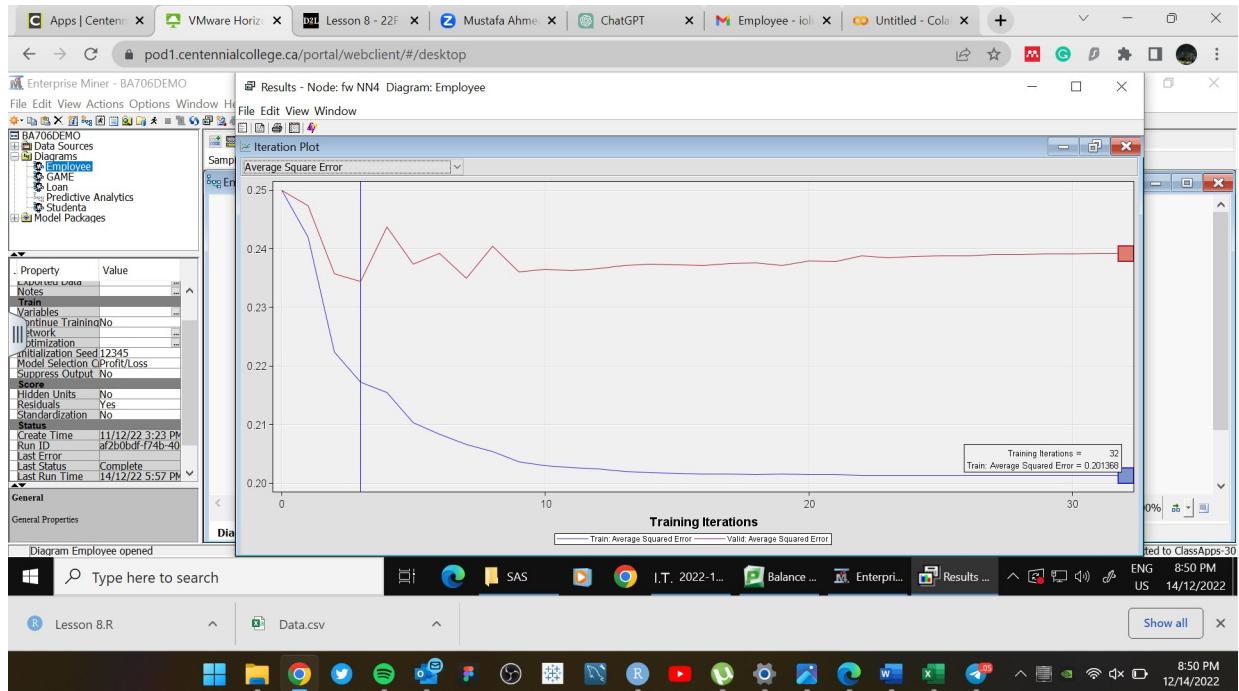
## NEURAL NETWORK WITH 4 HIDDEN UNITS

We continued to check for the best neural network model by increasing the number of hidden units to 4 as we noticed 3 hidden units was an improvement when compared to 2 hidden units. The result from the four hidden units neural network was even worse than that of the two hidden units with an ASE of 0.234428 .





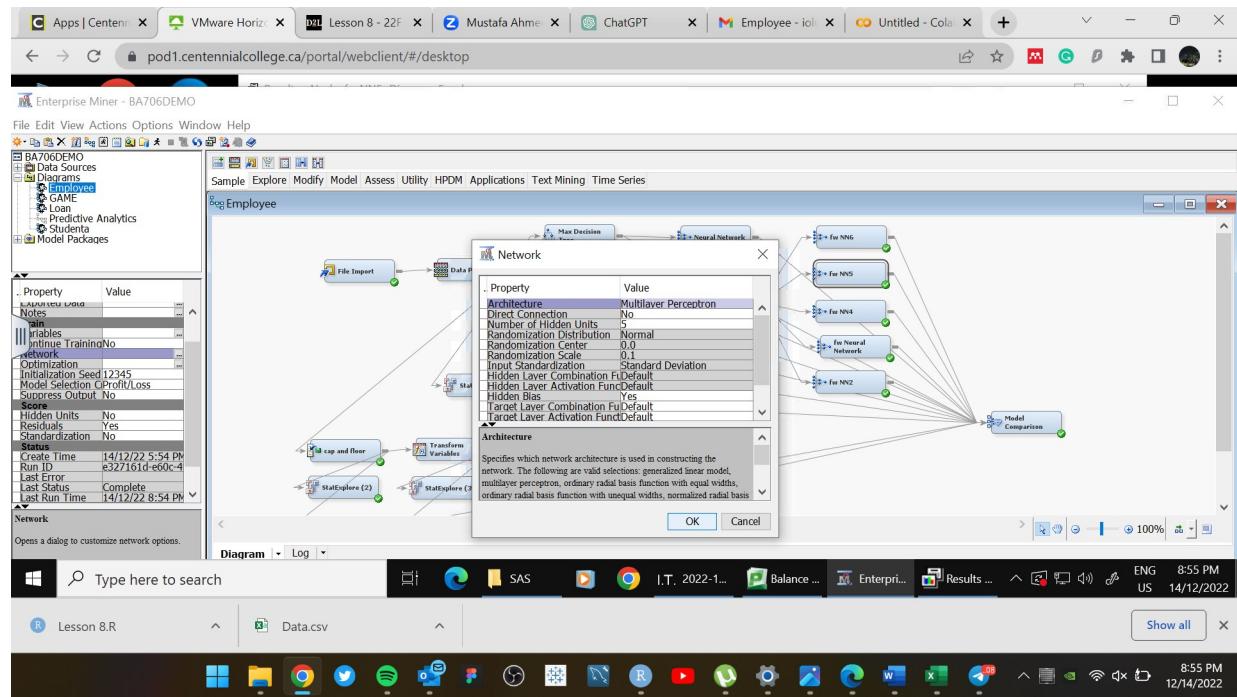
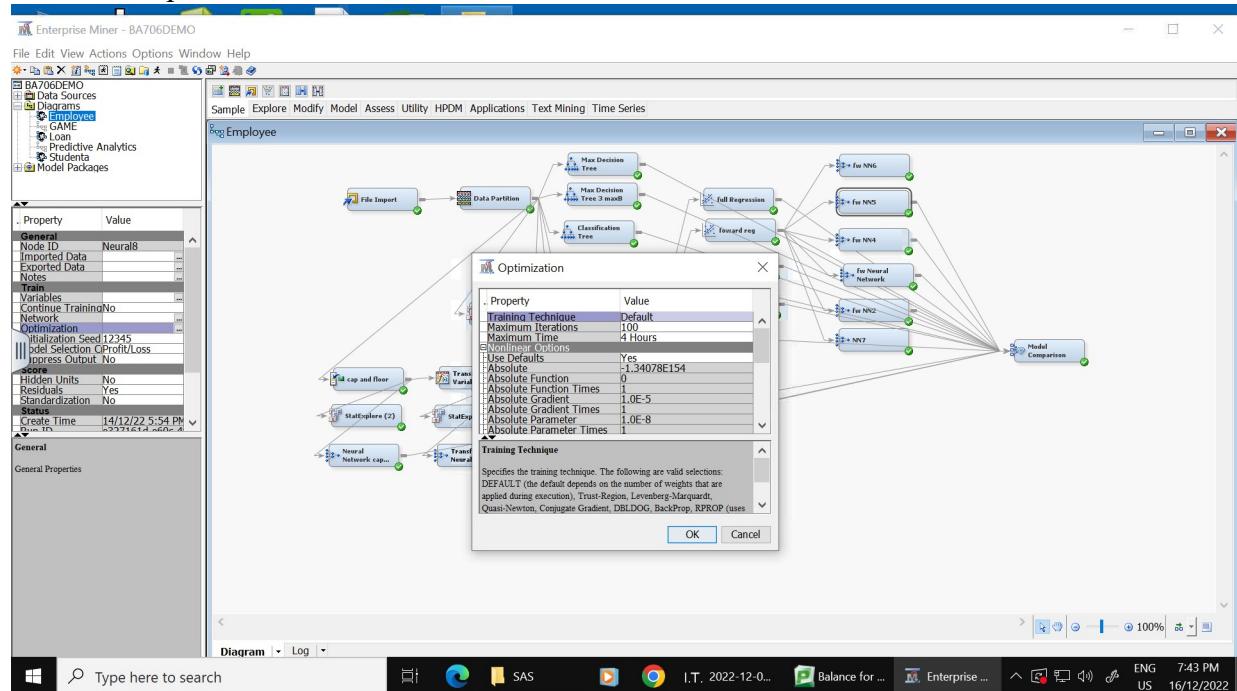
Here, we see that the average squared error of the training and validation data sets start off in the same direction of improvement and then the validation ASE starts to increase as a sign of overfitting. This occurs until the point of convergence at iteration 32.



## NEURAL NETWORK WITH 5 HIDDEN UNITS

We further increased the number of hidden units to 5 and the outcome ASE of 0.234067 was an improvement on the two hidden unit neural network but not as good as the three hidden unit neural network. This proves no further need to run additional hidden units as the aim is to find

## the least complex model with the best result.



Results - Node: fw NNS Diagram: Employee

File Edit View Window Help

Score Rankings Overlay: event

Cumulative Lift

Iteration Plot

Average Square Error

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	DFT	Total Degrees of Freedom	564			
event	DFE	Degrees of Freedom	463			
event	DFM	Model Degrees of Freedom	101			
event	NW	Number of Estimated Weights	101			
event	AIC	Akaike's Information Criterion	912.8829			
event	SBC	Schwarz's Bayesian Criterion	1350.723			
event	ASE	Average Squared Error	0.220765	0.234067		
event	MAX	Maximum Absolute Error	0.644101	0.875518		
event	DIV	Divisor for ASE	1128	1130		
event	NOBS	Sum of Frequencies	564	565		
event	RASE	Root Average Squared Error	0.469856	0.483804		
event	SSE	Sum of Squared Errors	249.0224	264.4952		
event	SUMW	Sum of Case Weights	1128	1130		
event	FPE	Final Prediction Error	0.317081			
event	MSL	Mean Squared Error	0.265923	0.234067		
event	RFPE	Root Final Prediction Error	0.405069			
event	RMSE	Root Mean Squared Error	0.518578	0.483804		
event	AVER	Average Error Function	0.630215	0.660831		
event	ERR	Error Rate	710.8529	745.6656		
event	MIC	Classification Rate	0.500475	0.407028		
event	WRONG	Number of Wrong Classes	205	230		

Output

```

1 User: 301259629
2 Date: December 14, 2022
3 Time: 20:54:53
4 -----
5 * Training Output
6 -----
7 -----
8 -----
9 -----
10 -----
11 -----
12 Variable Summary
13 -----
14 Measurement Level Frequency Count
15 Role -----

```

File Edit View Actions Options Window Help

Score

Hidden Units No

Residuals Yes

Standardization No

Status

Create Time 14/12/22 5:54 PM

Run ID e327161d-e60c-4

Last Error

Last Status Complete

Last Run Time 14/12/22 8:54 PM

Network

Open a dialog to customize network options.

Type here to search

I.T. 2022-1... Balance... Enterprise... Results... Show all

ENG US 8:55 PM 12/14/2022

Data.csv

Windows Taskbar

Results - Node: fw NNS Diagram: Employee

File Edit View Window Help

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
event	DFT	Total Degrees of Freedom	564			
event	DFE	Degrees of Freedom	463			
event	DFM	Model Degrees of Freedom	101			
event	NW	Number of Estimated Weights	101			
event	AIC	Akaike's Information Criterion	912.8829			
event	SBC	Schwarz's Bayesian Criterion	1350.723			
event	ASE	Average Squared Error	0.220765	0.234067		
event	MAX	Maximum Absolute Error	0.644101	0.875518		
event	DIV	Divisor for ASE	1128	1130		
event	NOBS	Sum of Frequencies	564	565		
event	RASE	Root Average Squared Error	0.469856	0.483804		
event	SSE	Sum of Squared Errors	249.0224	264.4952		
event	SUMW	Sum of Case Weights	1128	1130		
event	FPE	Final Prediction Error	0.317081			
event	MSL	Mean Squared Error	0.265923	0.234067		
event	RFPE	Root Final Prediction Error	0.405069			
event	RMSE	Root Mean Squared Error	0.518578	0.483804		
event	AVER	Average Error Function	0.630215	0.660831		
event	ERR	Error Rate	710.8529	745.6656		
event	MIC	Classification Rate	0.500475	0.407028		
event	WRONG	Number of Wrong Classes	205	230		

File Edit View Actions Options Window Help

Score

Hidden Units No

Residuals Yes

Standardization No

Status

Create Time 14/12/22 5:54 PM

Run ID e327161d-e60c-4

Last Error

Last Status Complete

Last Run Time 14/12/22 8:54 PM

Network

Open a dialog to customize network options.

Type here to search

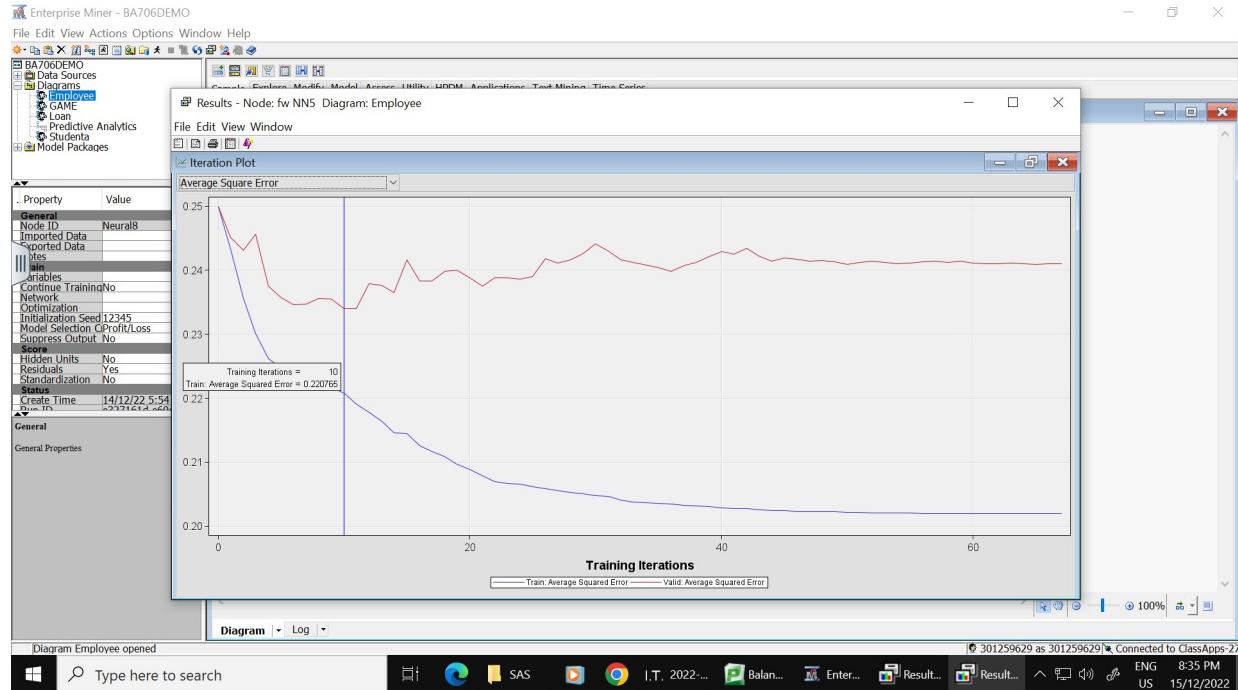
I.T. 2022-1... Balance... Enterprise... Results... Show all

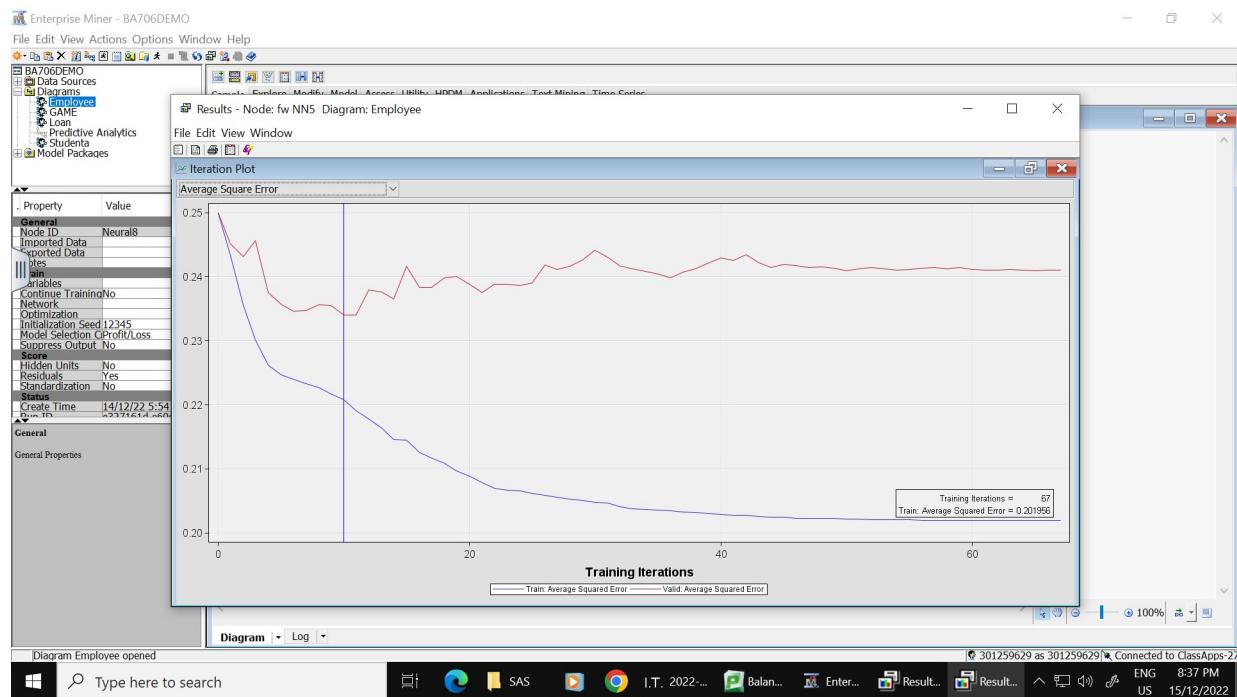
ENG US 8:56 PM 12/14/2022

Data.csv

Windows Taskbar

Here, we see that the average squared error of the training and validation data sets start off in the same direction of improvement till iteration 10 where the validation ASE starts to increase as a sign of overfitting. This occurs until the point of convergence at iteration 67.

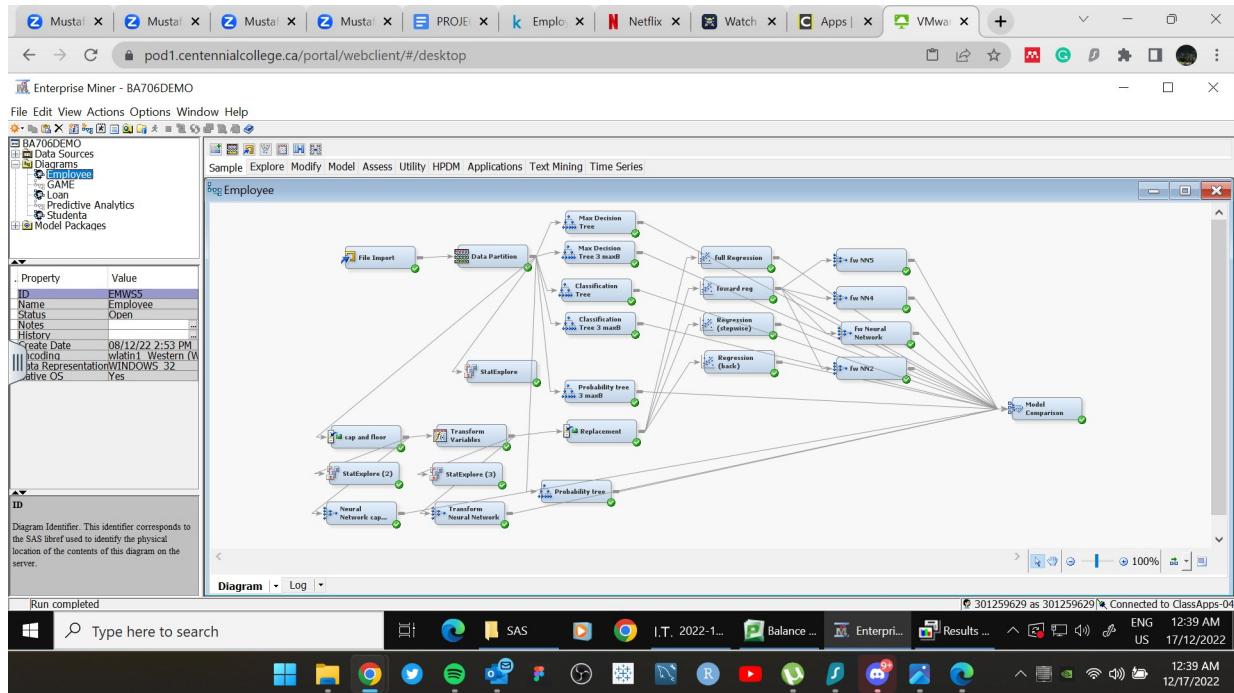




# ASSESSMENT

## MODEL COMPARISON

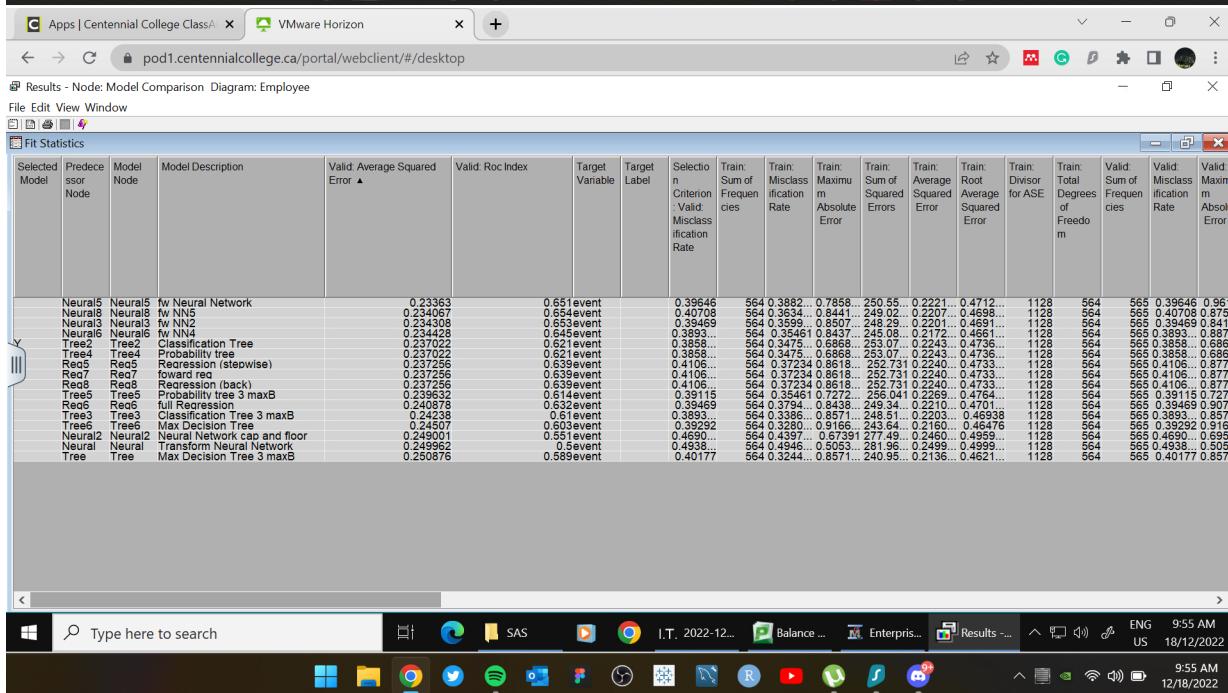
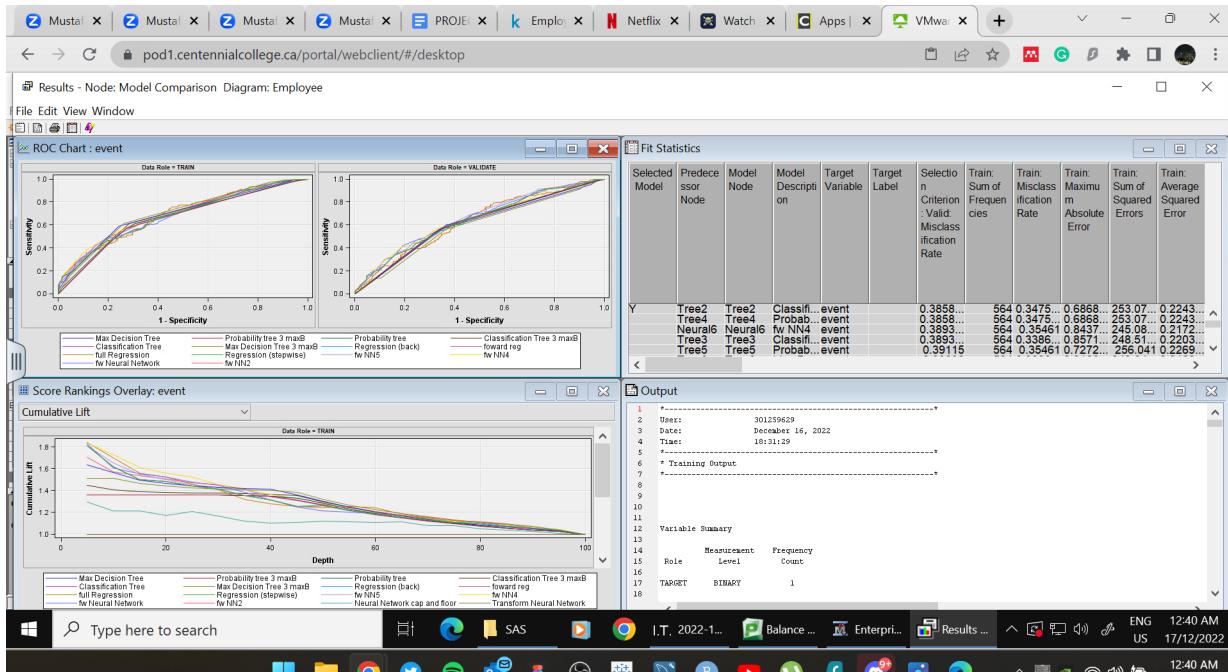
Finally, a model comparison node was introduced to the diagram workspace to which the different decision trees, regressions, and neural network models were connected. The results show a comparison of all models using several statistical tests. The below shows the final diagram of all the models we tested in this project:

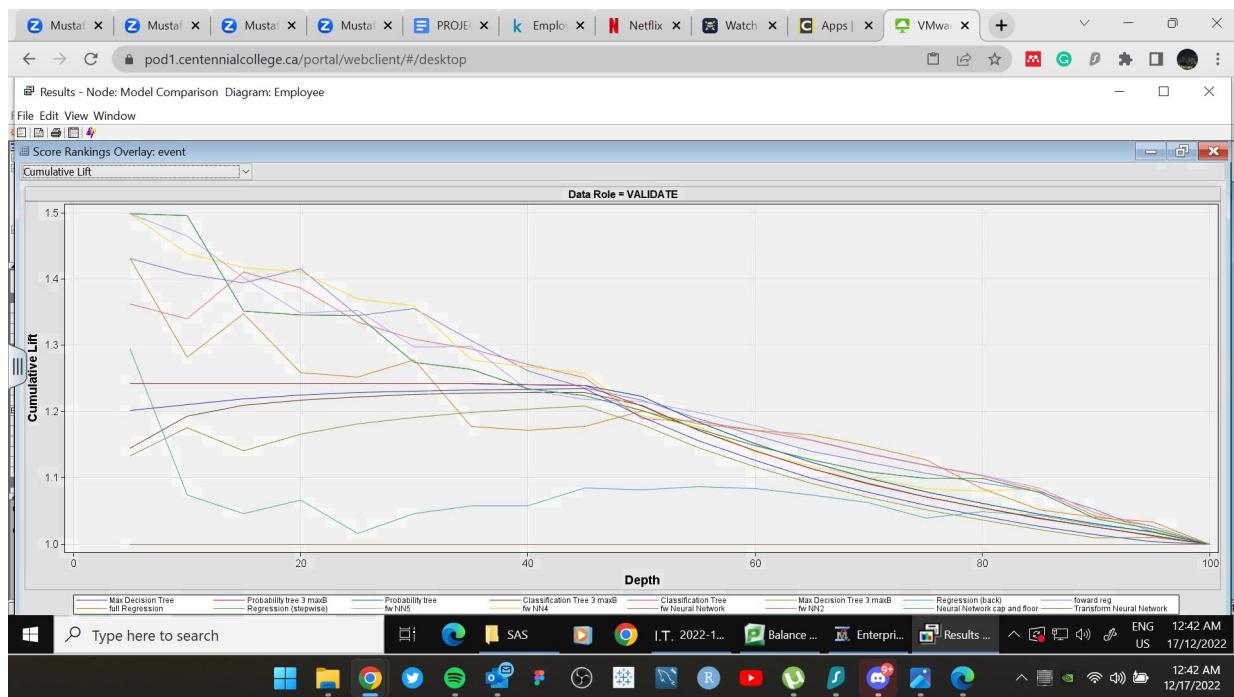


As stated earlier, we will be selecting the most fit model using the Average squared error. Below shows the summary of all models the Average squared error and ROC index. Using the ROC index as a selection criterion, the best model would be the Neural Network with 5 hidden units and ROC index of 0.654. However, our chosen method of selection is the Average squared error hence the Neural Network with 3 hidden units is the best model.

Model Type	Average Squared Error	ROC
<b>Neural Network 3 hidden units</b>	<b>0.233630</b>	0.651

Neural Network 5 hidden units	0.234067	0.654
Neural Network 2 hidden units	0.234308	0.653
Neural Network 4 hidden units	0.234428	0.645
Classification Tree 2 branch	0.237022	0.621
Probability Tree 2 branch	0.237022	0.621
Regression Stepwise	0.237256	0.639
Regression Forward	0.237256	0.639
Regression Backward	0.237256	0.639
Probability Tree 3 branch	0.239632	0.614
Regression Full	0.240878	0.632
Classification Tree 3 branch	0.242380	0.610
Maximal Decision Tree	0.245070	0.603





## CONCLUSION

The best model is the Neural Network with 3 hidden units which has a validation Average Square Error of 0.23363. This neural network model was derived from attaching a neural network node to the forward regression model which had the best ASE of 0.237256. However, because of the complexity of how the Neural network arrives at its result, we are unable to explain in detail the next steps to take to reduce attrition in the different industries using this model. For this reason, we will be using the next best model which is easy to interpret to make suggestions for improvement. From the comparison table, the next best models which are not neural network models are the classification tree and probability tree with two branches. They both have a Validation Average Square Error of 0.237022. Therefore, we will make our recommendations based on these trees. We will also use the forward regression model results to support our recommendations.

For all the trees used in this project, the first split occurred in the industry variable. This variable had the highest log worth and was considered the most important variable by the decision trees and interestingly by most of the regression models too. Other variables considered by the trees are Profession, Stag (duration of experience), and Way (means of transportation to the workplace).

The trees show that 60% of employees in Banks, Consult and State industries are likely to quit their jobs with 62.89% of these employees requiring a vehicle to get to work. Employees who walk to the office (most likely because they stay in close proximity) are less likely to quit their jobs. 43.99% of the employees in the power generation, and retail industries are likely to quit with a higher percentage than those in the commercial, sales, and finance professions.

The forward regression model used the transport industry as a means of comparison for all other industries. It appears to be the most stable industry that is able to retain its employees after the IT industry with the least likelihood of attrition. Employees in HR and IT profession when compared with sales are also less likely to quit their jobs.

Below are the insights obtained from the selected models and our recommendations:

1. **Banks, Consult, and State Industry:** With 60% of employees leaving their jobs, a deep dive could show specific reasons for attrition. From the forward regression model, those in

banks, consult, and state are 4.6, 4.4 and 3.1 times respectively more likely to quit their jobs when compared with those in the transport industry. There might also be a need to study the transport industry to see why it seems more stable and able to retain its employees.

We suggest a drive for work-life balance is implemented in these industries. Incentives should be given for working longer than usual work hours, compensation given for achieving goals far and beyond expectations, and recognition for assignments well executed. A program for continued benefits after retirement should also be introduced. All these might be useful in encouraging professionals within these industries to stay on the job for a longer period or till retirement.

2. **Employees who come to work by bus or car:** Proximity to the office location also appears to be an important factor from both the tree models and regression models. People who have to get to work with a vehicle are 2.5 times more likely to quit their jobs. We advise that attention is given to how employees get to work. A transportation initiative can be introduced whereby employees have a choice of joining a staff bus that does a pick-up and drop-off for staff who stay a distance from the workplace. This will reduce the stress of driving to work or joining the public transportation system on a daily basis. In addition, the employees can be given the option of working remotely on some days in the week. This option has become a widely accepted work arrangement in the past two years which has improved the work-life balance of employees.
  
3. **Commercial, Sales and Finance Profession:** From the regression models, we see that employees in all other professions including sales are more likely to quit their jobs when compared with HR and IT professionals. We would say carrying out an investigation on the two professions might give an insight as to the reason for their stability. In the mean time, suggestions made for the industries above will also be profitable when applied to the profession variable. In addition, we suggest significance should be attached to attaining specific levels of achievement. This can be an incentive for employees to spend more time on their jobs while they pursue higher levels in their career paths.

## APPROVAL OF DATASET

Group 9 Applied Analytic Modeling

IT Ibiyengha Tobin  
To: David Parent  
Cc: Ibukunoluwa Olukoko; Emmanuella Adegbenjo

turnover.csv 81 KB

Good Day David,  
This is Group 9 for Applied Analytic Modeling.

Please you already approved our dataset 'Employee Turnover' in class which we got from kaggle.com.  
We have attached this dataset for your written approval.  
Thank You.

Best Regards,  
Group 9.

Reply Reply all Forward

## REFERENCE

<https://www.kaggle.com/datasets/davinwijaya/employee-turnover>