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 resignfrom 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, youjs,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)

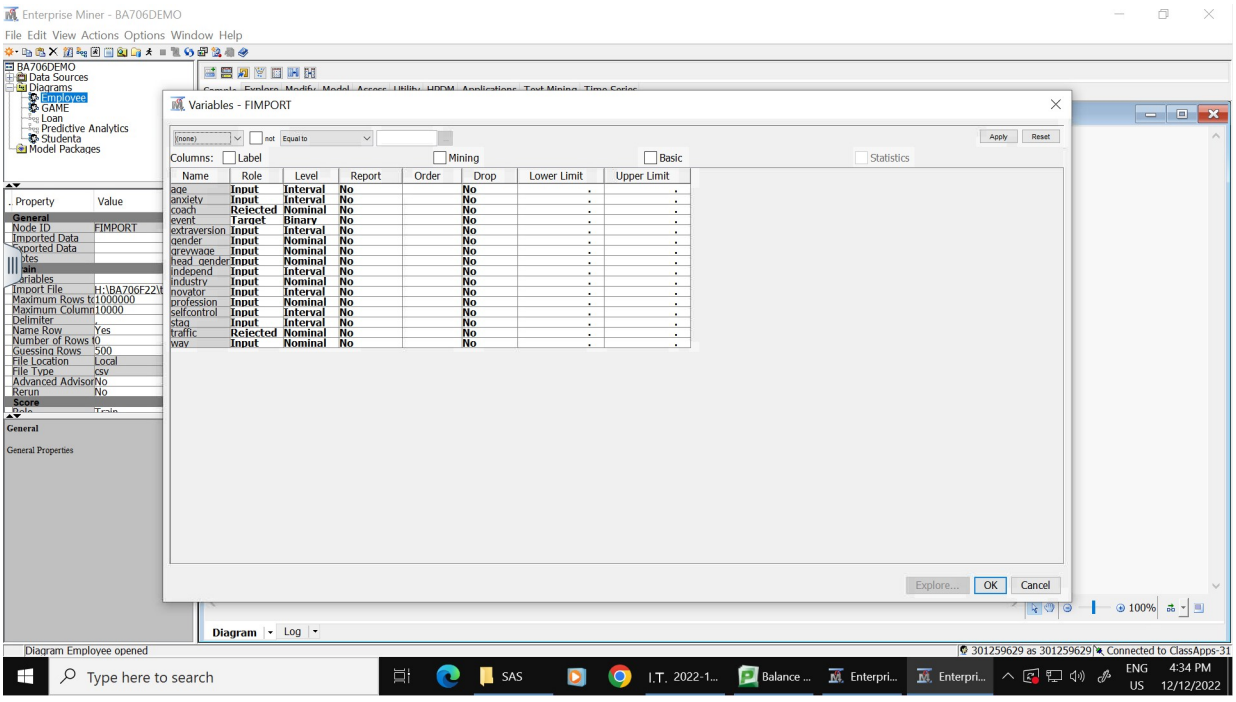
Description of the Traffic and Greywage variables

TRAFFIC	*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.*
GREYWAGE	***Grey-wage**: in Russia or Ukraine means that the employer pays just a tiny bit amount of salary above the white-wage* ***White wage:** minimum wage*

DATA SETUP/EXPLORATION

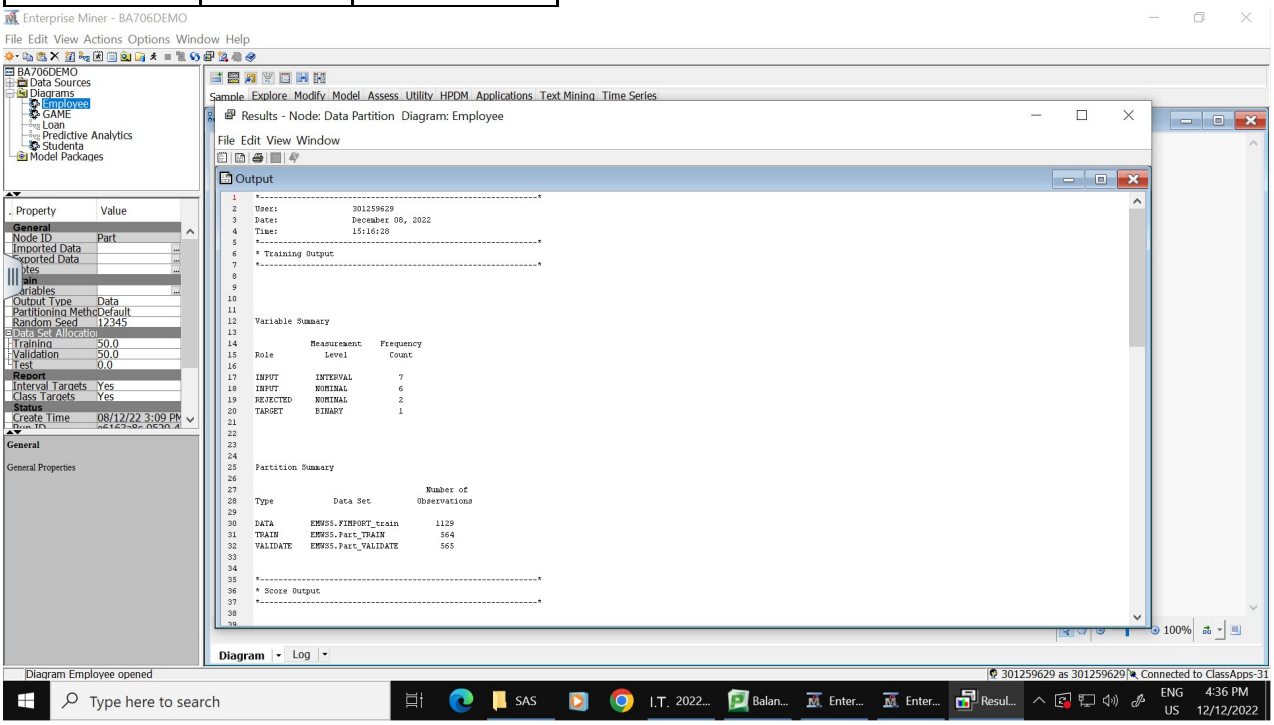
We first imported the **employee dataset** and then made **'event'** our target variable as it is a binary variable that indicates whether an employee resigned from the company or not.

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.



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



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:

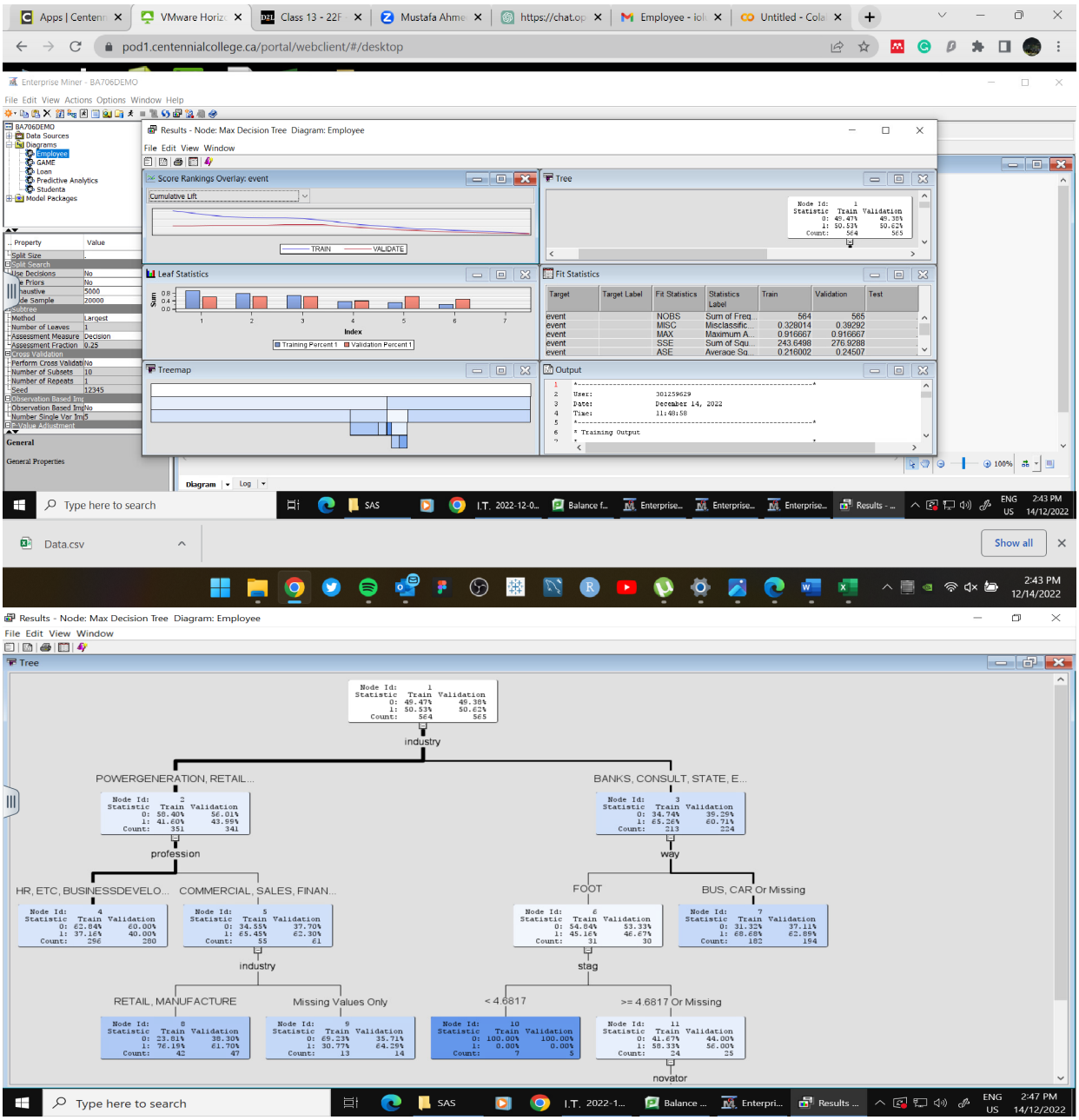
1. Maximal Trees(Two and Three Branch)

2. Classification Trees(Two and Three Branch)

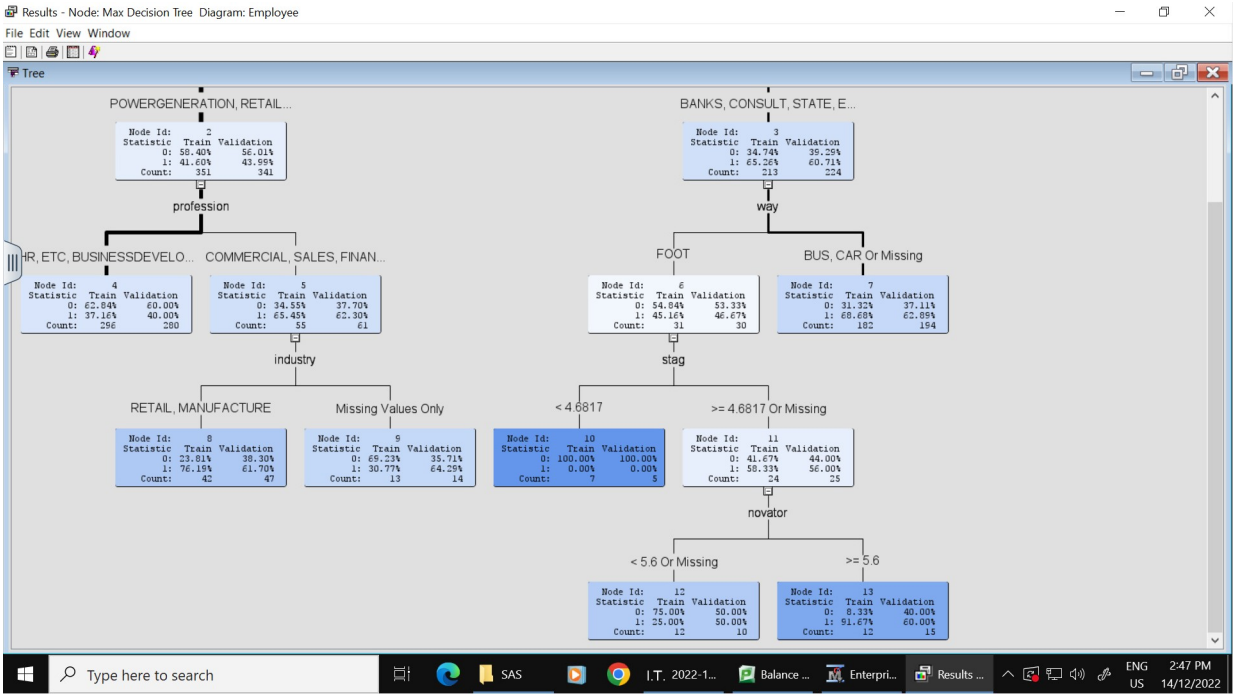
3. Probability Trees(Two and Three Branch)

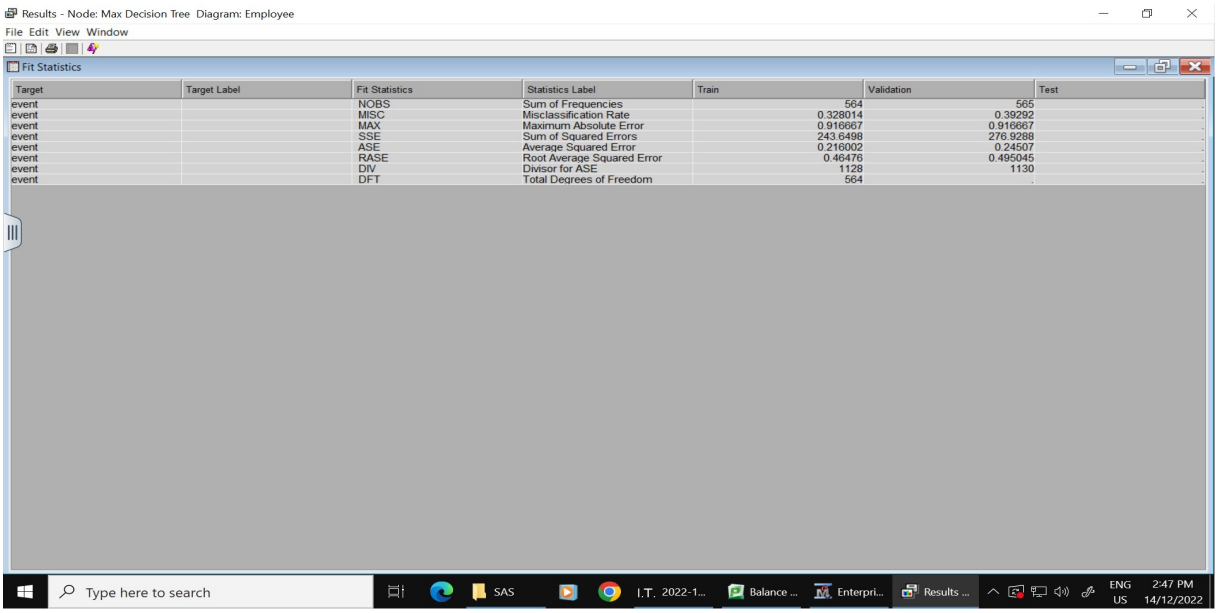
TWO-BRANCH MAXIMAL TREE

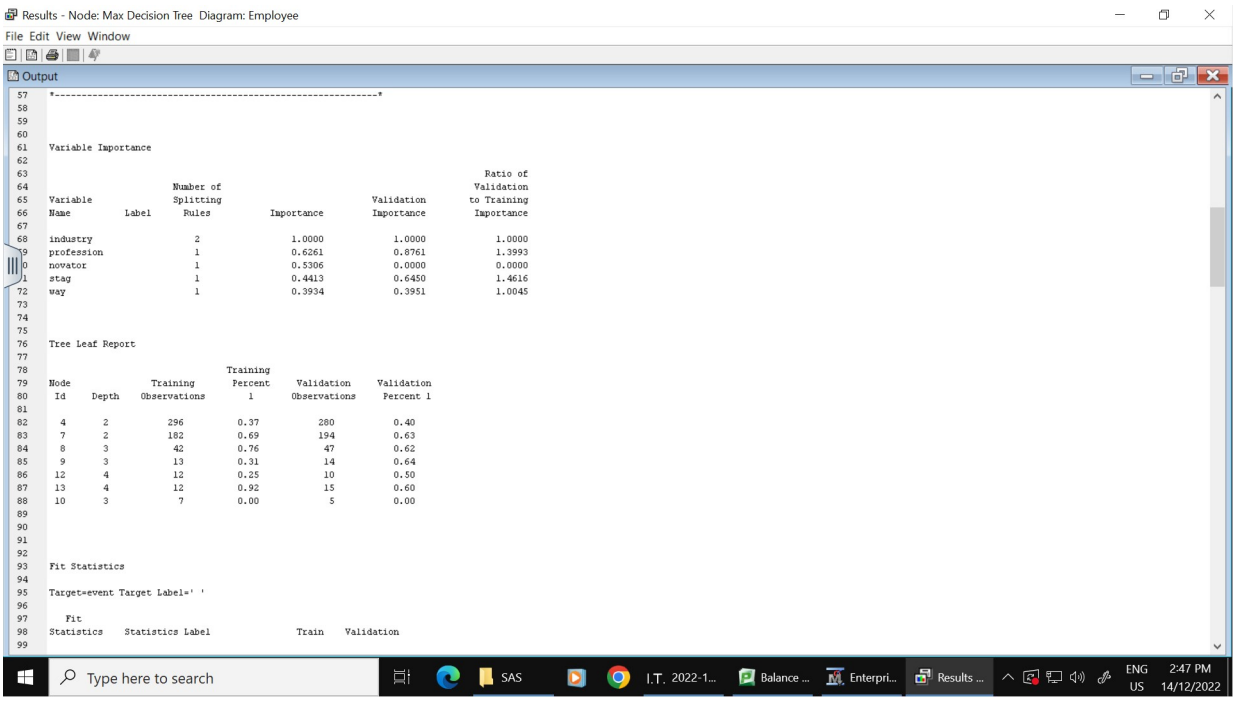
The decision tree node was attached to the data partition node with a maximum of two branches indicated, ‘largest’ used as the method and ‘decision’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.


pod1.centennialcollege.ca/portal/webclient/#/desktop
Enterprise Miner - BA706DEMO
Results - Node: Max Decision Tree Diagram: Employee
Score Rankings Overlay: event
Cumulative Lift
TRAIN
VALIDATE
Tree
Leaf Statistics
Training Percent 1
Validation Percent 1
Fit Statistics
Treemap
Output
Data.csv
Show all
Type here to search
2:43 PM
12/14/2022

Results - Node: Max Decision Tree Diagram: Employee
Tree
industry
POWERGENERATION, RETAIL...
BANKS, CONSULT, STATE, E...
profession
way
HR, ETC, BUSINESSDEVELO...
COMMERCIAL, SALES, FINAN...
FOOT
BUS, CAR Or Missing
industry
stag
RETAIL, MANUFACTURE
Missing Values Only
< 4.6817
>= 4.6817 Or Missing
novator
Type here to search
2:47 PM
14/12/2022


Results - Node: Max Decision Tree Diagram: Employee
File Edit View Window
Tree
POWERGENERATION, RETAIL...
Node Id: 2
Statistic Train Validation
0: 58.40% 56.01%
1: 41.60% 43.99%
Count: 351 341
profession
HR, ETC, BUSINESSDEVELO...
Node Id: 4
Statistic Train Validation
0: 62.84% 60.00%
1: 37.16% 40.00%
Count: 256 280
COMMERCIAL, SALES, FINAN...
Node Id: 5
Statistic Train Validation
0: 34.55% 37.70%
1: 65.45% 62.30%
Count: 55 61
industry
RETAIL, MANUFACTURE
Node Id: 8
Statistic Train Validation
0: 23.81% 38.30%
1: 76.19% 61.70%
Count: 42 47
Missing Values Only
Node Id: 9
Statistic Train Validation
0: 69.23% 35.71%
1: 30.77% 64.29%
Count: 13 14
BANKS, CONSULT, STATE, E...
Node Id: 3
Statistic Train Validation
0: 34.74% 39.29%
1: 65.26% 60.71%
Count: 213 224
way
FOOT
Node Id: 6
Statistic Train Validation
0: 54.84% 53.33%
1: 45.16% 46.67%
Count: 31 30
stag
< 4.6817
Node Id: 10
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 7 5
>= 4.6817 Or Missing
Node Id: 11
Statistic Train Validation
0: 41.67% 44.00%
1: 58.33% 56.00%
Count: 24 25
novator
< 5.6 Or Missing
Node Id: 12
Statistic Train Validation
0: 75.00% 50.00%
1: 25.00% 50.00%
Count: 12 10
>= 5.6
Node Id: 13
Statistic Train Validation
0: 8.33% 40.00%
1: 91.67% 60.00%
Count: 12 15
BUS, CAR Or Missing
Node Id: 7
Statistic Train Validation
0: 31.32% 37.11%
1: 68.68% 62.89%
Count: 182 194
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG
US
2:47 PM
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 | | MISC | Misclassification Rate | 0.328014 | 0.39292 |
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 | 0.24507 |
event | | RASE | Root Average Squared Error | 0.46476 | 0.495045 |
event | | DIV | Divisor for ASE | 1128 | 1130 |
event | | DFT | Total Degrees of Freedom | 564 | |
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG
US
2:47 PM
14/12/2022

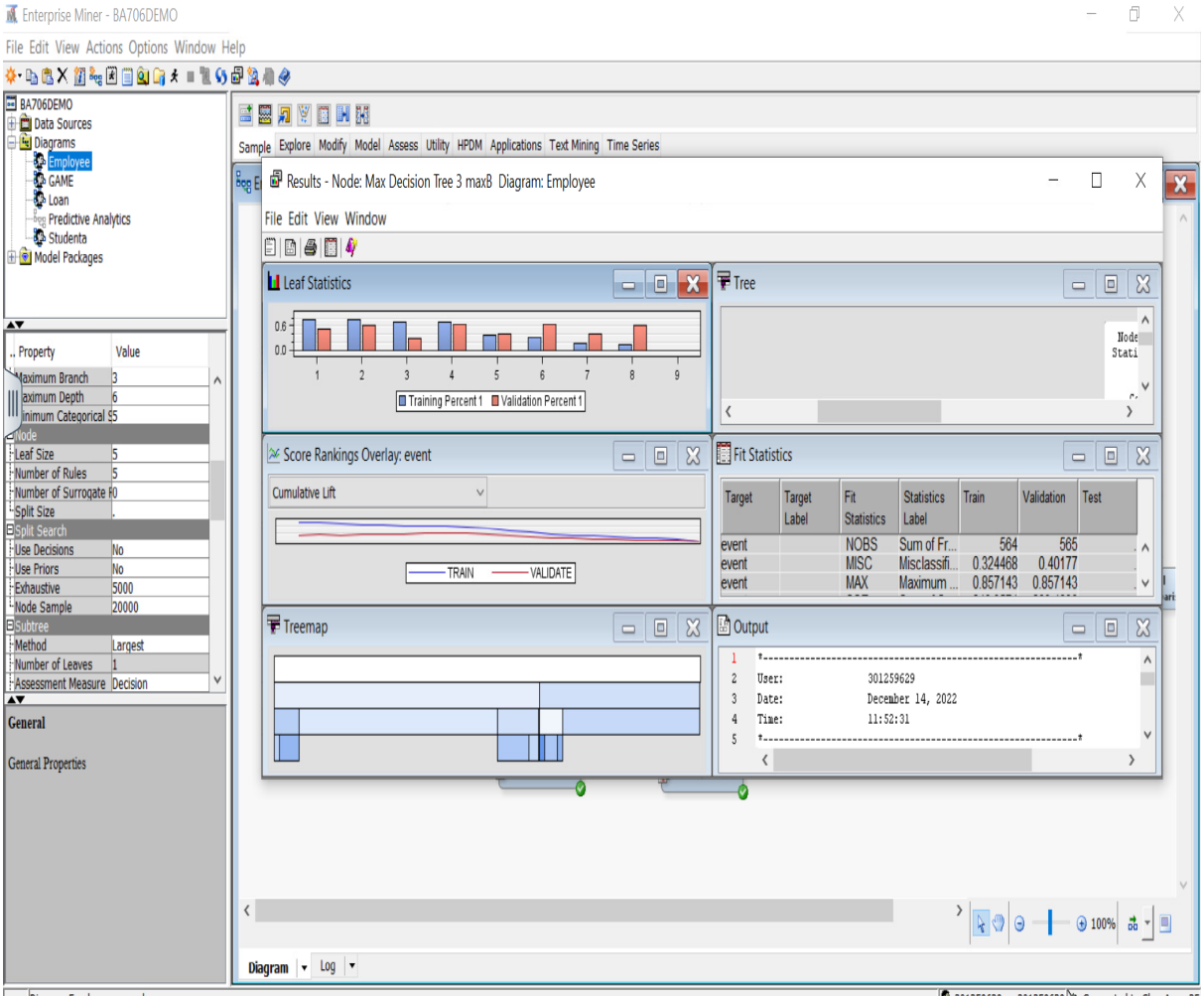


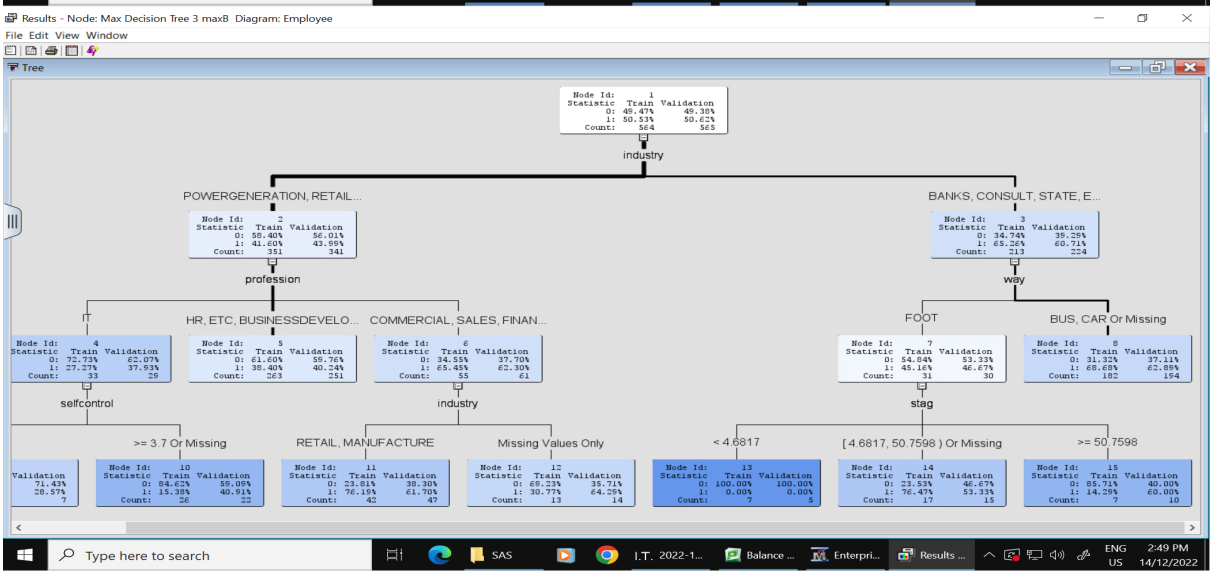
The tree had seven leaves with five variable splits. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit 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.

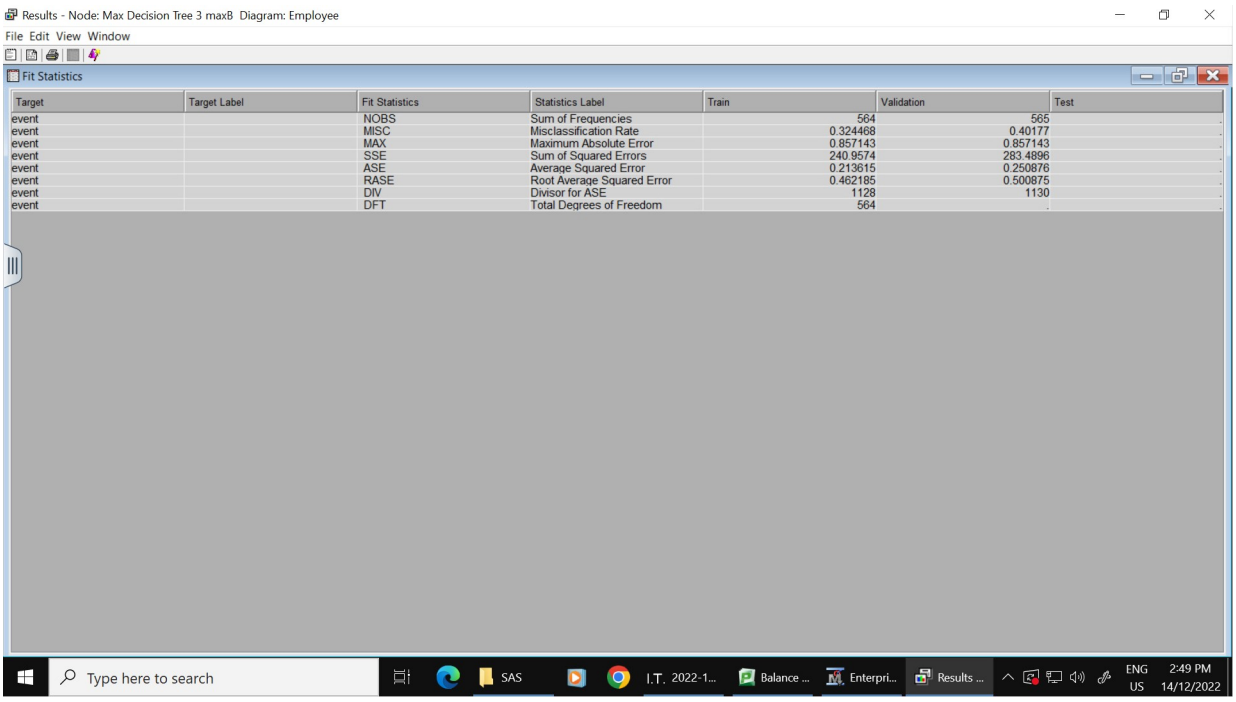
The Validation Average Square Error for the two-branch Maximal Tree is 0.24507.

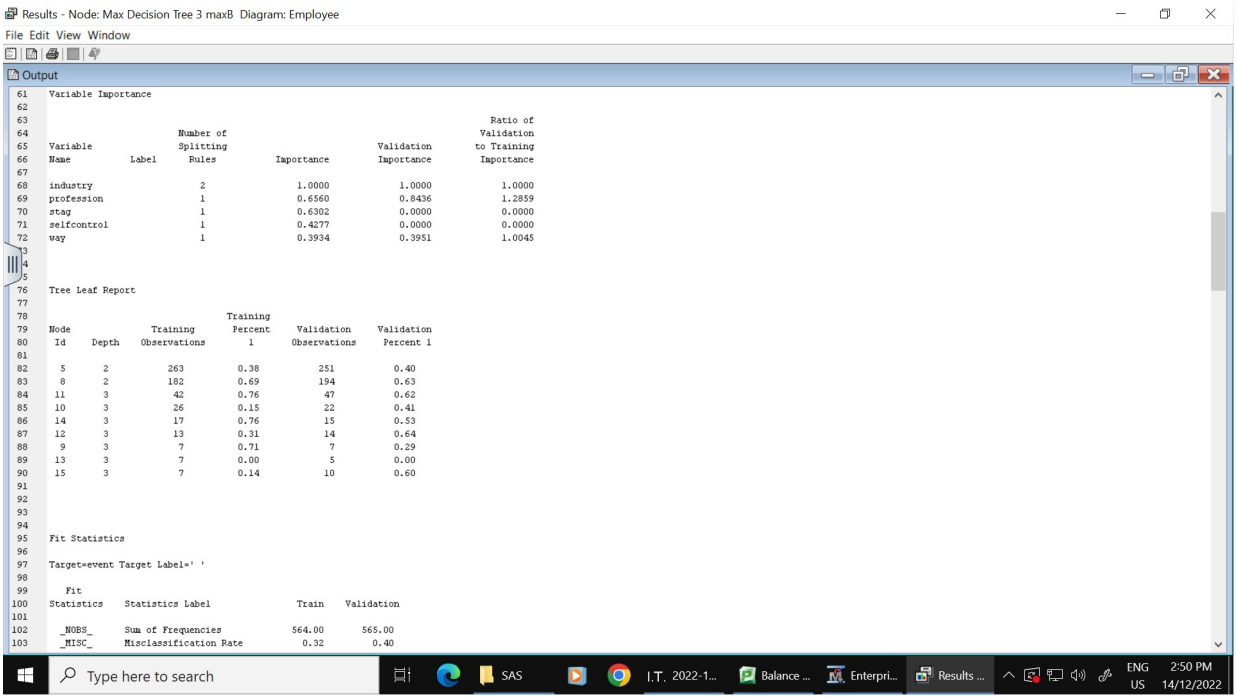
THREE-BRANCH MAXIMAL TREE

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
Studenta
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Results - Node: Max Decision Tree 3 maxB Diagram: Employee
File Edit View Window
Leaf Statistics
Training Percent 1
Validation Percent 1
Tree
Score Rankings Overlay: event
Cumulative Lift
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event NOBS Sum of Fr... 564 565
event MISC Misclassifi... 0.324468 0.40177
event MAX Maximum ... 0.857143 0.857143
Treemap
Output
User: 301259629
Date: December 14, 2022
Time: 11:52:31
Property
Value
Maximum Branch 3
Maximum Depth 6
Minimum Categorical Size 5
Node
Leaf Size 5
Number of Rules 5
Number of Surrogate Rules 0
Split Size
Split Search
Use Decisions No
Use Priors No
Exhaustive 5000
Node Sample 20000
Subtree
Method Largest
Number of Leaves 1
Assessment Measure Decision
General
General Properties
Diagram
Log
Diagram Employee opened
Connected to ClassApps-25
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG US
2:49 PM
14/12/2022


Results - Node: Max Decision Tree 3 maxB Diagram: Employee
File Edit View Window
Tree
industry
POWERGENERATION, RETAIL...
BANKS, CONSULT, STATE, E...
profession
way
IT
HR, ETC, BUSINESSDEVELO...
COMMERCIAL, SALES, FINAN...
FOOT
BUS, CAR Or Missing
selfcontrol
industry
stag
>= 3.7 Or Missing
RETAIL, MANUFACTURE
Missing Values Only
< 4.6817
[4.6817, 50.7598) Or Missing
>= 50.7598
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG US
2:49 PM
14/12/2022



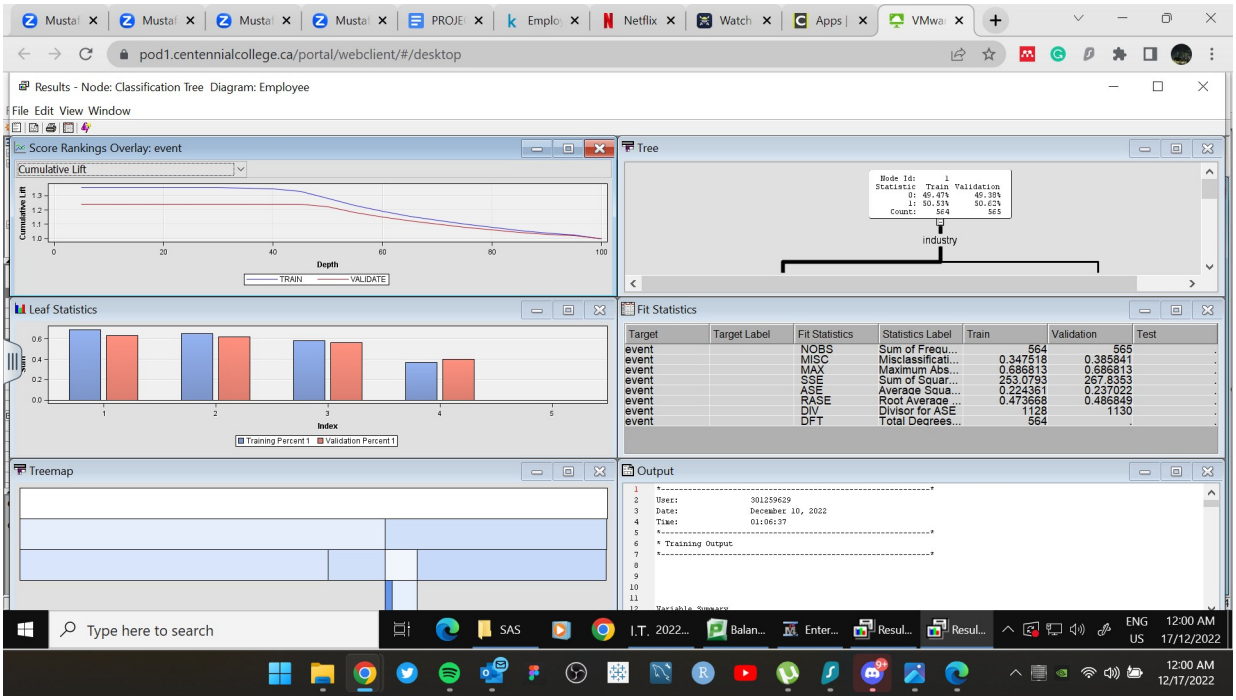


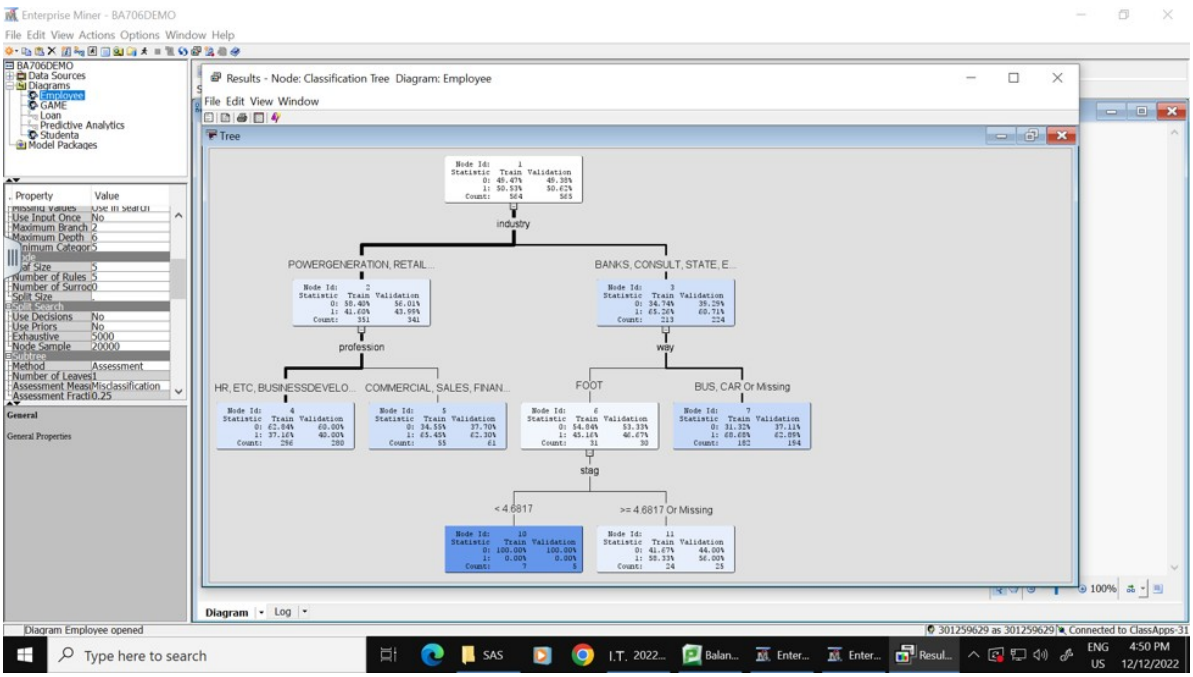
The tree had nine leaves with five variable splits. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their 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.

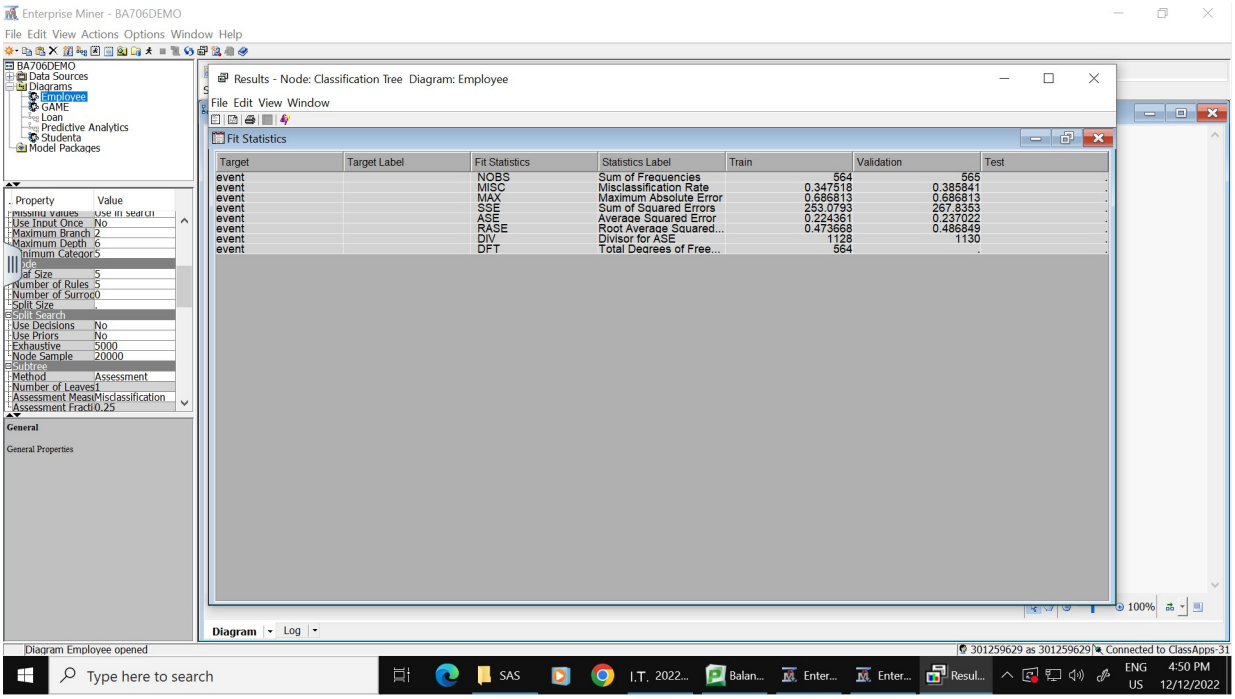
The Validation Average Square Error for the three-branch Maximal Tree is 0.250876. We observed that the Two Branch Maximal Tree is a better model than the Three Branch Maximal tree, as this has a higher Validation Average Square Error.

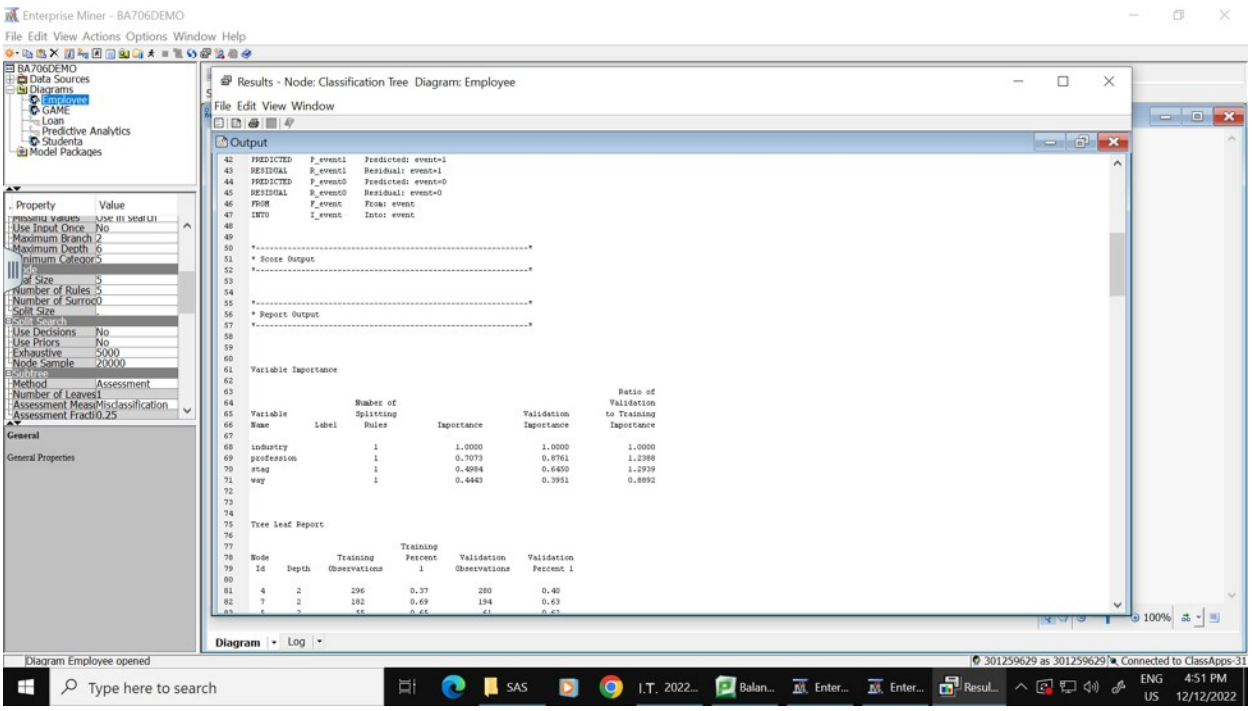
TWO-BRANCH CLASSIFICATION TREE

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.




Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Results - Node: Classification Tree Diagram: Employee
File Edit View Window
Tree
industry
POWERGENERATION, RETAIL...
BANKS, CONSULT, STATE, E...
profession
way
HR, ETC, BUSINESSDEVELO...
COMMERCIAL, SALES, FINAN...
FOOT
BUS, CAR Or Missing
stag
< 4.6817
>= 4.6817 Or Missing
Diagram
Log
Diagram Employee opened
Type here to search


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
Results - Node: Classification 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 | | MISC | Misclassification Rate | 0.347518 | 0.385841 |
event | | MAX | Maximum Absolute Error | 0.686813 | 0.686813 |
event | | SSE | Sum of Squared Errors | 253.0793 | 267.8353 |
event | | ASE | Average Squared Error | 0.224361 | 0.237022 |
event | | RASE | Root Average Squared... | 0.473668 | 0.486849 |
event | | DIV | Divisor for ASE | 1128 | 1130 |
event | | DFT | Total Degrees of Free... | 564 | |
Diagram
Log
Diagram Employee opened
Type here to search

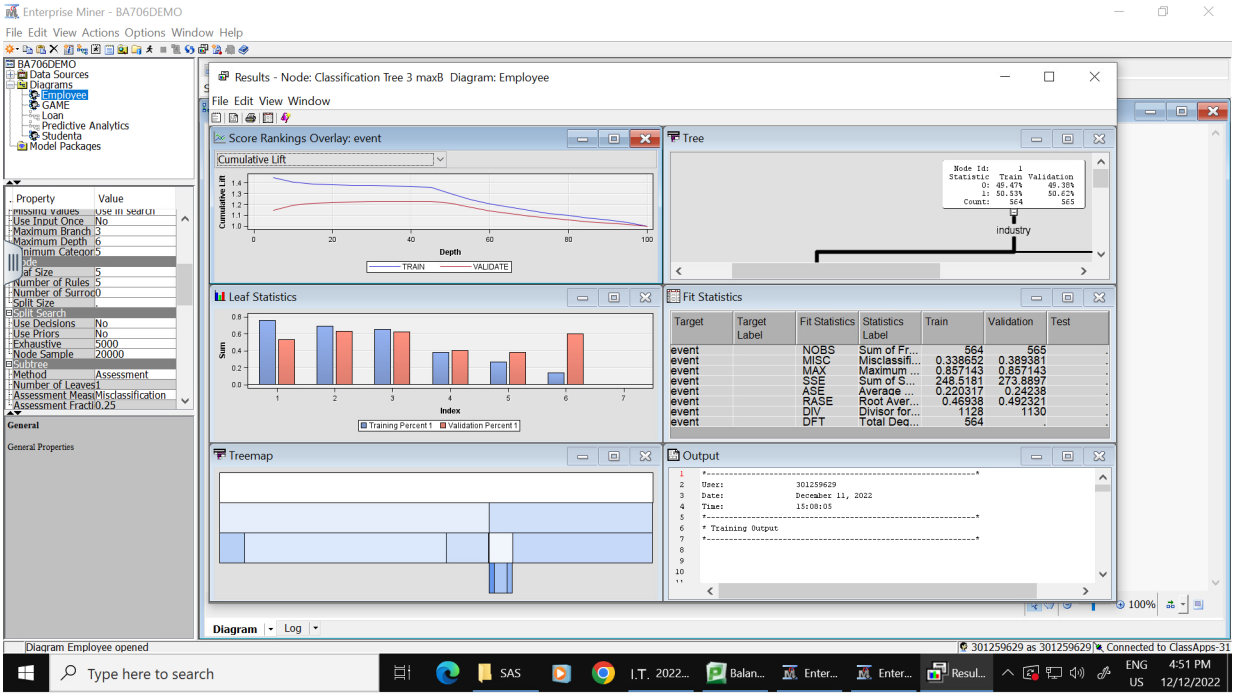


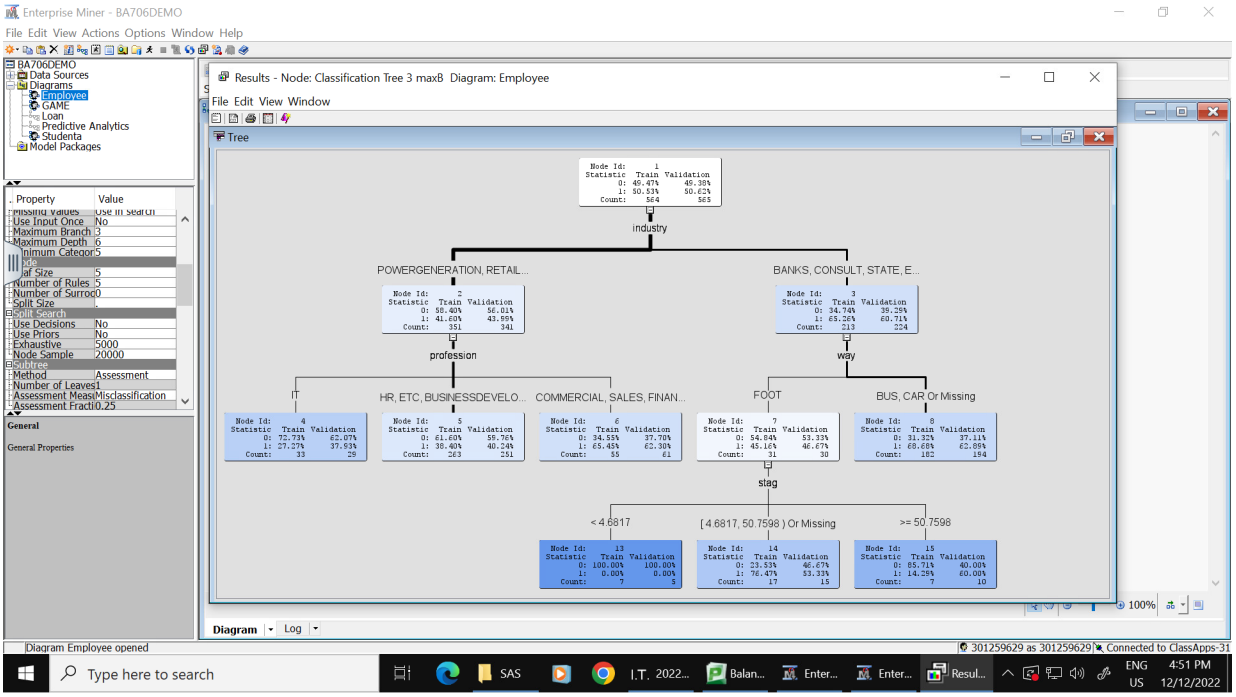
The tree has five leaves with four variable splits. This reduction is a result of pruning using the misclassification measure. The first split occurs at the Industry variable with the highest logworth when compared to the other variables selected in the model. The tree shows that 62.89% of employees 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.

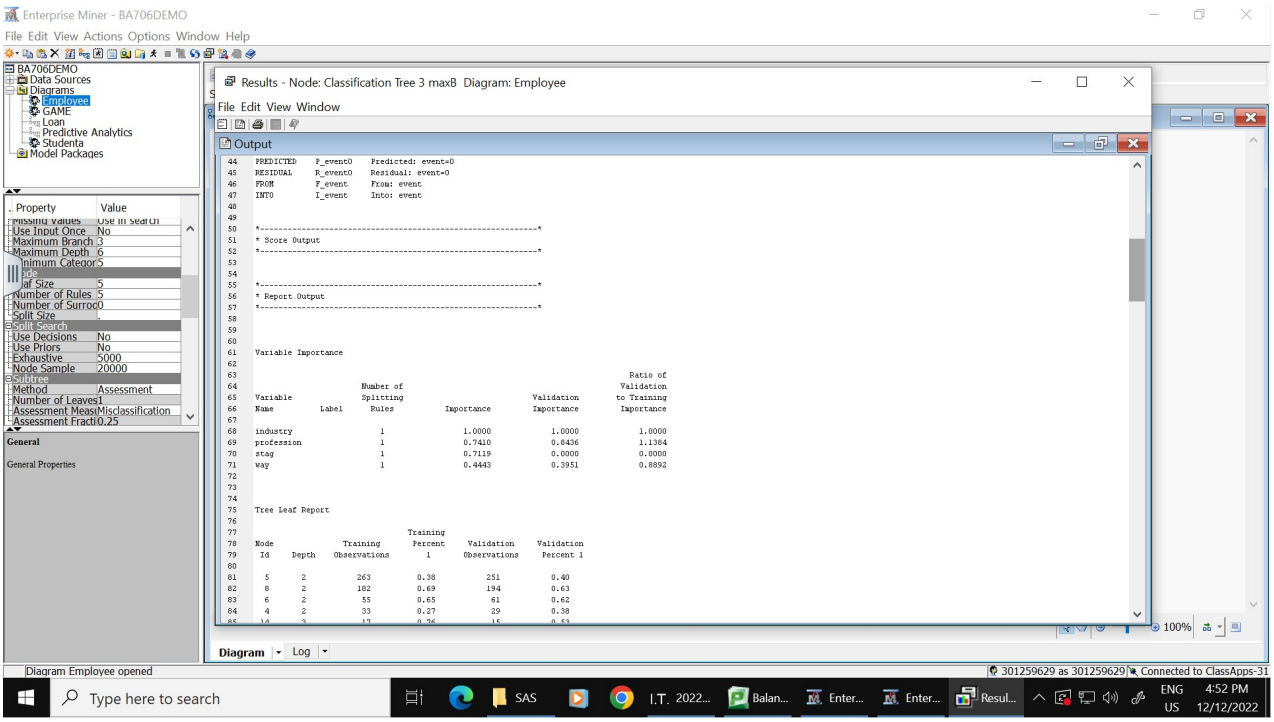
The Validation Average Square Error for the two-branch classification Tree is 0.237022.

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.


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property
Value
Results - Node: Classification Tree 3 maxB Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Tree
industry
Leaf Statistics
Training Percent 1
Validation Percent 1
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event
NOBS
Sum of Fr...
564
565
event
MISC
Misclassifi...
0.338652
0.389381
event
MAX
Maximum ...
0.857143
0.857143
event
SSE
Sum of S...
248.5181
273.8897
event
ASE
Average ...
0.220317
0.24238
event
RASE
Root Aver...
0.46938
0.492321
event
DIV
Divisor for...
1128
1130
event
DFT
Total Deg...
564
Treemap
Output
Diagram
Log
Diagram Employee opened
301259629 as 301259629
Connected to ClassApps-31
Type here to search
SAS
I.T. 2022...
Balan...
Enter...
Enter...
Resul...
ENG
US
4:51 PM
12/12/2022


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property
Value
Results - Node: Classification Tree 3 maxB Diagram: Employee
File Edit View Window
Tree
industry
POWERGENERATION, RETAIL...
BANKS, CONSULT, STATE, E...
profession
way
IT
HR, ETC, BUSINESSDEVELO...
COMMERCIAL, SALES, FINAN...
FOOT
BUS, CAR Or Missing
stag
< 4.6817
[4.6817, 50.7598) Or Missing
>= 50.7598
Diagram
Log
Diagram Employee opened
301259629 as 301259629
Connected to ClassApps-31
Type here to search
SAS
I.T. 2022...
Balan...
Enter...
Enter...
Resul...
ENG
US
4:51 PM
12/12/2022



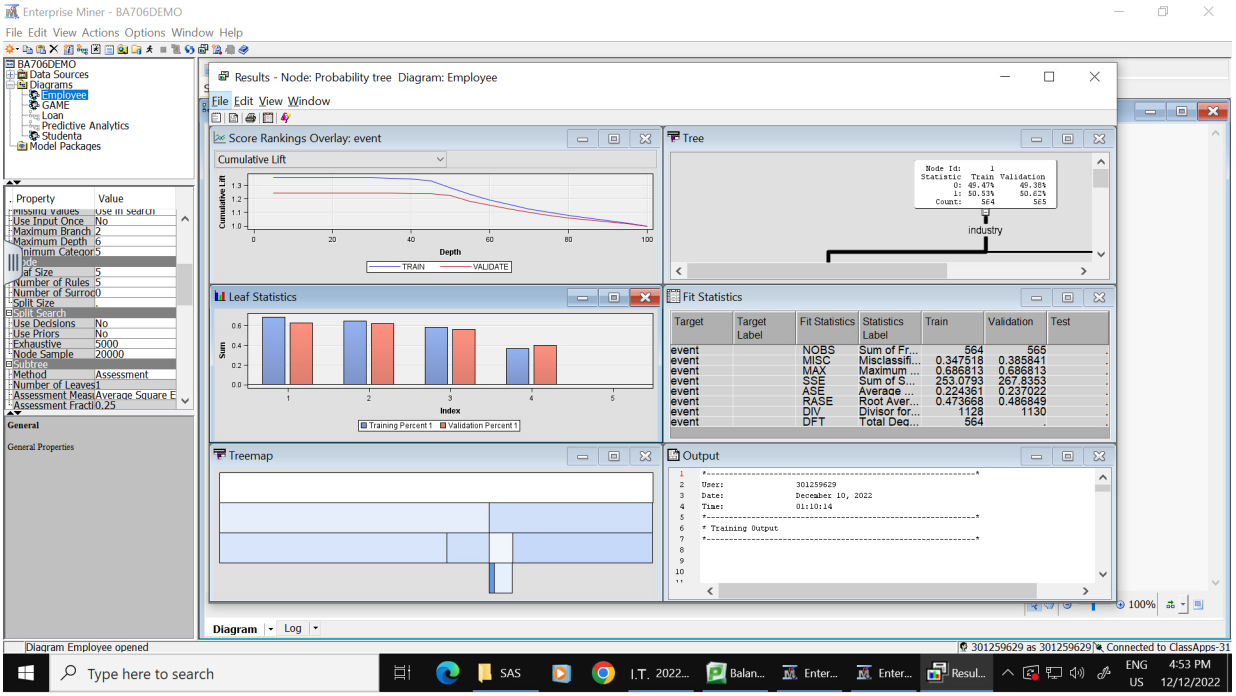
The tree has seven leaves with four variable splits. The increased number of branches also increased the number of leaves when compared to the two branch model. The tree shows that 62.89% of employees in Banks, Consult and State who go to work in vehicles are likely to quit their jobs. 60% of employees 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.

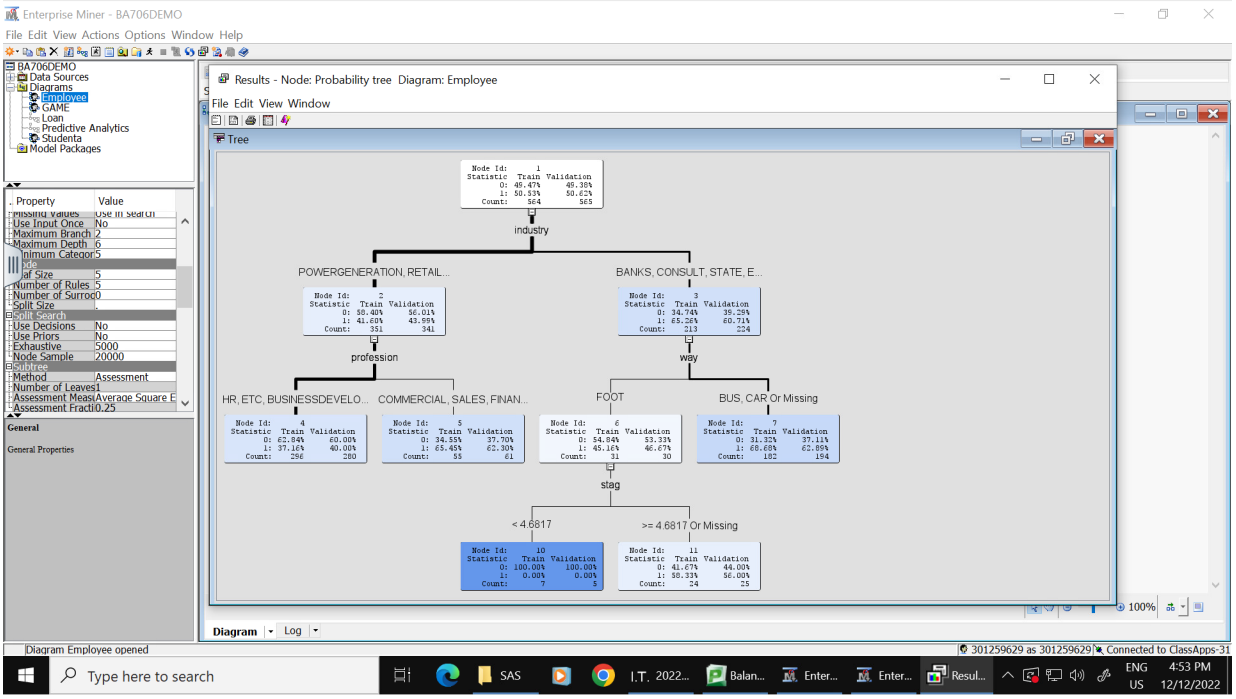
The Validation Average Square Error for the three-branch classification Tree is 0.24238. The Two Branch Classification Tree is a better model than the Three Branch Classification tree, as it has a lower Validation Average Square Error.

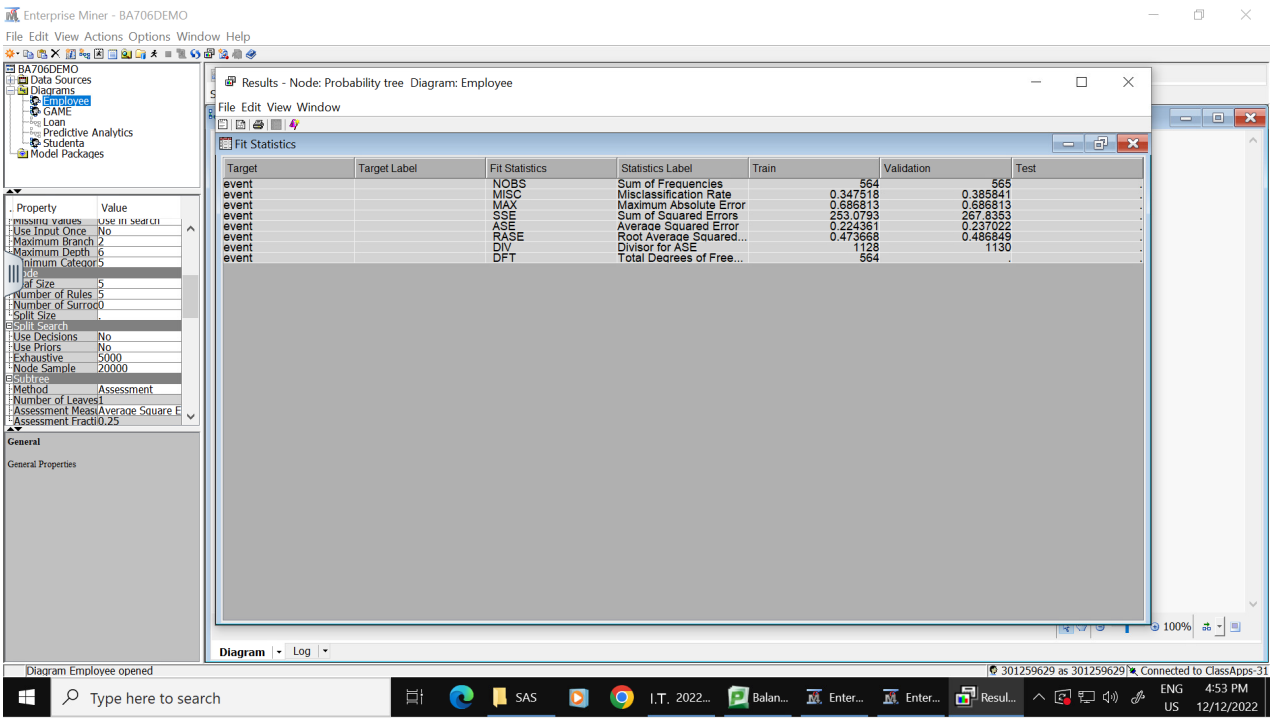
.

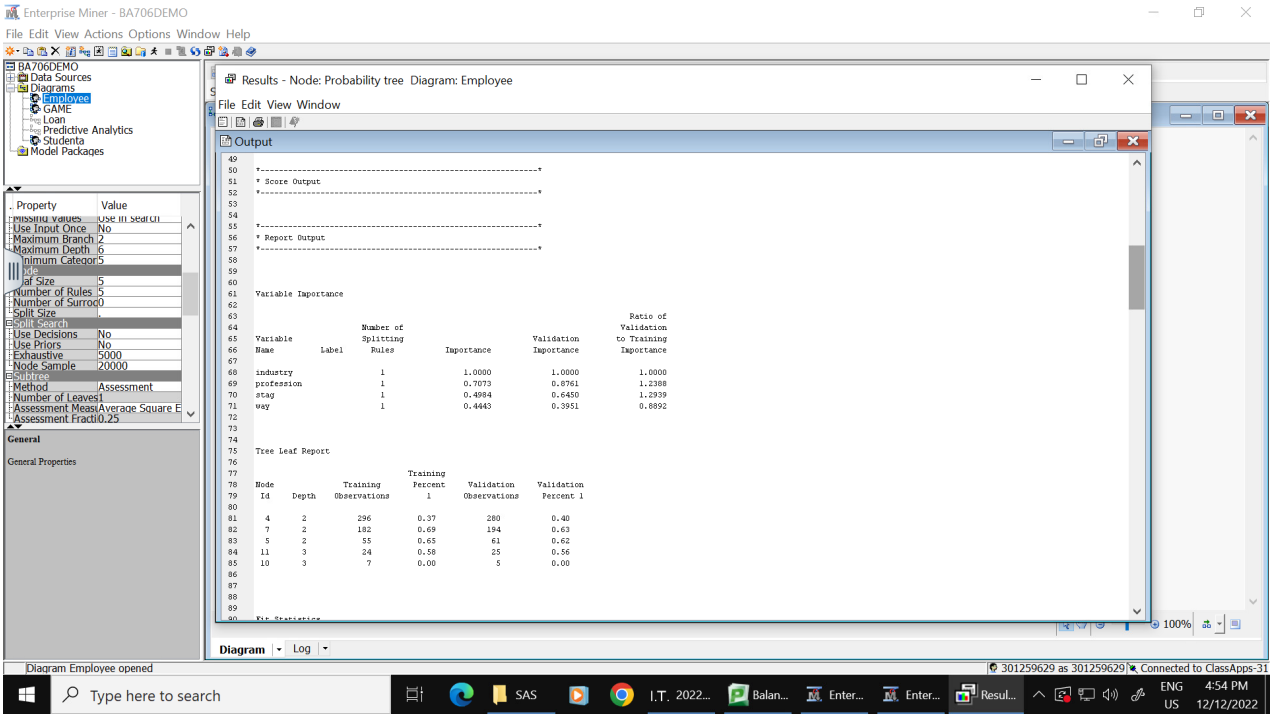
TWO BRANCH PROBABILITY TREE

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.


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Results - Node: Probability tree Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Tree
industry
Leaf Statistics
Training Percent 1
Validation Percent 1
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event NOBS Sum of Fr... 564 565
event MISC Misclassifi... 0.347518 0.385841
event MAX Maximum ... 0.686813 0.686813
event SSE Sum of S... 253.0793 267.8353
event ASE Average ... 0.224361 0.237022
event RASE Root Aver... 0.473668 0.486849
event DIV Divisor for... 1128 1130
event DFT Total Deg... 564
Treemap
Output
Diagram
Log
Diagram Employee opened
301259629 as 301259629
Connected to ClassApps-31
Type here to search
SAS
I.T. 2022...
Balan...
Enter...
Enter...
Resul...
ENG
US
4:53 PM
12/12/2022


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Results - Node: Probability tree Diagram: Employee
File Edit View Window
Tree
industry
POWERGENERATION, RETAIL...
BANKS, CONSULT, STATE, E...
profession
way
HR, ETC, BUSINESSDEVELO...
COMMERCIAL, SALES, FINAN...
FOOT
BUS, CAR Or Missing
stag
< 4.6817
>= 4.6817 Or Missing
Diagram
Log
Diagram Employee opened
301259629 as 301259629
Connected to ClassApps-31
Type here to search
SAS
I.T. 2022...
Balan...
Enter...
Enter...
Resul...
ENG
US
4:53 PM
12/12/2022



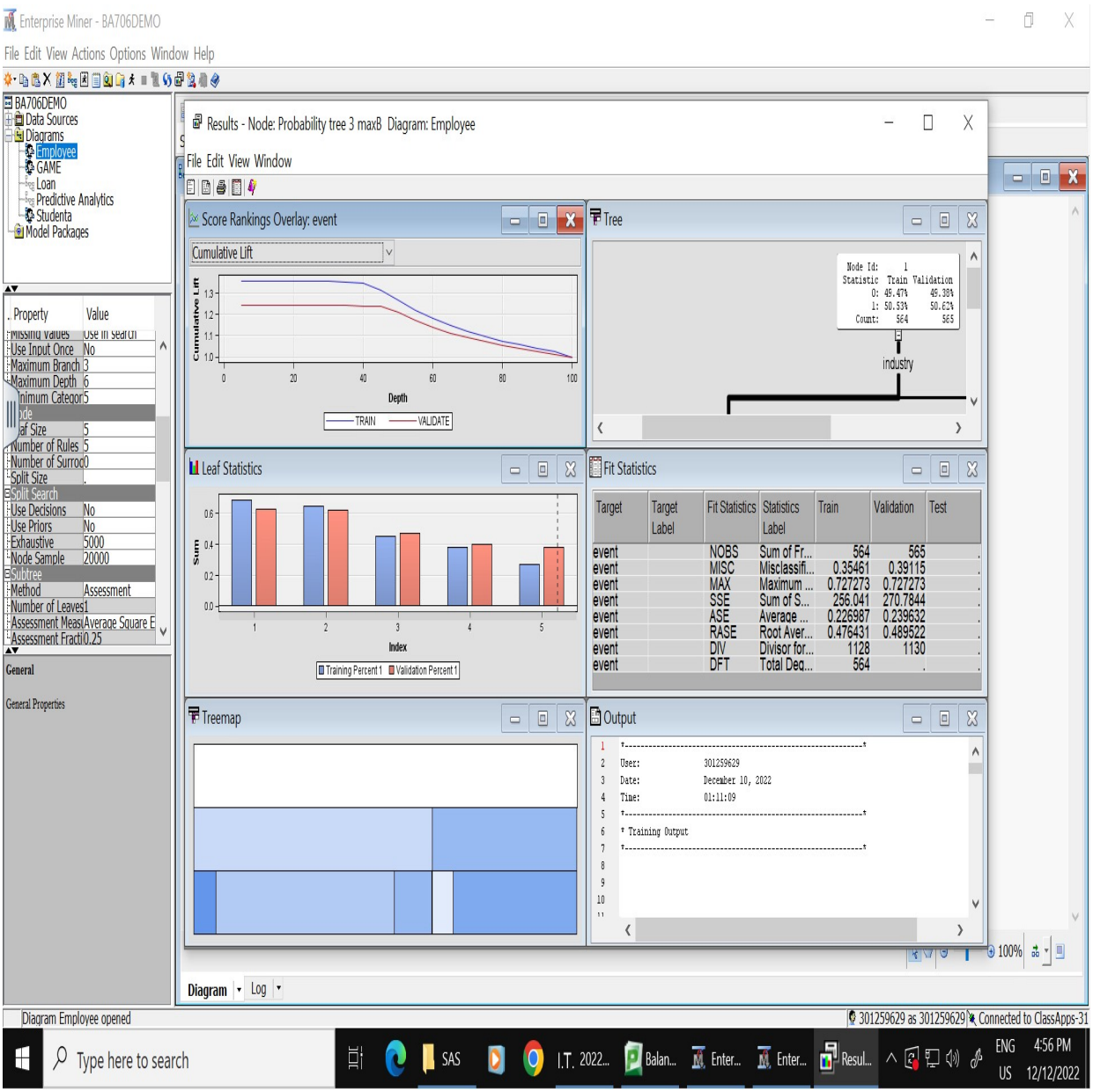


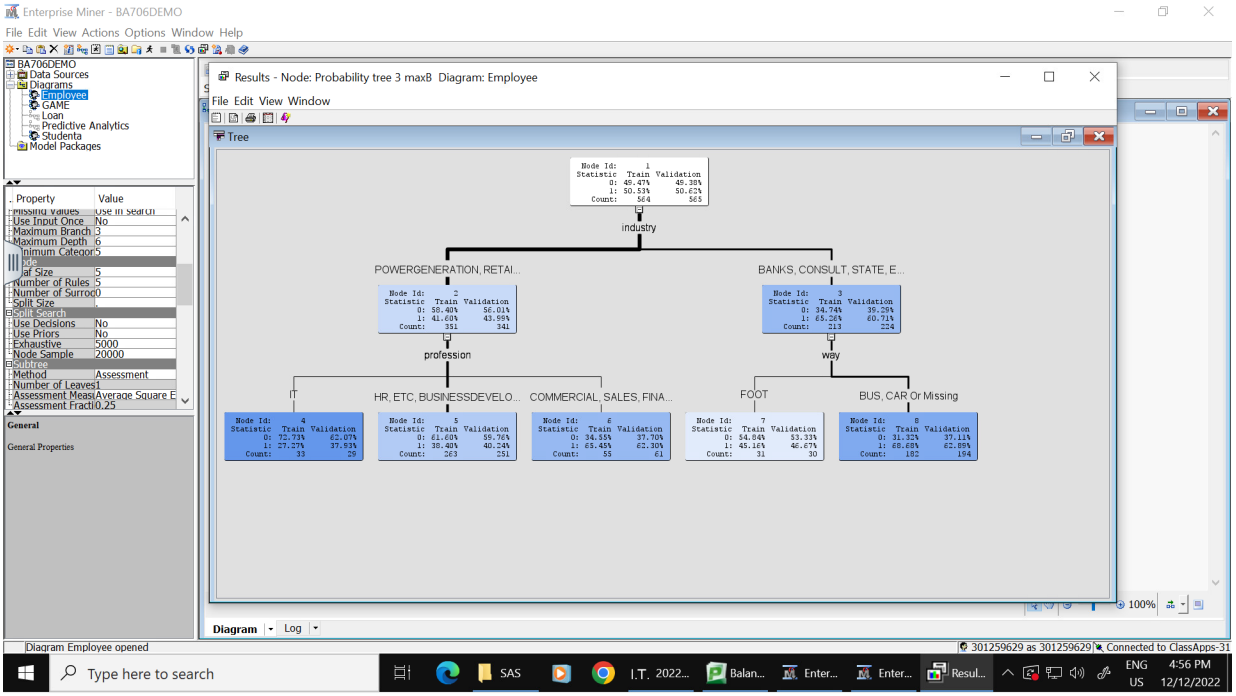
The tree has five leaves with four variable splits. this reduction is a result of pruning using the average squared error measure. The tree shows that 62.89% of employees in Banks, Consult and State industries who go to work in vehicles are likely to quit their jobs. 56% of employees who have worked 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, and retail industry are also likely to quit.

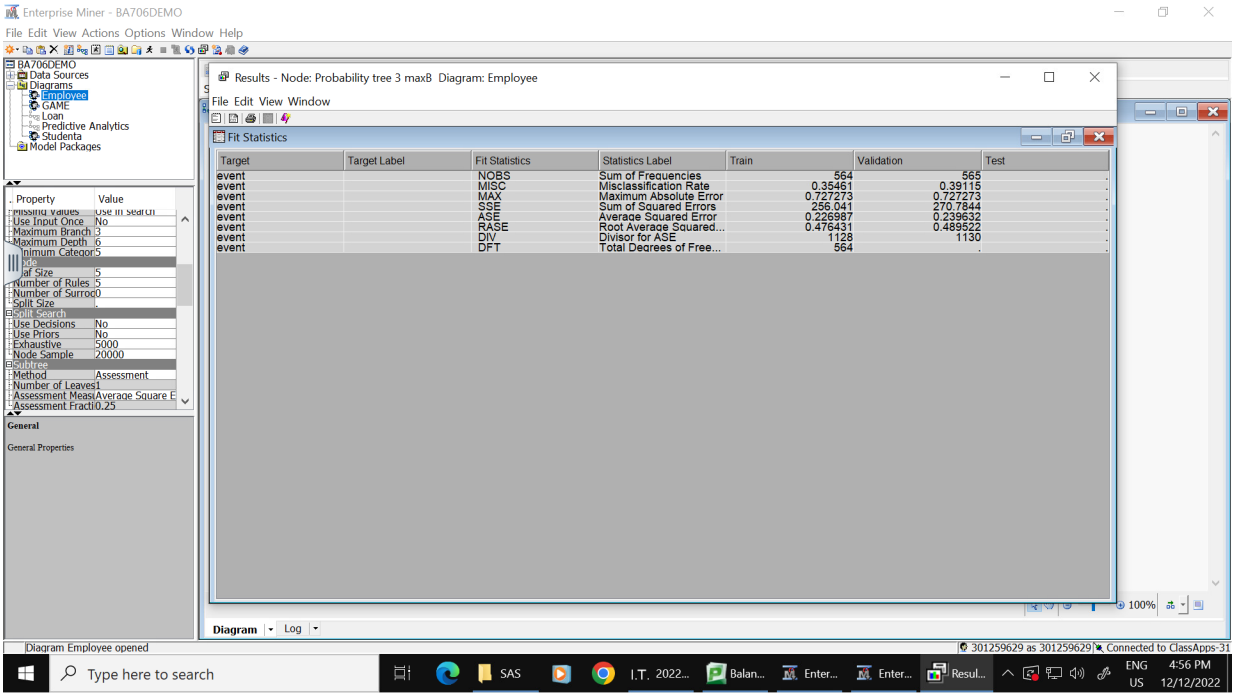
The Validation Average Square Error for the two-branch classification Tree is 0.237022. This has a similar validation ASE as the two-branch classification tree model.

THREE BRANCH PROBABILITY TREE

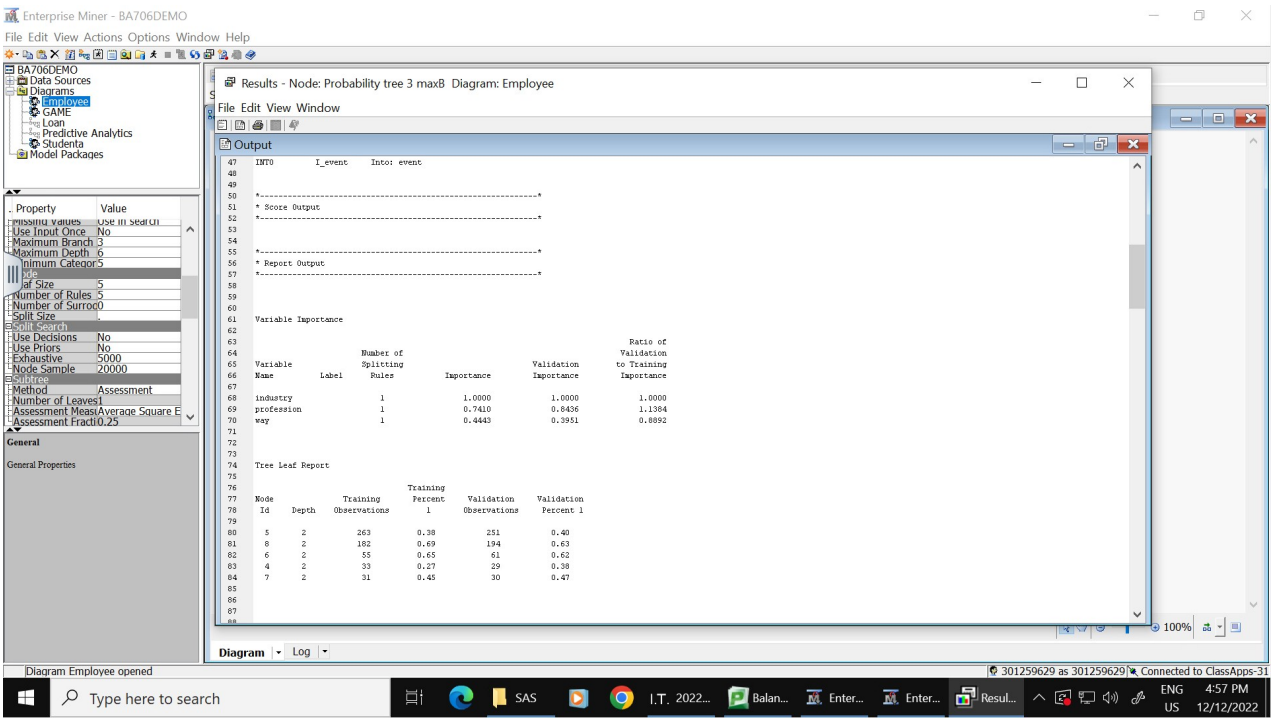
The decision tree node was attached to the data partition node with a maximum branch of three, ‘assessment’ used as the method and ‘average squared error’ as the assessment measure. Below shows the outcome after this node was run with the specifications noted.


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property
Value
Use Input Once
No
Maximum Branch
3
Maximum Depth
6
Use Decisions
No
Use Priors
No
Exhaustive
5000
Node Sample
20000
Method
Assessment
Number of Leaves
1
Assessment Fracti
0.25
General
General Properties
Results - Node: Probability tree 3 maxB Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Cumulative Lift
Depth
TRAIN
VALIDATE
Tree
Node Id: 1
Statistic Train Validation
0: 45.47% 45.38%
1: 50.53% 50.62%
Count: 564 565
industry
Leaf Statistics
Sum
Index
Training Percent 1
Validation Percent 1
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event NOBS Sum of Fr... 564 565
event MISC Misclassifi... 0.35461 0.39115
event MAX Maximum ... 0.727273 0.727273
event SSE Sum of S... 256.041 270.7844
event ASE Average ... 0.226987 0.239632
event RASE Root Aver... 0.476431 0.489522
event DIV Divisor for... 1128 1130
event DFT Total Deg... 564
Treemap
Output
User: 301259629
Date: December 10, 2022
Time: 01:11:09
* Training Output
Diagram
Log
Diagram Employee opened
301259629 as 301259629
Connected to ClassApps-31
Type here to search
SAS
I.T. 2022...
Balan...
Enter...
Enter...
Resul...
ENG
US
4:56 PM
12/12/2022





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 Absolute Error	0.727273	0.727273	.
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	.
event		DFT	Total Degrees of Free...	564	.	.

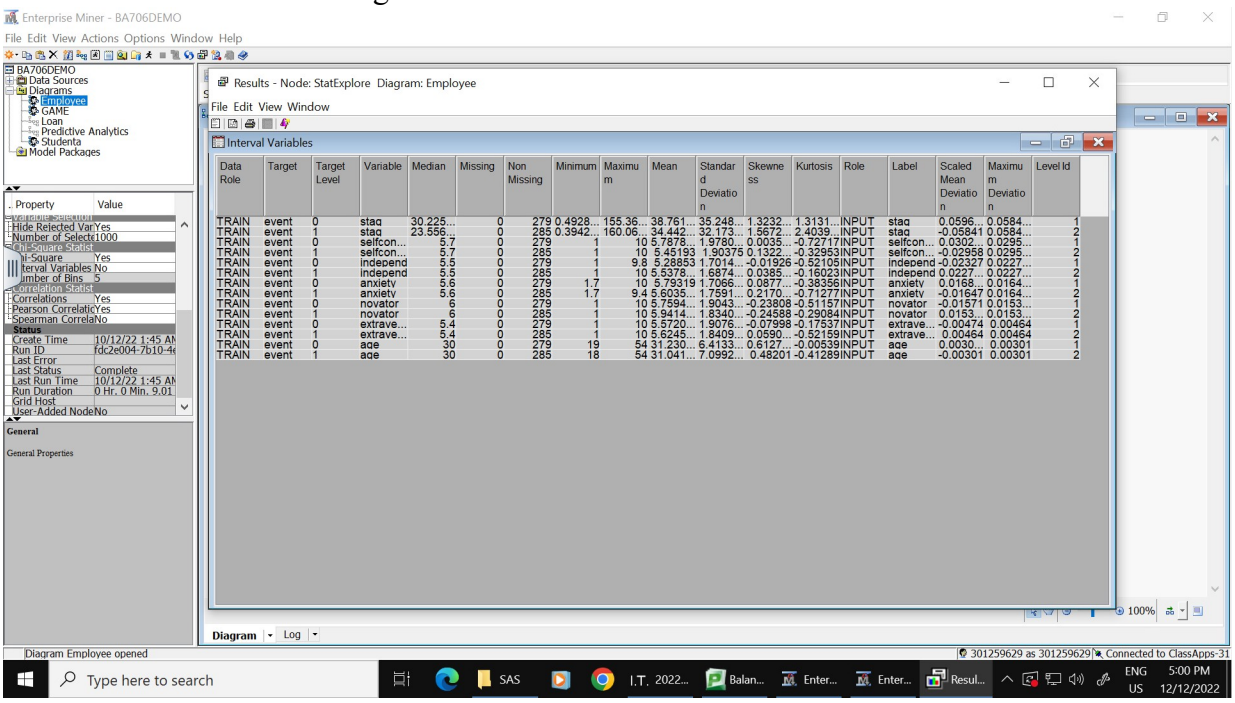


The tree has five leaves with three variable splits. This model has the least amount of significant variables when compared to all the other tree models. The stag variable which shows the amount of time an employee has worked was not considered significant enough to be included. The tree shows that 62.89% 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.

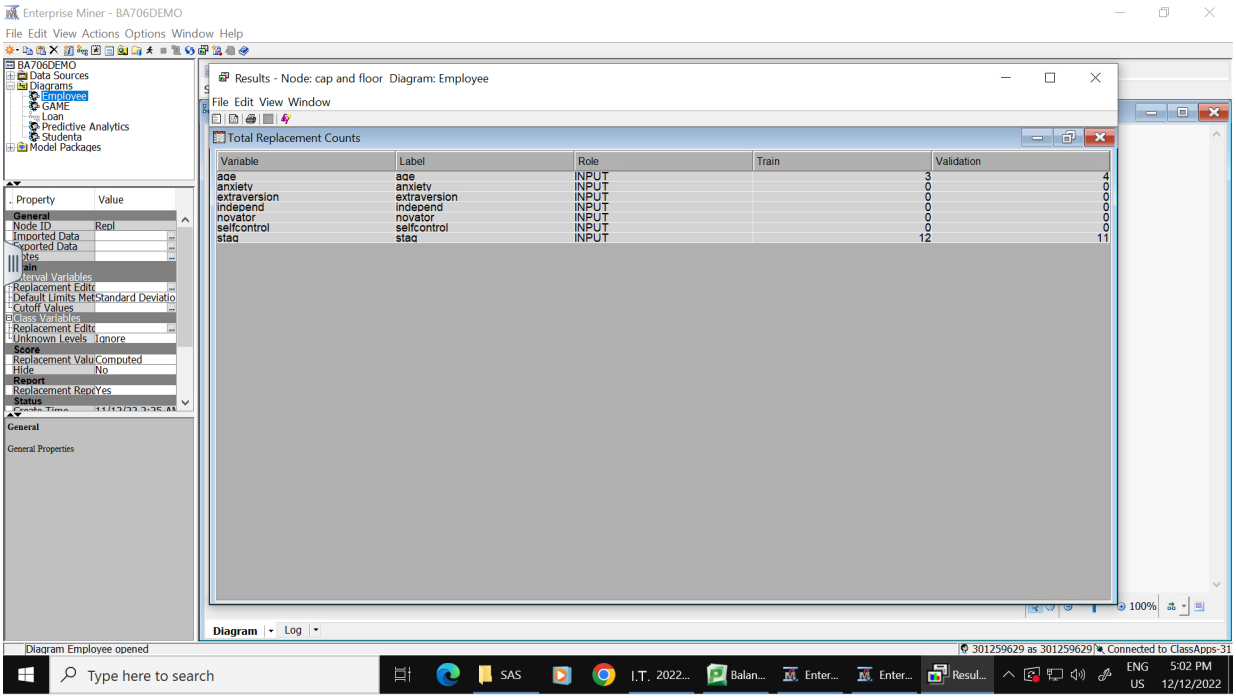
EXPLORING OUR DATA

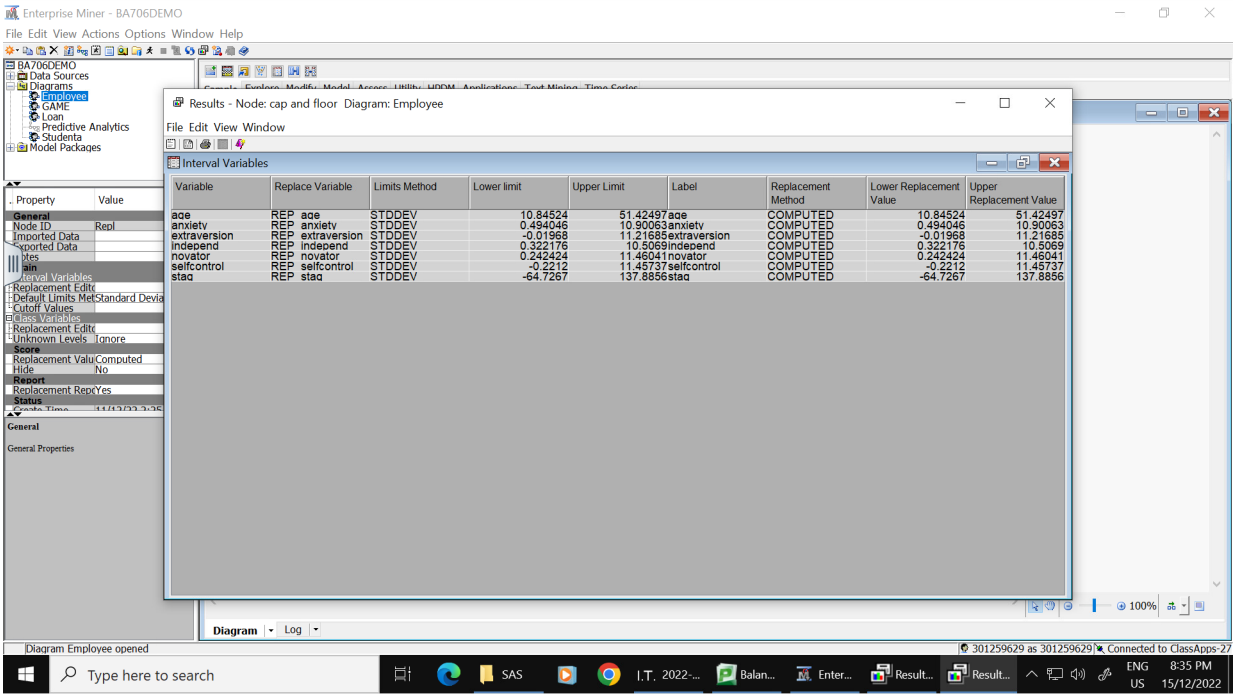
After we were done with analyzing our decision trees we decided to do a state explore to see if we have any missing values in the dataset. As decision trees are forgiving with respect to data, they are not affected by missing values, this step was not necessary before the tree models. However, for the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



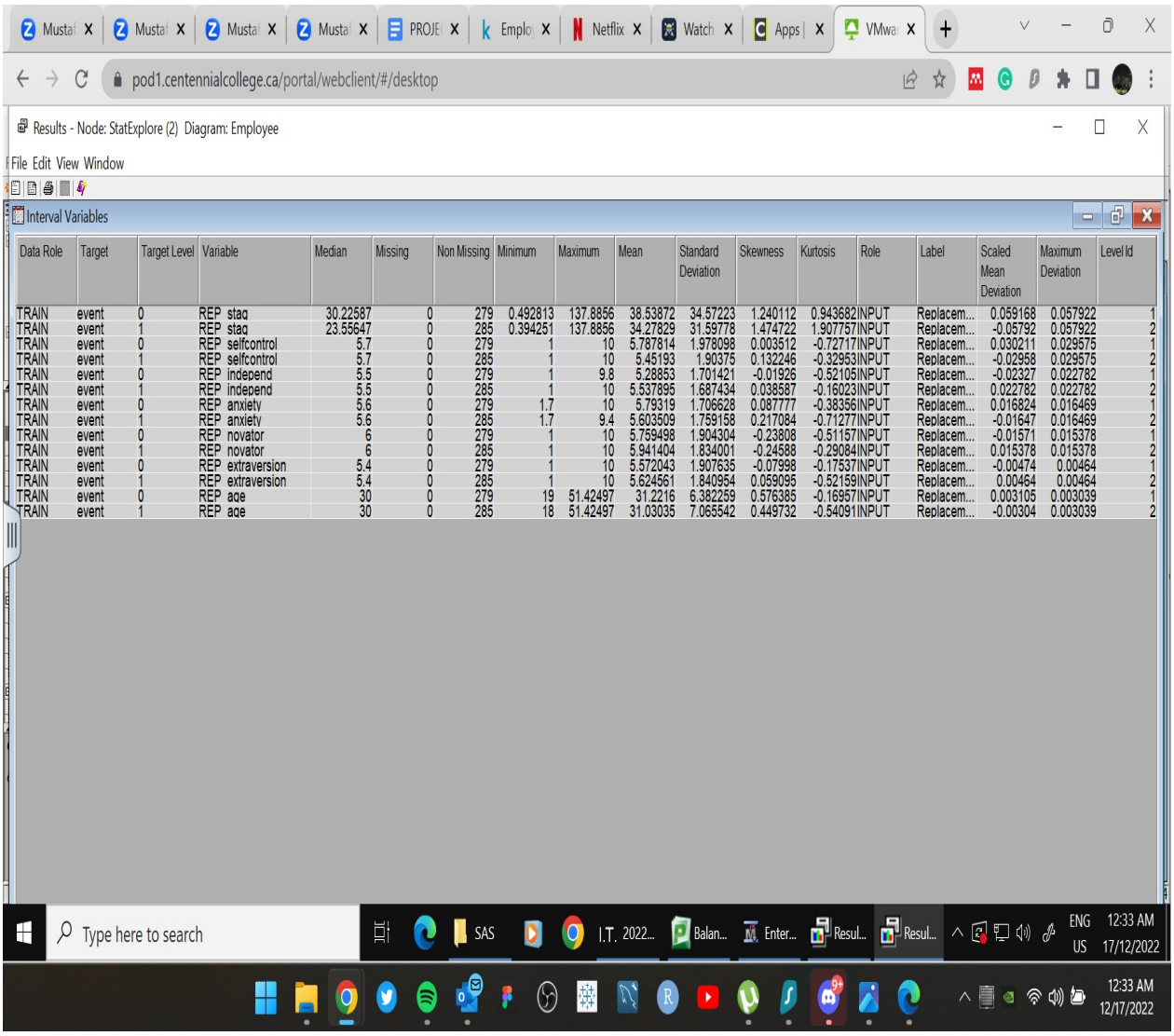
CAP AND FLOOR

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.



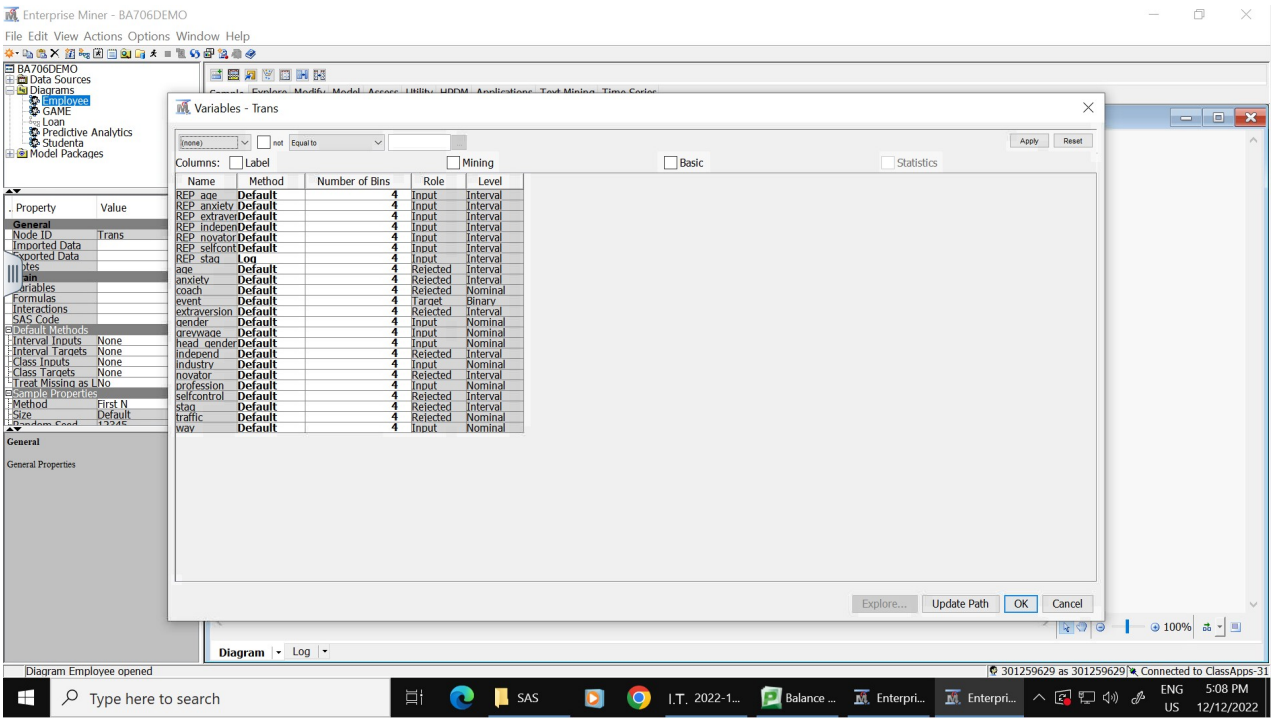


A stat explore node was connected to the replacement node to see the effect of the cap and floor on the skewed values.

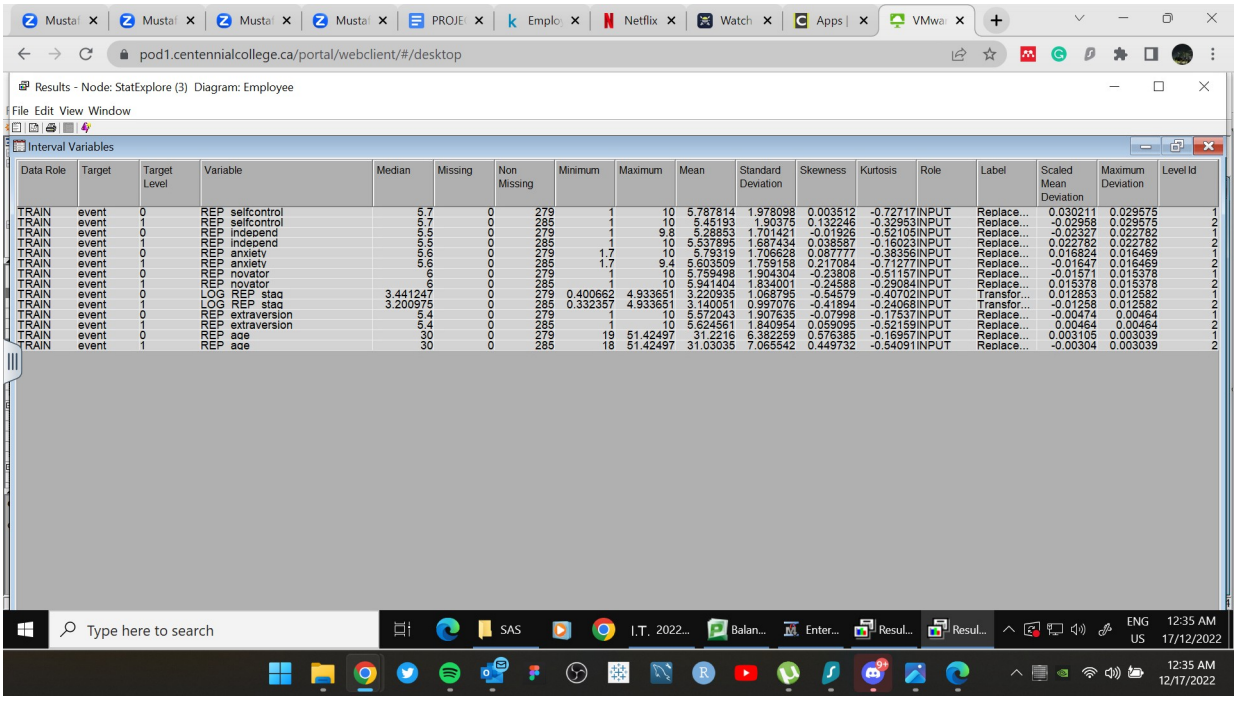


Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	event	0	REP_stag	30.22587	0	279	0.492813	137.8856	38.53872	34.57223	1.240112	0.943682	INPUT	Replacem...	0.059168	0.057922	1
TRAIN	event	1	REP_stag	23.55647	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	1.687434	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_age	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_age	30	0	285	18	51.42497	31.03035	7.065542	0.449732	-0.54091	INPUT	Replacem...	-0.00304	0.003039	2

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.



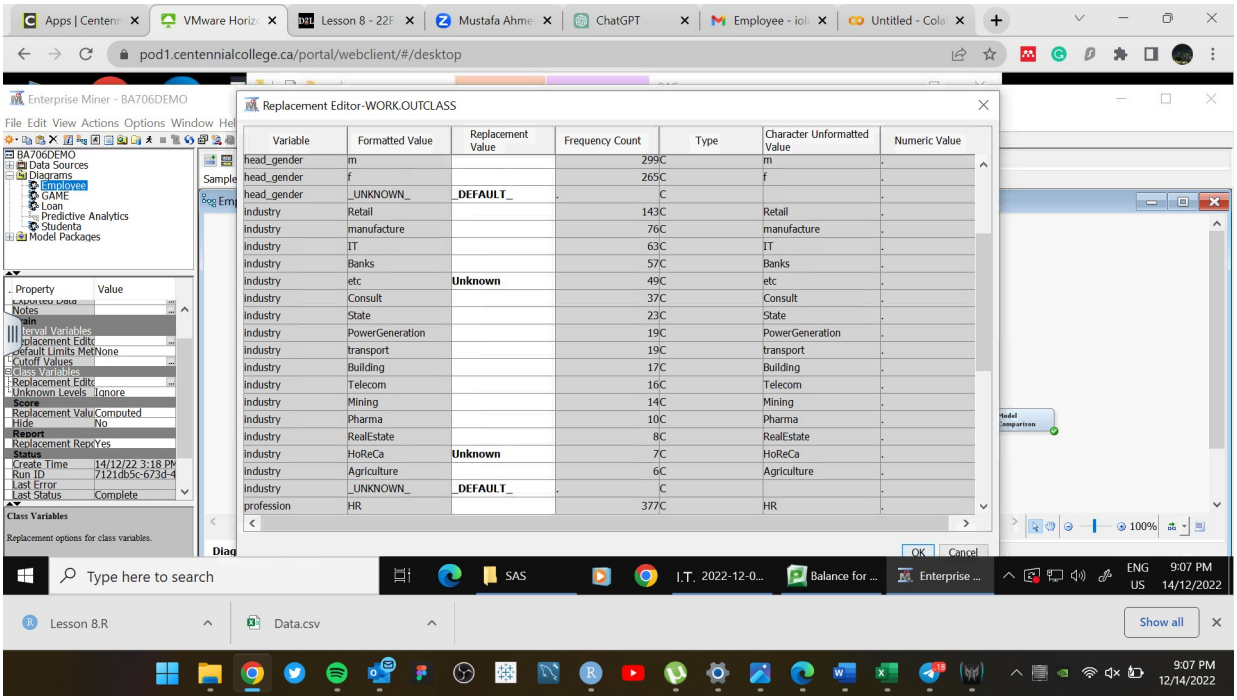
Next was to stat explore the data and we noticed the skewness was sorted as shown below:



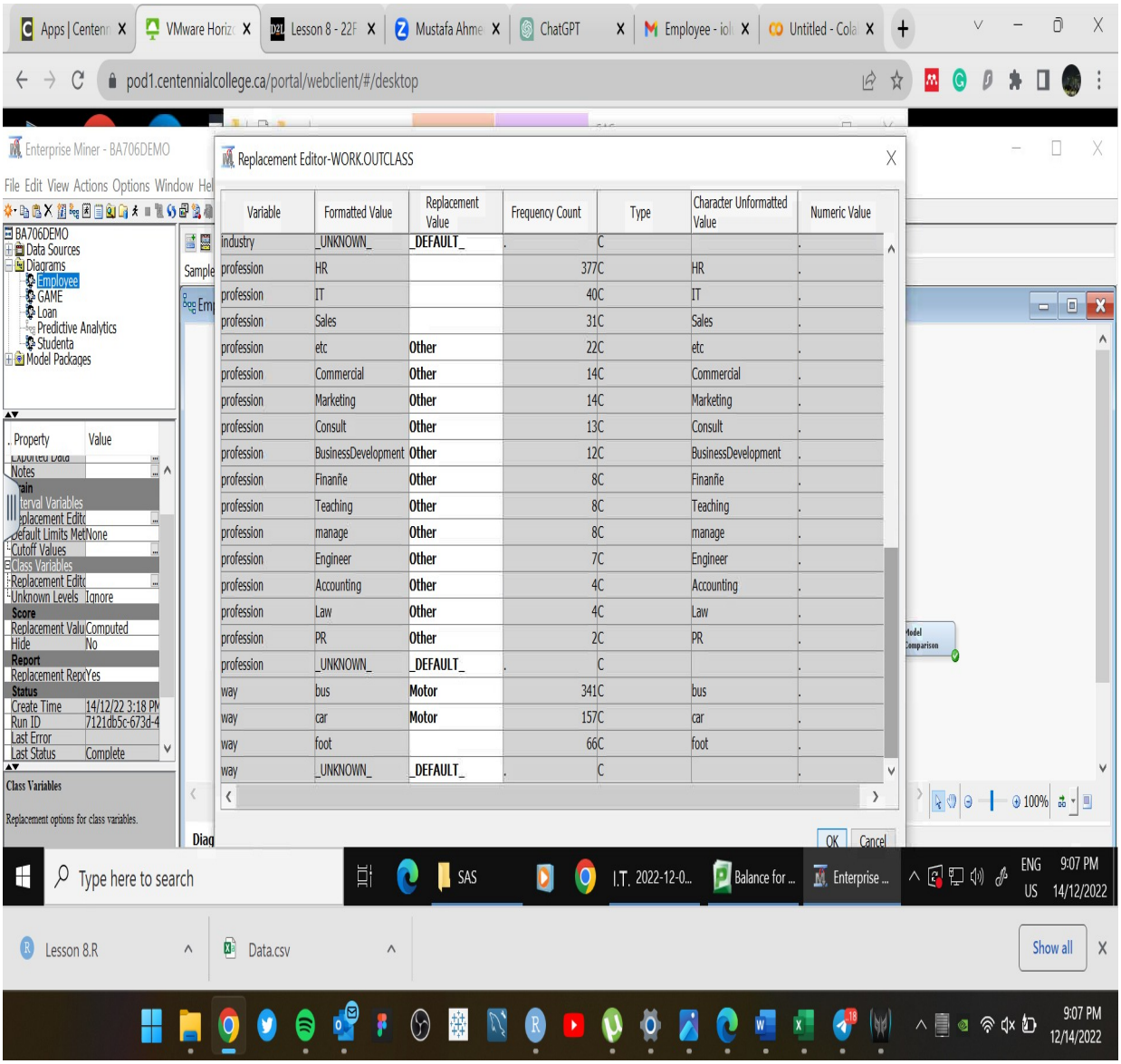
REPLACEMENT

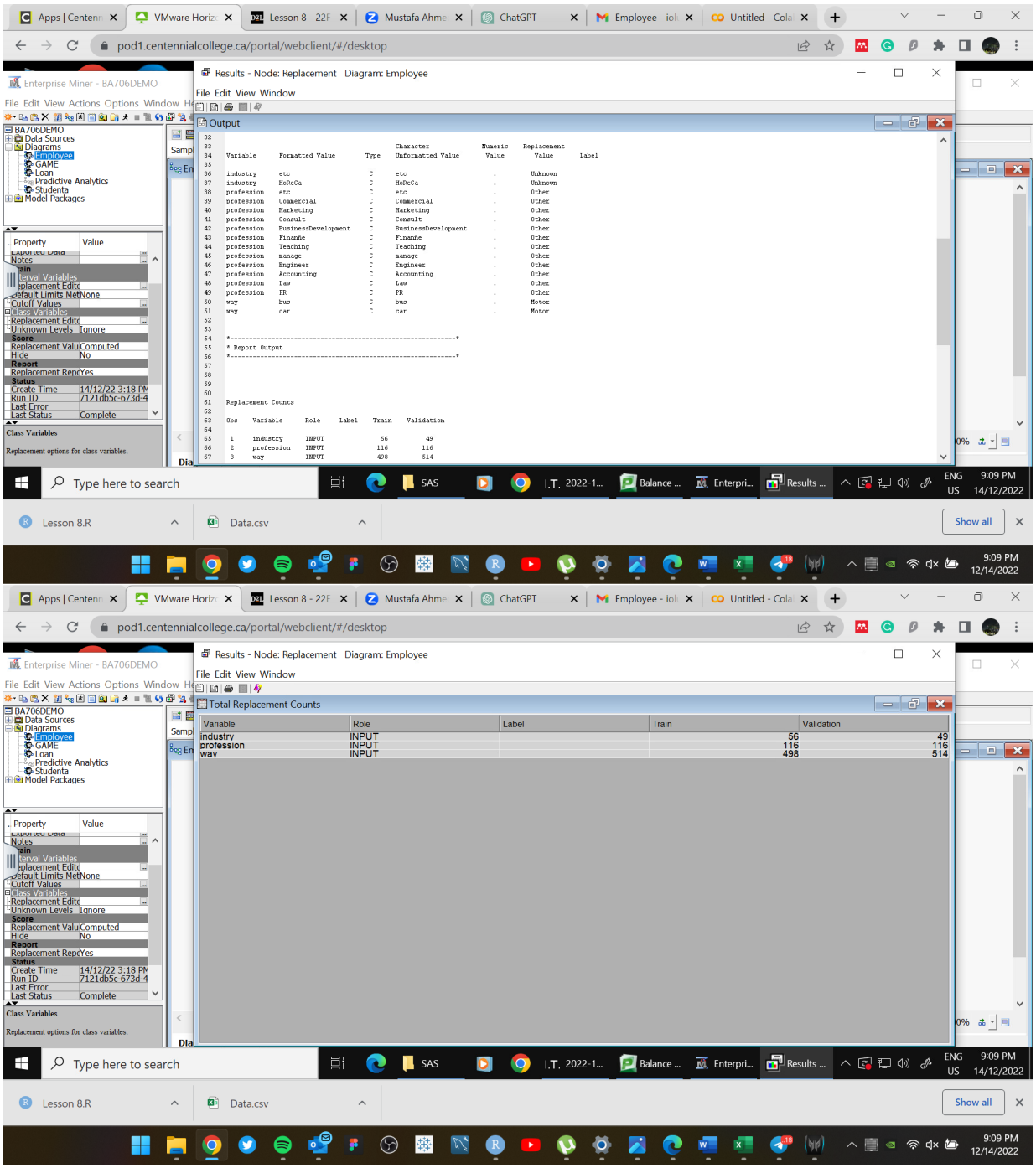
In logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and increases the stability and significance of the coefficients.

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



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’.





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:

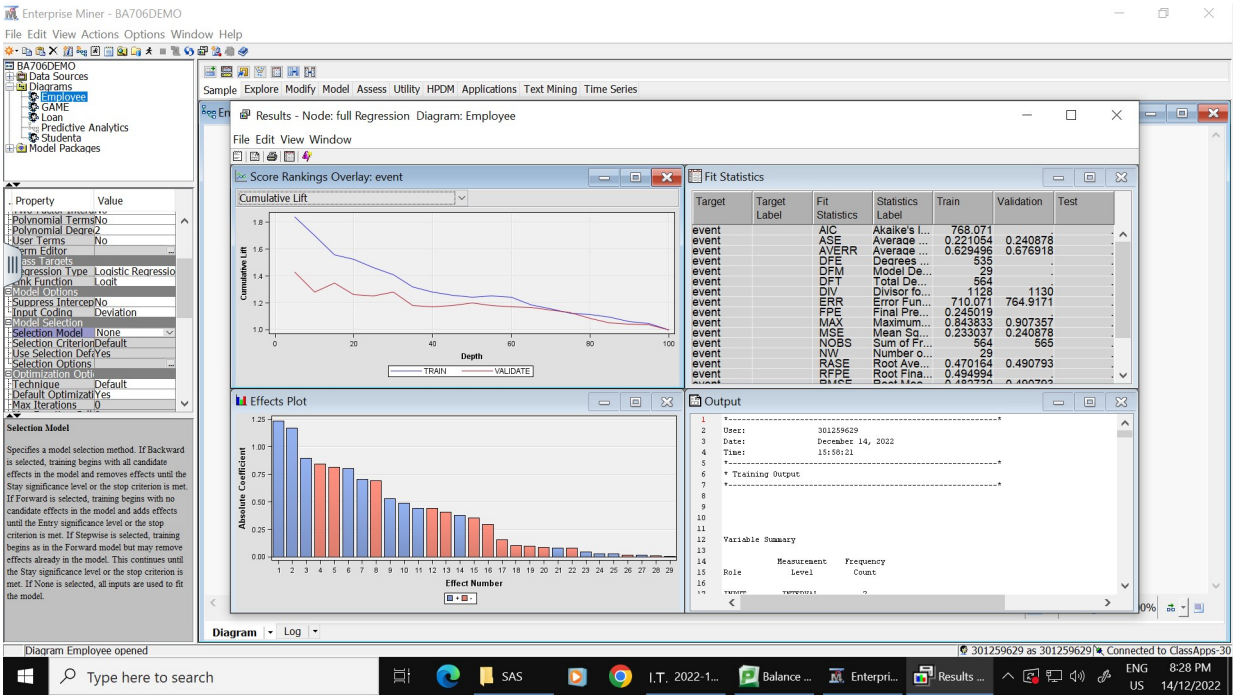
FORWARD REGRESSION

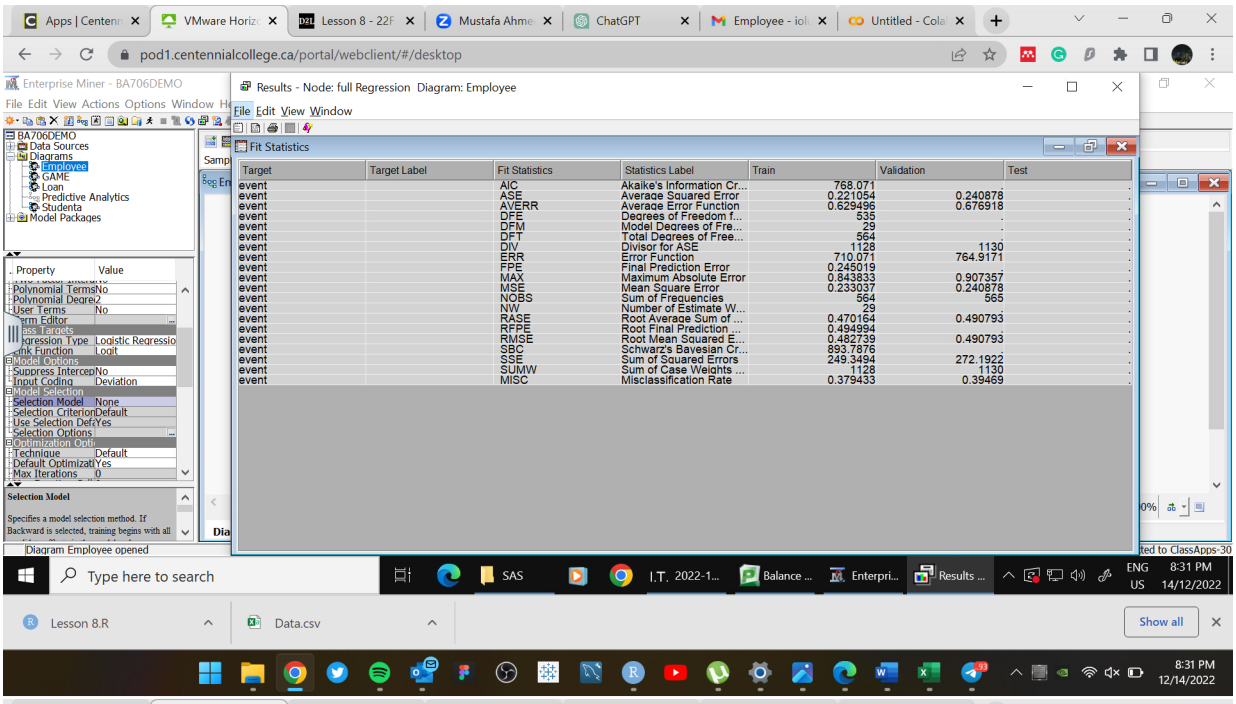
STEPWISE REGRESSION

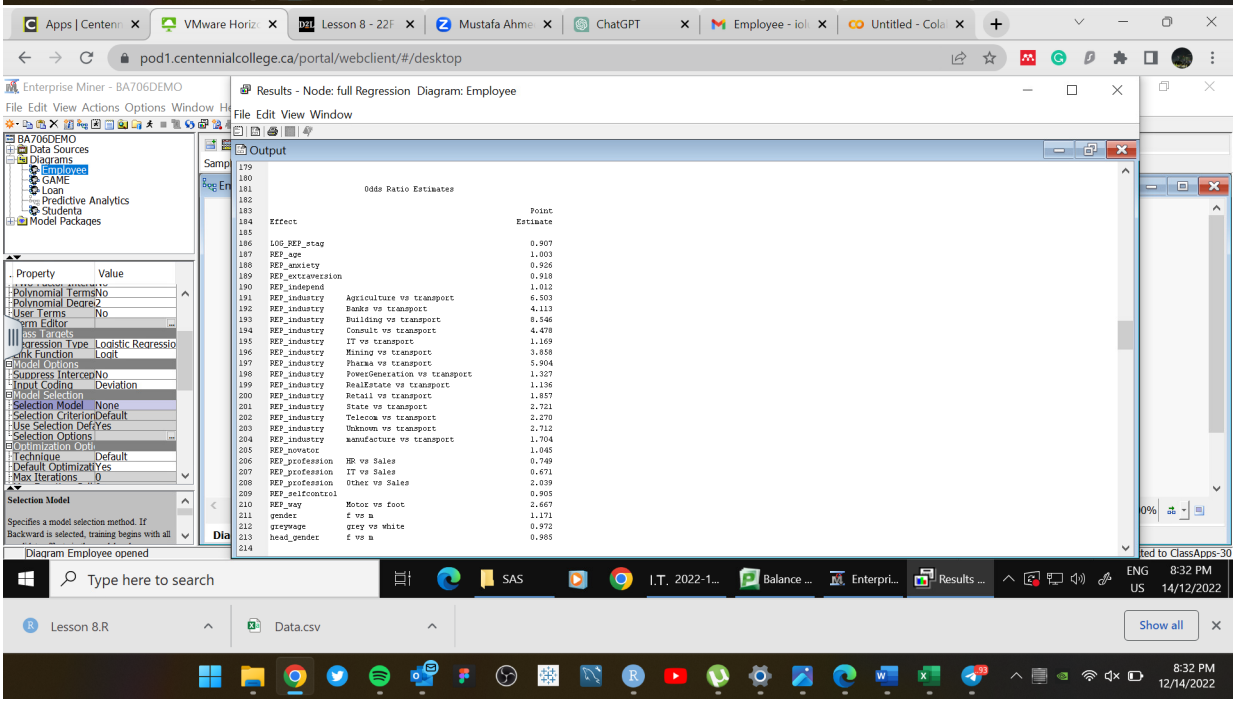
BACKWARD REGRESSION

FULL REGRESSION

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







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 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 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.
			When compared with Sales Profession
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.
			Other Variables

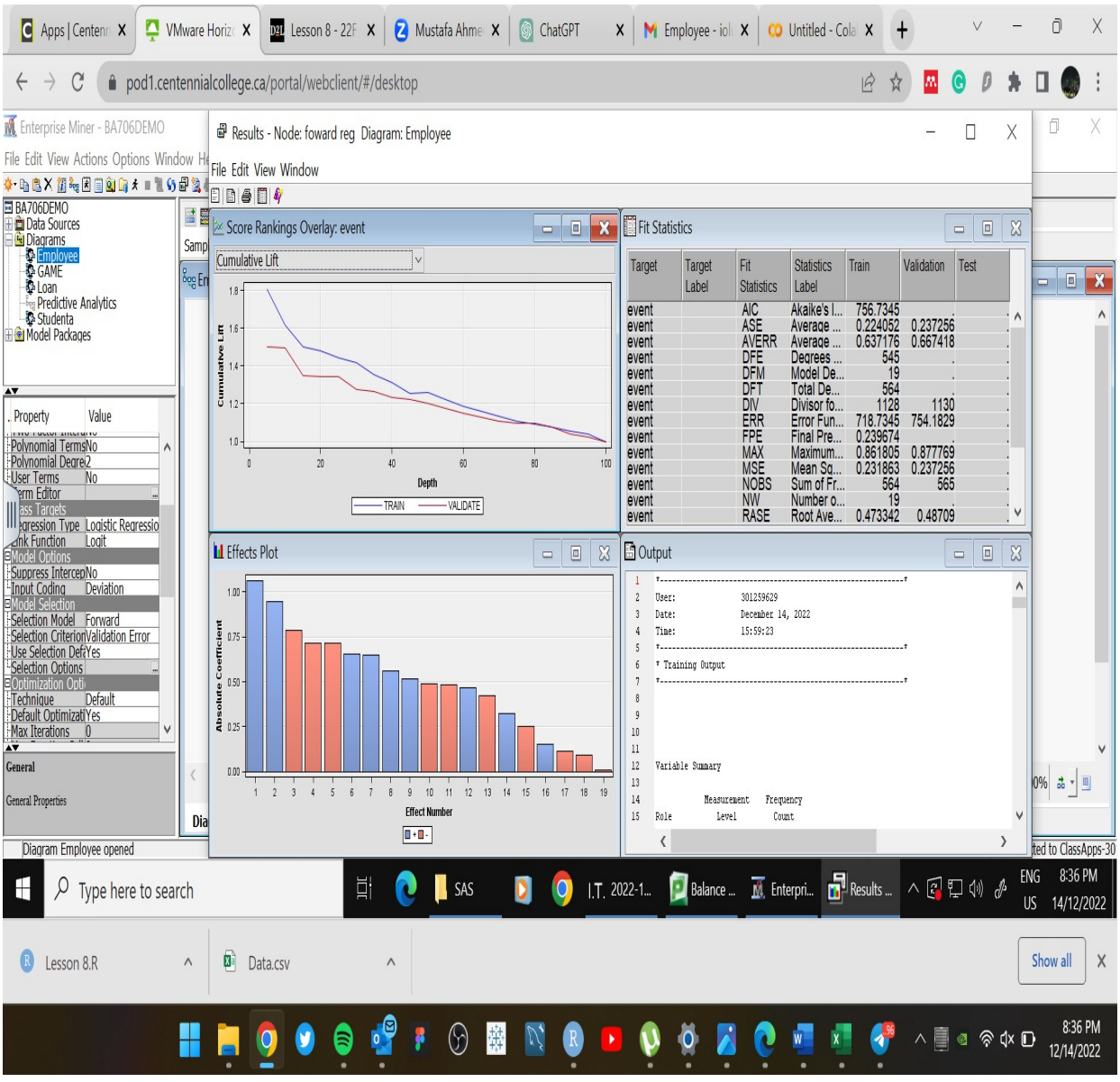
REP_way	Motor vs foot	2.667	Employees who come by Motor are 2.7 times more likely quit than those who come by foot
gender	f vs m	1.171	Female employees are 17.1% more likely to quit than male employees
REP_novator		1.045	For each 1 score the odds of quitting changed by a factor 1.045, a 4.5% increase
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

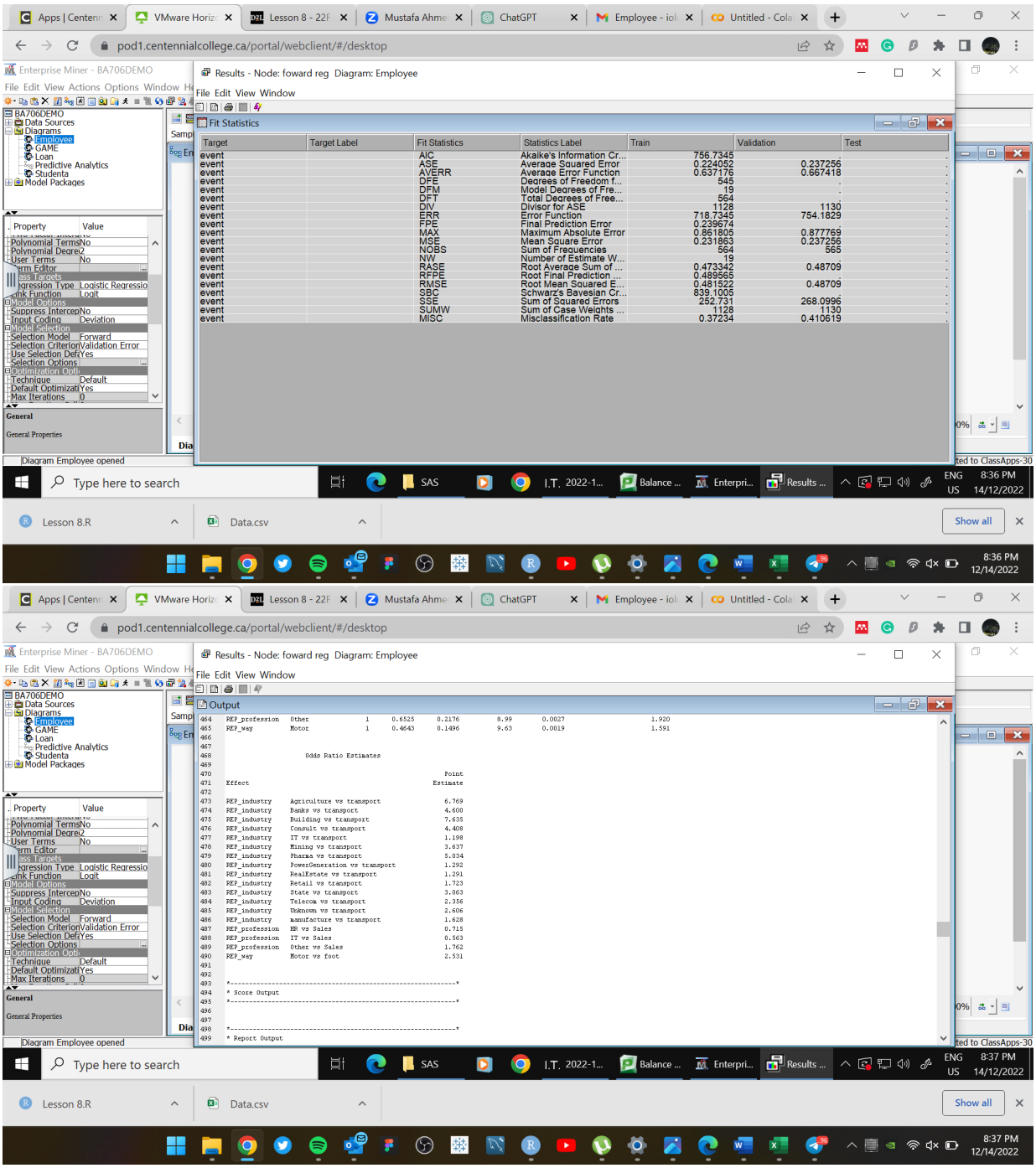
REP_selfcontrol		0.905	For each 1 score the odds of quitting changed by a factor of 0.905, a 9.5% decrease

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.


Results - Node: foward reg Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
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
Absolute Coefficient
Effect Number
Output
User: 301259629
Date: December 14, 2022
Time: 15:59:23
Training Output
Variable Summary
Measurement Frequency
Role Level Count
Enterprise Miner - BA706DEMO
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property
Value
Diagram Employee opened
Lesson 8.R
Data.csv
Show all


Results - Node: foward reg Diagram: Employee
File Edit View Window
Fit Statistics
Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test
event | AIC | Akaike's Information Cr... | 756.7345 |
event | ASE | Average Squared Error | 0.224052 | 0.237256
event | AVERR | Average Error Function | 0.637176 | 0.667418
event | DFE | Degrees of Freedom f... | 545 |
event | DFM | Model Degrees of Fre... | 19 |
event | DFT | Total Degrees of Free... | 564 |
event | DIV | Divisor for ASE | 1128 | 1130
event | ERR | Error Function | 718.7345 | 754.1829
event | FPE | Final Prediction Error | 0.239674 |
event | MAX | Maximum Absolute Error | 0.861805 | 0.877769
event | MSE | Mean Square Error | 0.231863 | 0.237256
event | NOBS | Sum of Frequencies | 564 | 565
event | NW | Number of Estimate 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 | SBC | Schwarz's Bayesian Cr... | 839.1005 |
event | SSE | Sum of Squared Errors | 252.731 | 268.0996
event | SUMW | Sum of Case Weights ... | 1128 | 1130
event | MISC | Misclassification Rate | 0.37234 | 0.410619
Enterprise Miner - BA706DEMO
Diagram Employee opened
Type here to search
Lesson 8.R
Data.csv
Show all
8:36 PM
12/14/2022

Results - Node: foward reg Diagram: Employee
File Edit View Window
Output
Odds Ratio Estimates
Point
Effect 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 Mining vs transport 3.637
REP_industry Pharma vs transport 5.034
REP_industry PowerGeneration vs transport 1.292
REP_industry RealEstate vs transport 1.291
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 2.606
REP_industry manufacture vs transport 1.628
REP_profession HR vs Sales 0.715
REP_profession IT vs Sales 0.563
REP_profession Other vs Sales 1.762
REP_way Motor vs foot 2.531
* Score Output
* Report Output
Enterprise Miner - BA706DEMO
Diagram Employee opened
Type here to search
Lesson 8.R
Data.csv
Show all
8:37 PM
12/14/2022

Interpretation of Forward Regression

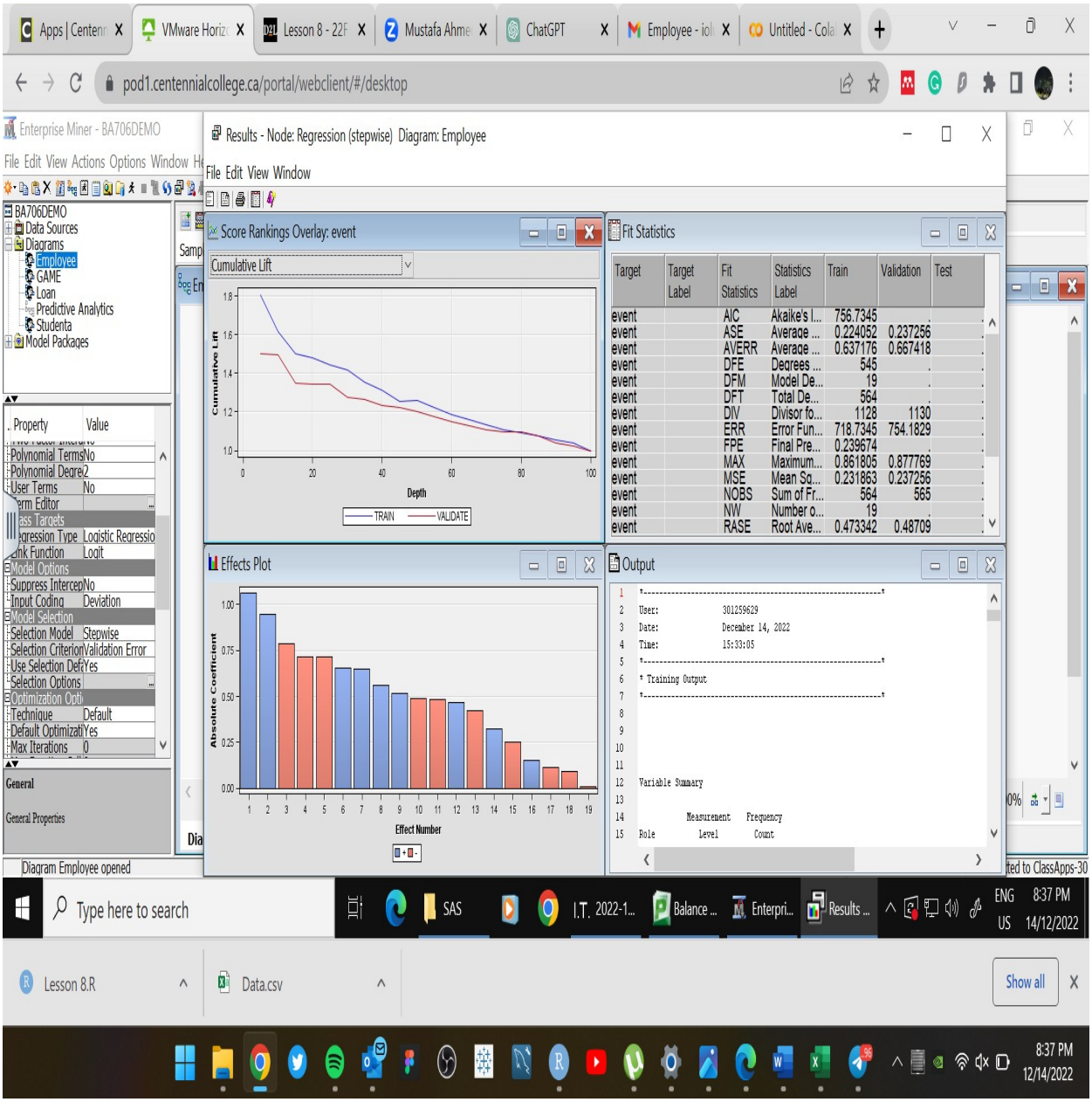
		PoinT Estimate	When compared with the Transport industry
REP_industry	Building vs transport	7.635	Employees in the Building industry are 7.6 times more likely to quit.
REP_industry	Agric vs transport	6.769	Employees in the Agriculture industry are 6.8 times more likely to quit.
REP_industry	Pharma vs 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.

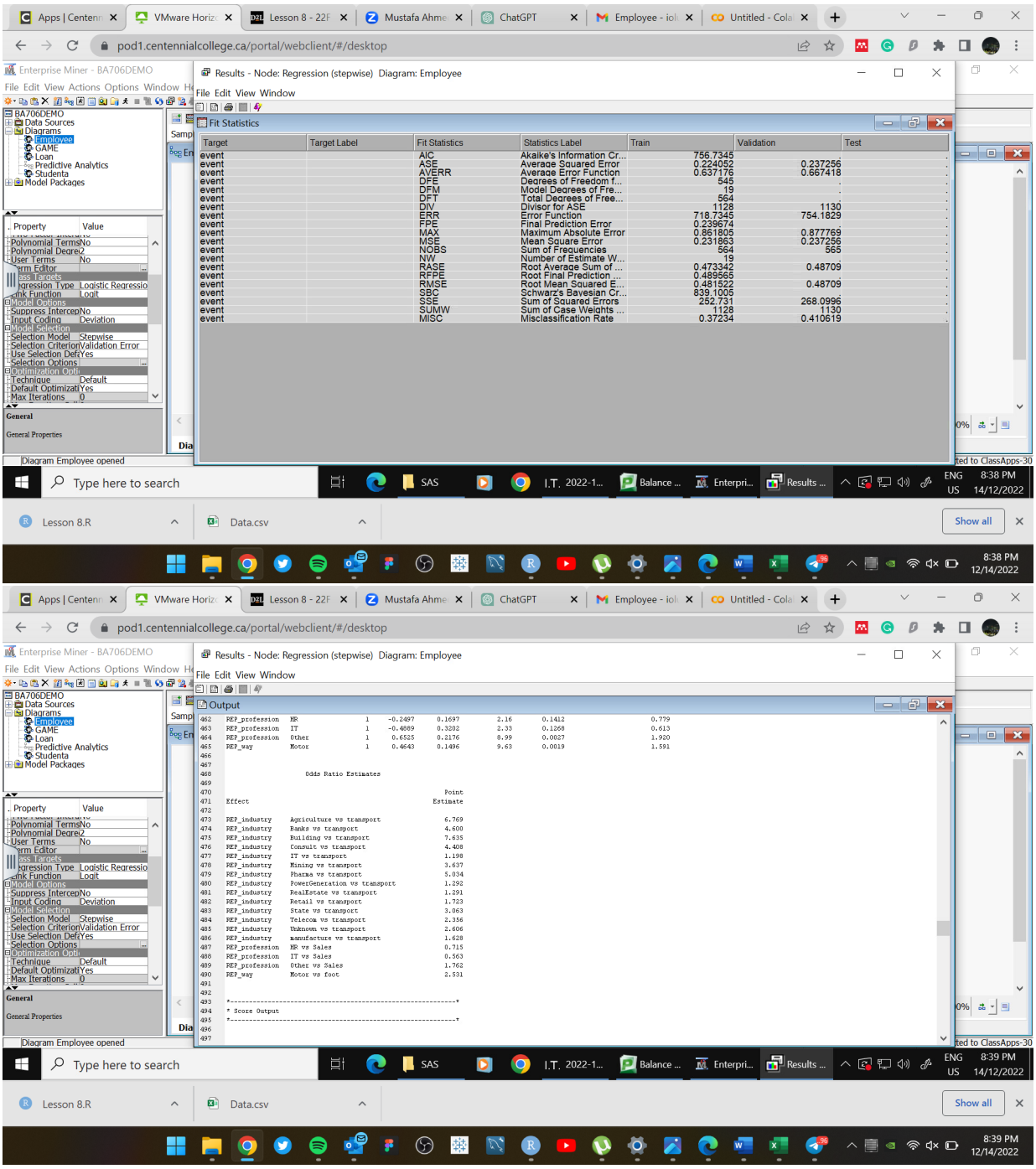
REP_industry	Unknown vs transport	2.606	Employees in the all other industries not listed are 2.6 times more likely to quit.
REP_industry	Telecom vs transport	2.356	Employees in the Telecom industry are 2.4% more likely to quit.
REP_industry	Retail vs transport	1.723	Employees in the Retail industry are 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	r 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	loyees 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.





Interpretation of Stepwise Regression

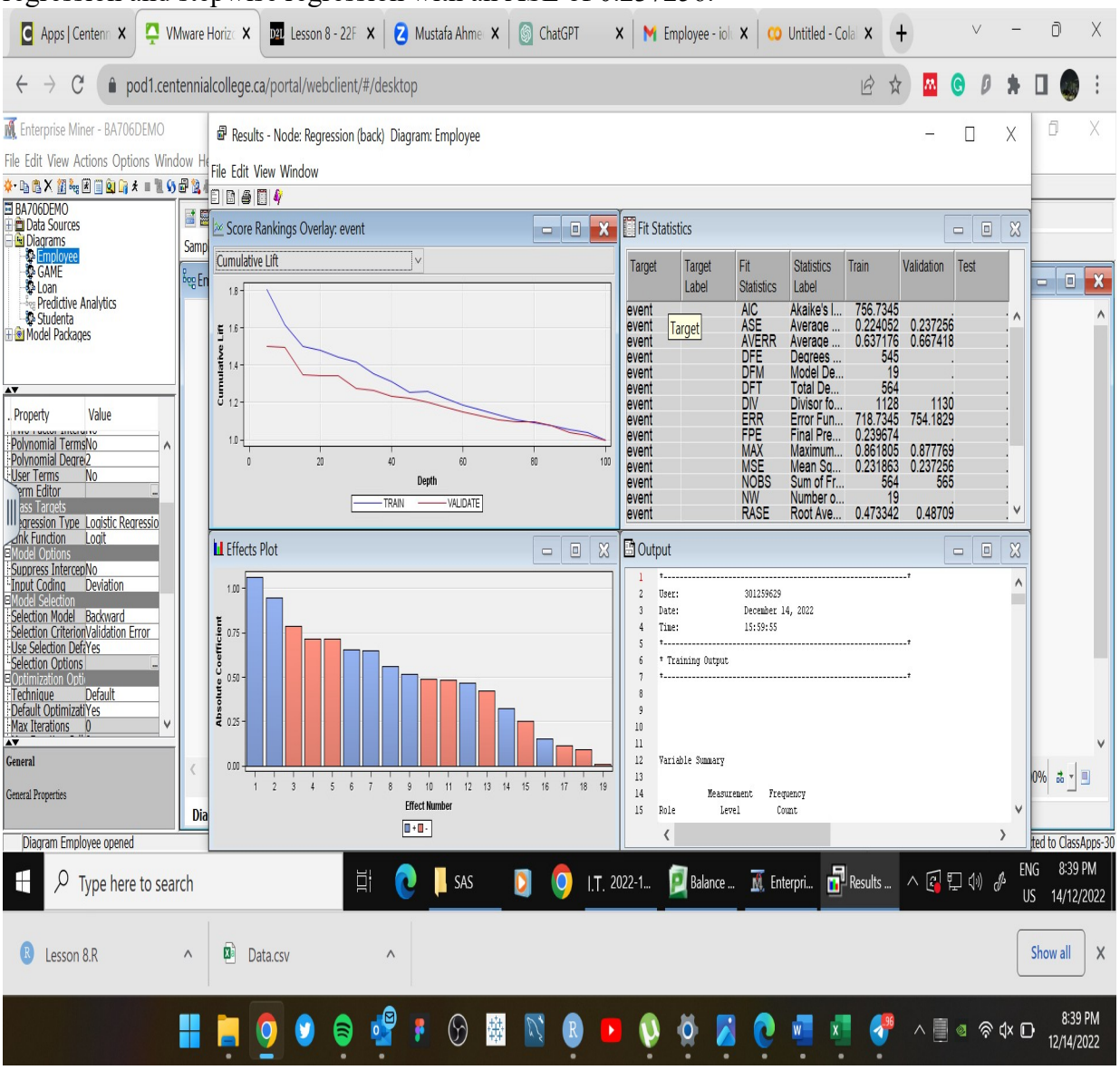
Variables		Point Estimate	When compared to Transport Industry
Rep_Industry	Agriculture vs transport	6.769	Employees in the

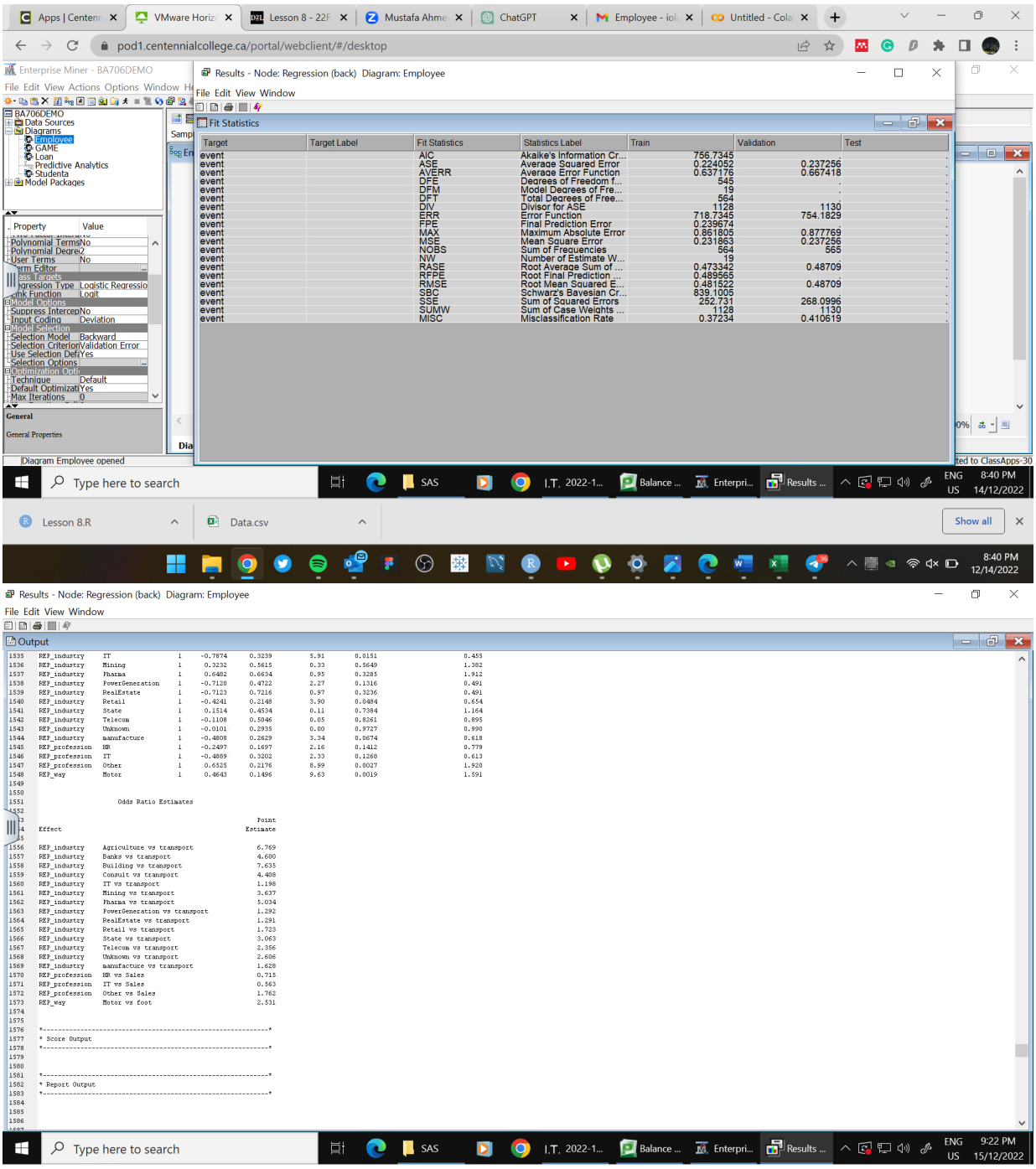
			Agriculture industry are 6.8 times more likely to quit
Rep_Industry	Banks vs transport	4.600	Employees in the Bank industry are 4.6 times more likely to quit
Rep_Industry	Building vs transport	7.635	Employees in the Building industry are 7.6 times more likely to quit
Rep_Industry	Consult vs transport	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
			When compared with Sales Profession
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
			Other Variable
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.




Enterprise Miner - BA706DEMO
Results - Node: Regression (back) Diagram: Employee
File Edit View Window
Fit Statistics
Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test
event | | AIC | Akaike's Information Cr... | 756.7345 | . | .
event | | ASE | Average Squared Error | 0.224052 | 0.237256 | .
event | | AVERR | Average Error Function | 0.637176 | 0.667418 | .
event | | DFE | Degrees of Freedom f... | 545 | . | .
event | | DFM | Model Degrees of Fre... | 19 | . | .
event | | DFT | Total Degrees of Free... | 564 | . | .
event | | DIV | Divisor for ASE | 1128 | 1130 | .
event | | ERR | Error Function | 718.7345 | 754.1829 | .
event | | FPE | Final Prediction Error | 0.239674 | . | .
event | | MAX | Maximum Absolute Error | 0.861805 | 0.877769 | .
event | | MSE | Mean Square Error | 0.231863 | 0.237256 | .
event | | NOBS | Sum of Frequencies | 564 | 565 | .
event | | NW | Number of Estimate 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 | | SBC | Schwarz's Bayesian Cr... | 839.1005 | . | .
event | | SSE | Sum of Squared Errors | 252.731 | 268.0996 | .
event | | SUMW | Sum of Case Weights ... | 1128 | 1130 | .
event | | MISC | Misclassification Rate | 0.37234 | 0.410619 | .
Diagram Employee opened
Type here to search
Lesson 8.R
Data.csv
Show all
8:40 PM 12/14/2022
Output
Odds Ratio Estimates
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 Mining vs transport 3.637
REP_industry Pharma vs transport 5.034
REP_industry PowerGeneration vs transport 1.292
REP_industry RealEstate vs transport 1.291
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 2.606
REP_industry manufacture vs transport 1.628
REP_profession HR vs Sales 0.715
REP_profession IT vs Sales 0.563
REP_profession Other vs Sales 1.762
REP_way Motor vs foot 2.531
* Score Output
* Report Output
9:22 PM 15/12/2022

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
			When compared with Sales Profession
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
			Other Variable
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

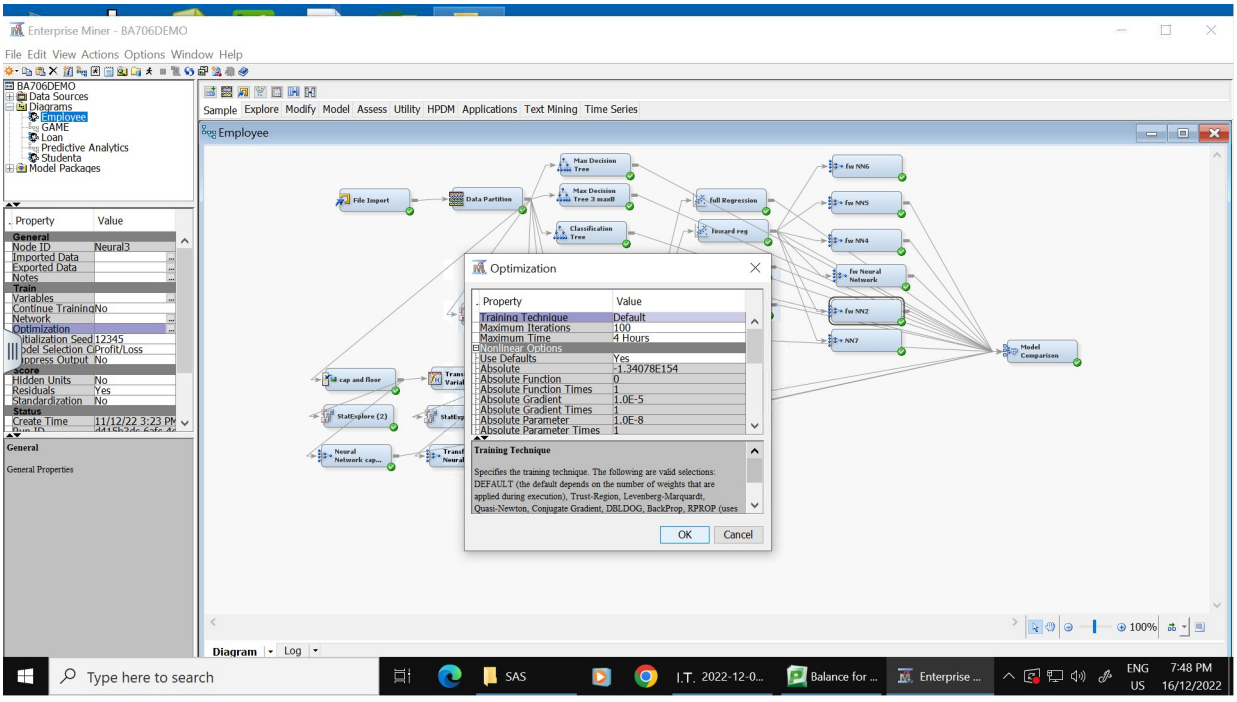
With the full regression having an ASE of 0.240878 and Forward, Stepwise , and Backward regression models all having the same ASE of 0.237256 we have an option of choosing from any of the three with the better average squared error. We chose the Forward regression model because of its additive style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

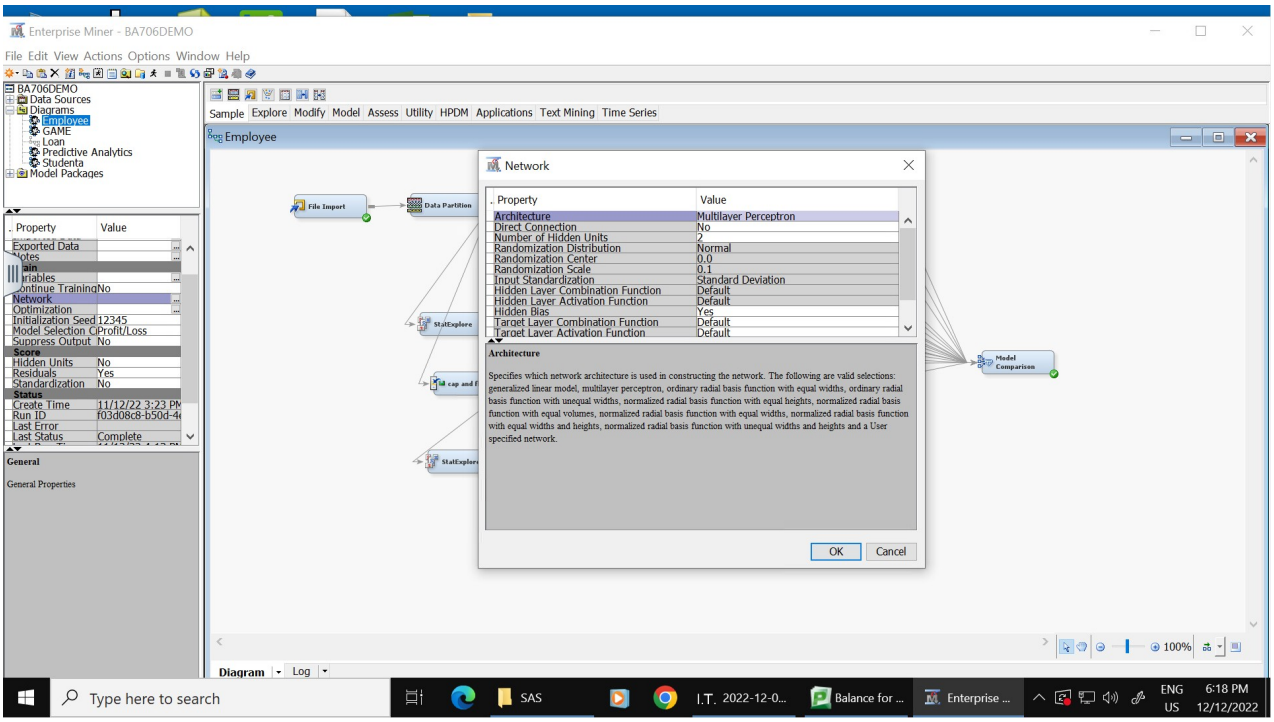
NEURAL NETWORK

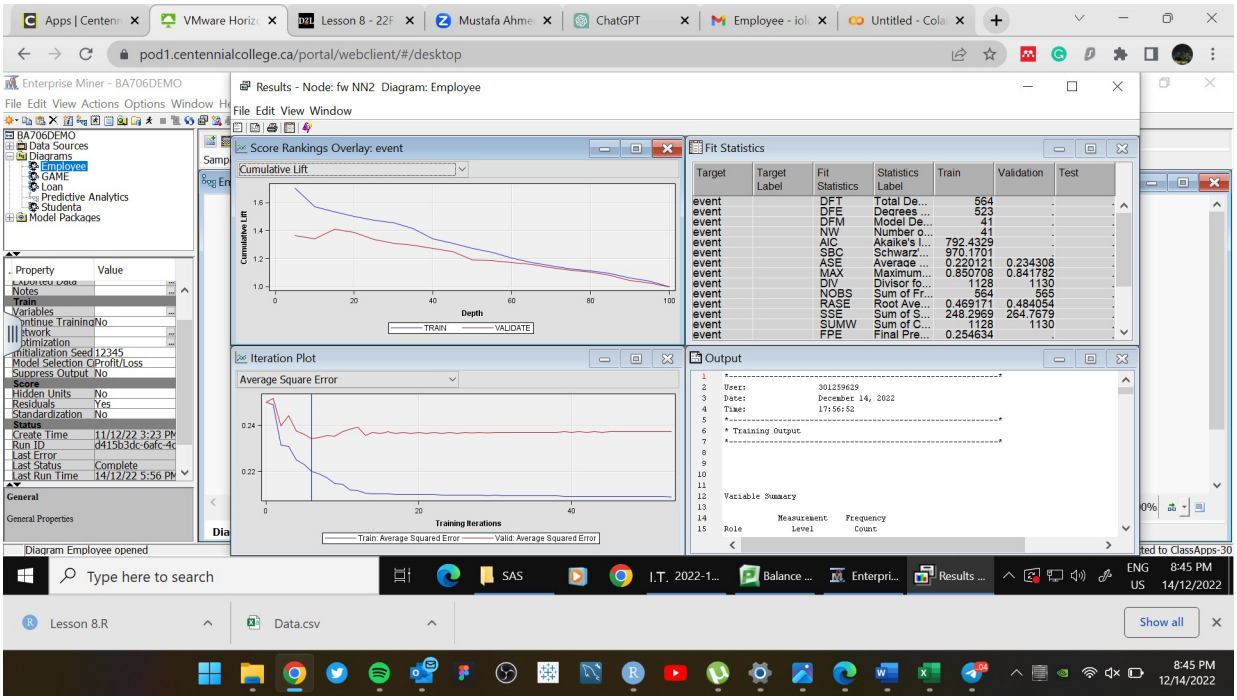
All neural networks were run using 100 iterations, and preliminary training was turned off. The distinguishing factor for all the neural networks run was the number of hidden units used.

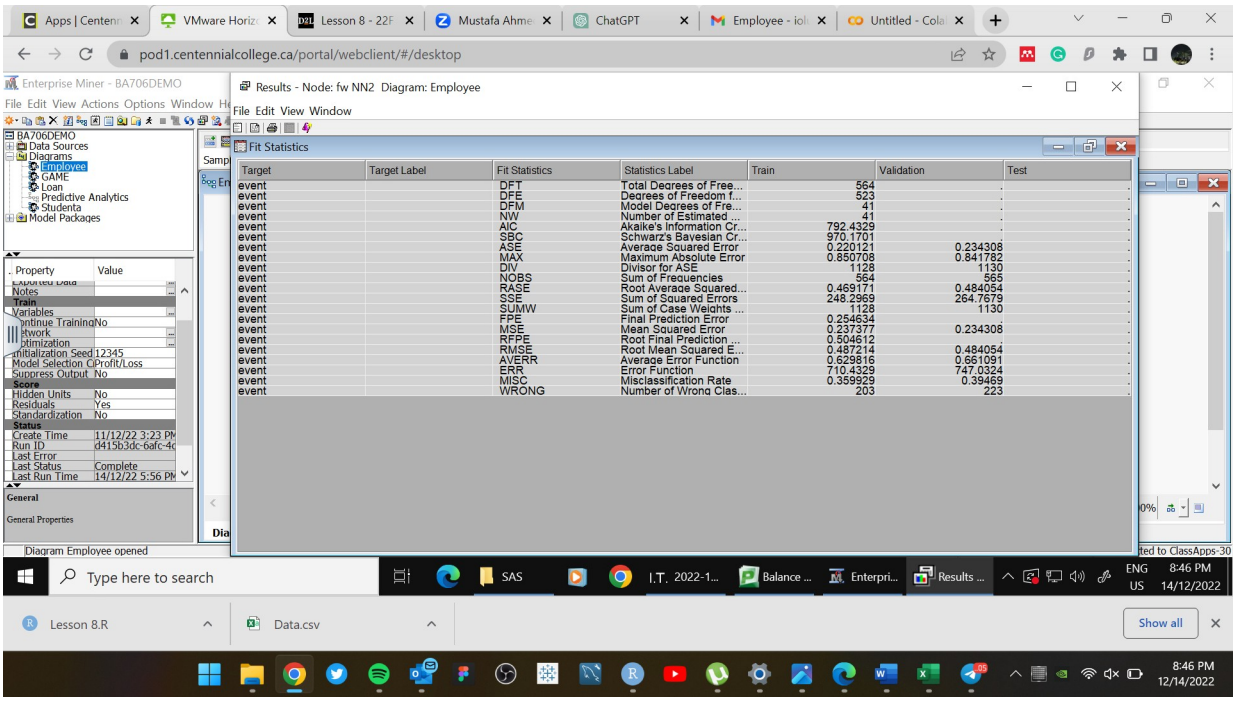
NEURAL NETWORK WITH 2 HIDDEN UNITS

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

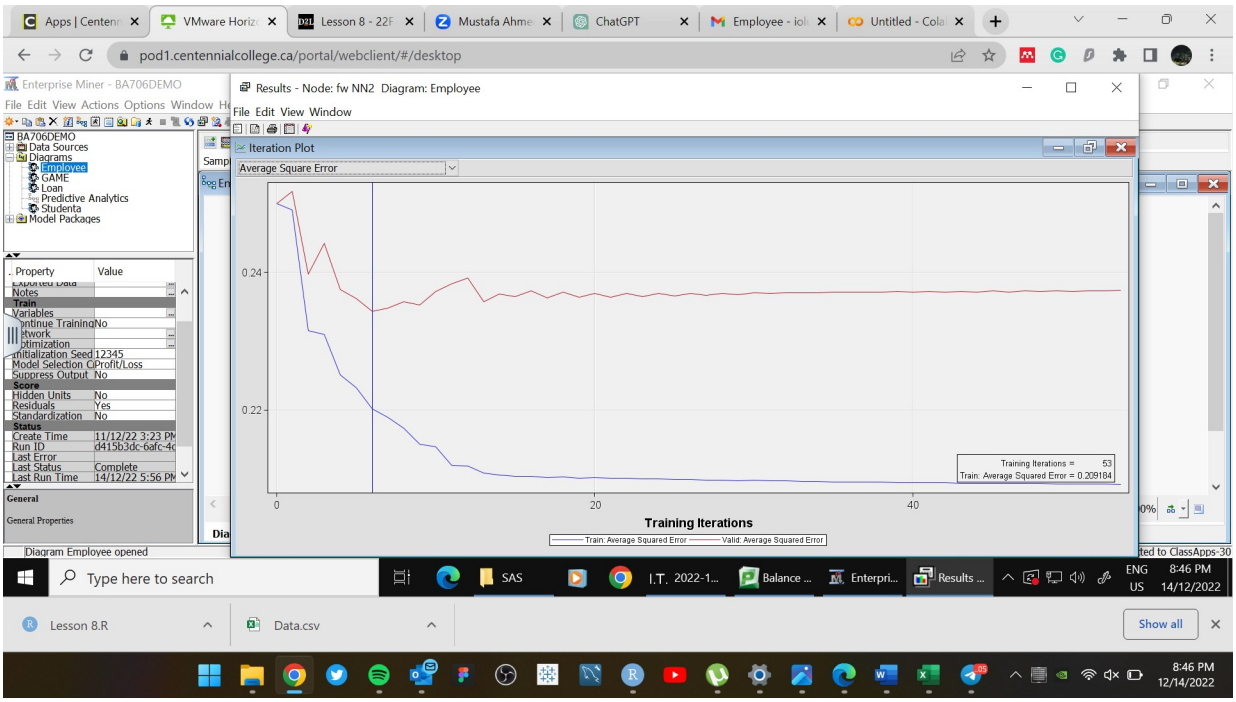



Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Employee
Network
Property
Value
Architecture
Multilayer Perceptron
Direct Connection
No
Number of Hidden Units
2
Randomization Distribution
Normal
Randomization Center
0.0
Randomization Scale
0.1
Input Standardization
Standard Deviation
Hidden Layer Combination Function
Default
Hidden Layer Activation Function
Default
Hidden Bias
Yes
Target Layer Combination Function
Default
Target Layer Activation Function
Default
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
File Import
Data Partition
Model Comparison
Diagram
Log
Type here to search
SAS
I.T. 2022-12-0...
Balance for ...
Enterprise ...
ENG
US
6:18 PM
12/12/2022


Apps | Centenn
VMware Horiz
Lesson 8 - 22F
Mustafa Ahme
ChatGPT
Employee - iol
Untitled - Cola
pod1.centennialcollege.ca/portal/webclient/#/desktop
Enterprise Miner - BA706DEMO
Results - Node: fw NN2 Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
Iteration Plot
Average Square Error
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error
Output
User: 301259629
Date: December 14, 2022
Time: 17:56:52
* Training Output
Variable Summary
Measurement Level
Frequency Count
Role
Diagram Employee opened
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG
US
8:45 PM
14/12/2022
Lesson 8.R
Data.csv
Show all
8:45 PM
12/14/2022

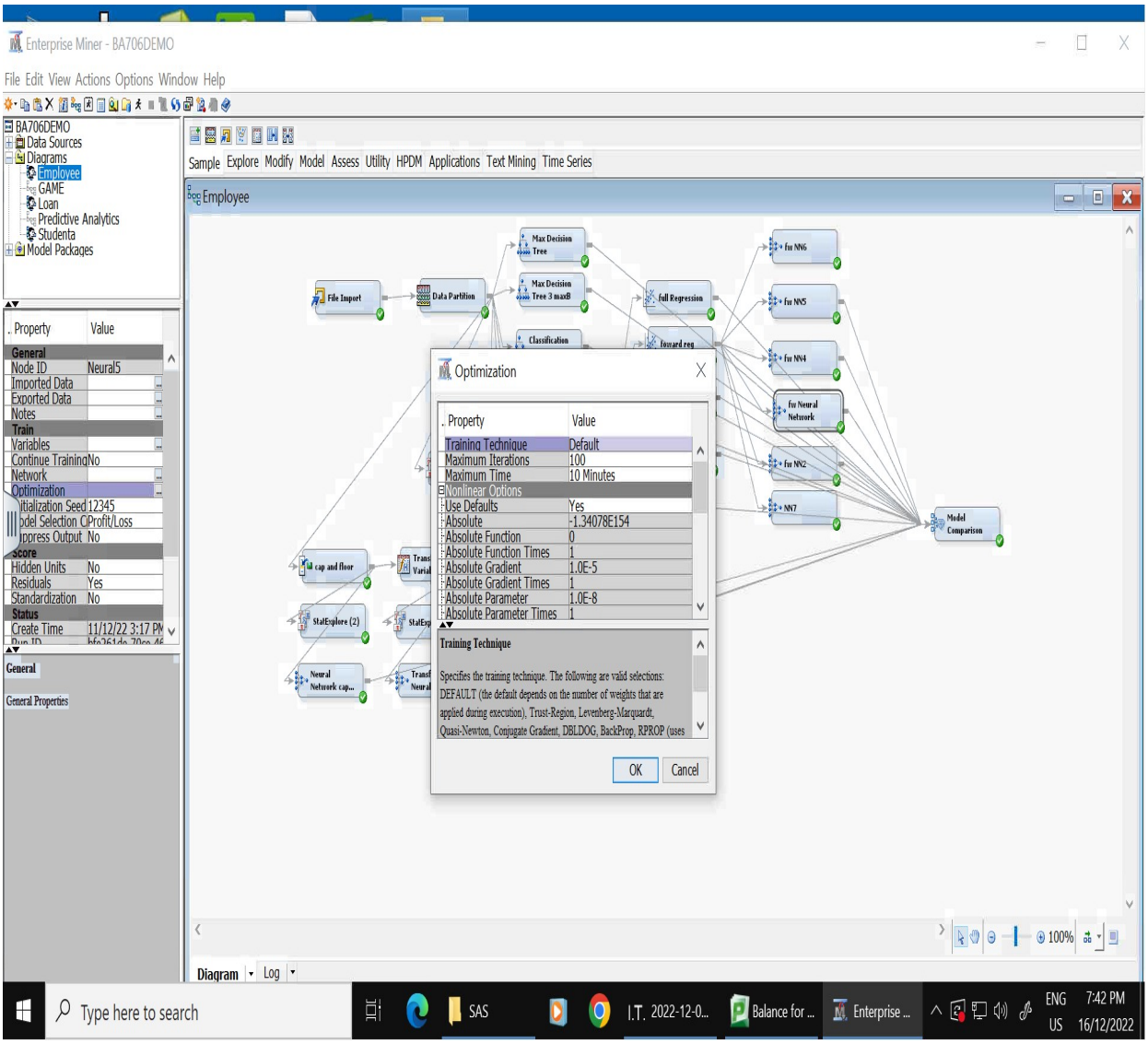


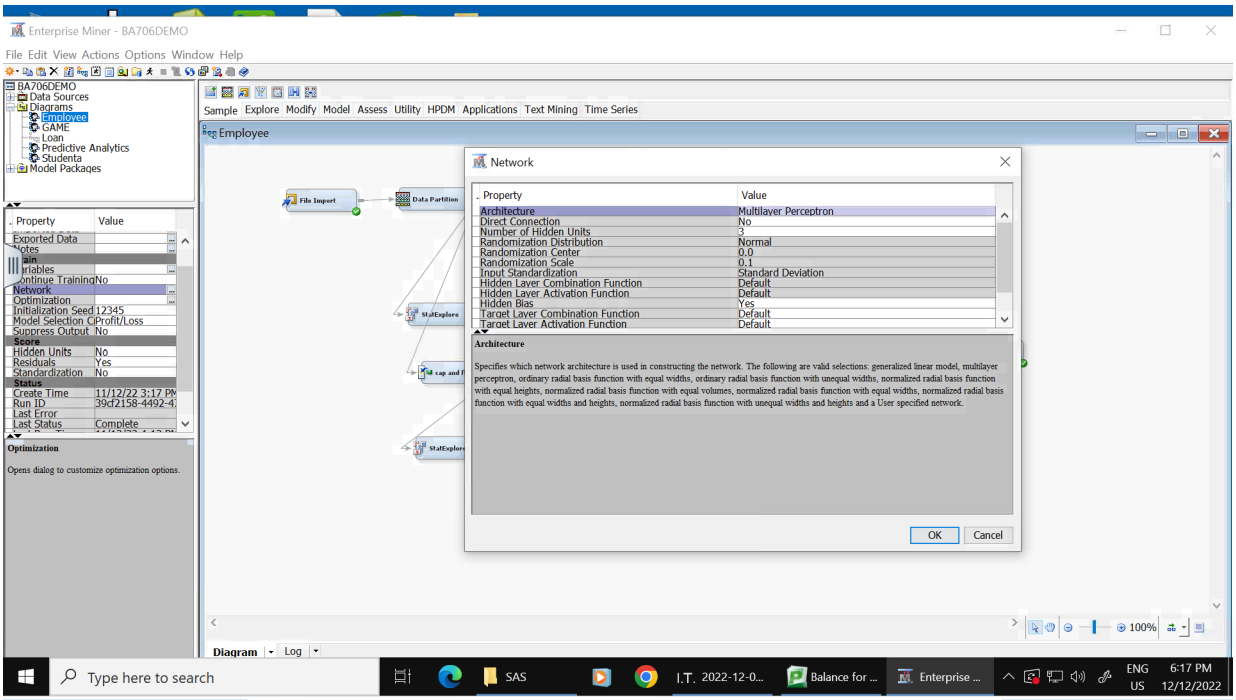
Here, we see that the average squared error of the training and validation data sets start off in opposite directions with the validation getting worse before it changes direction for improvement. The most improvement for the validation ASE was at iteration 6 then a slight increase occurred which is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

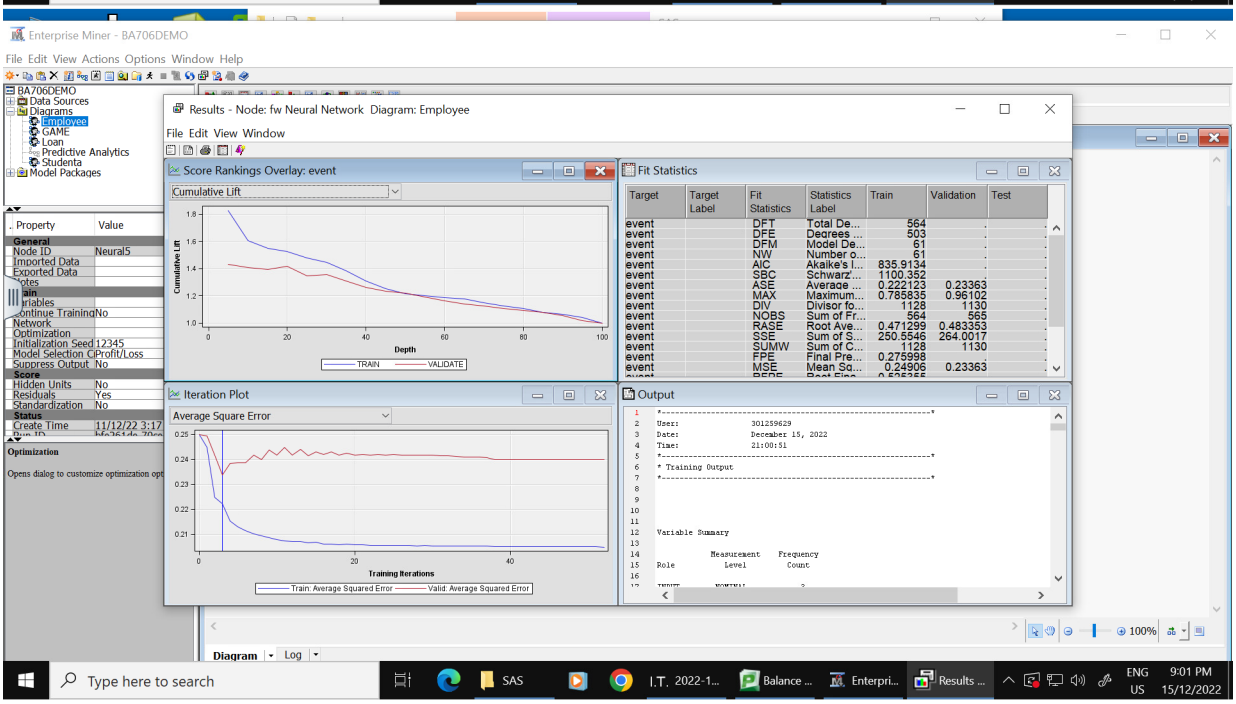


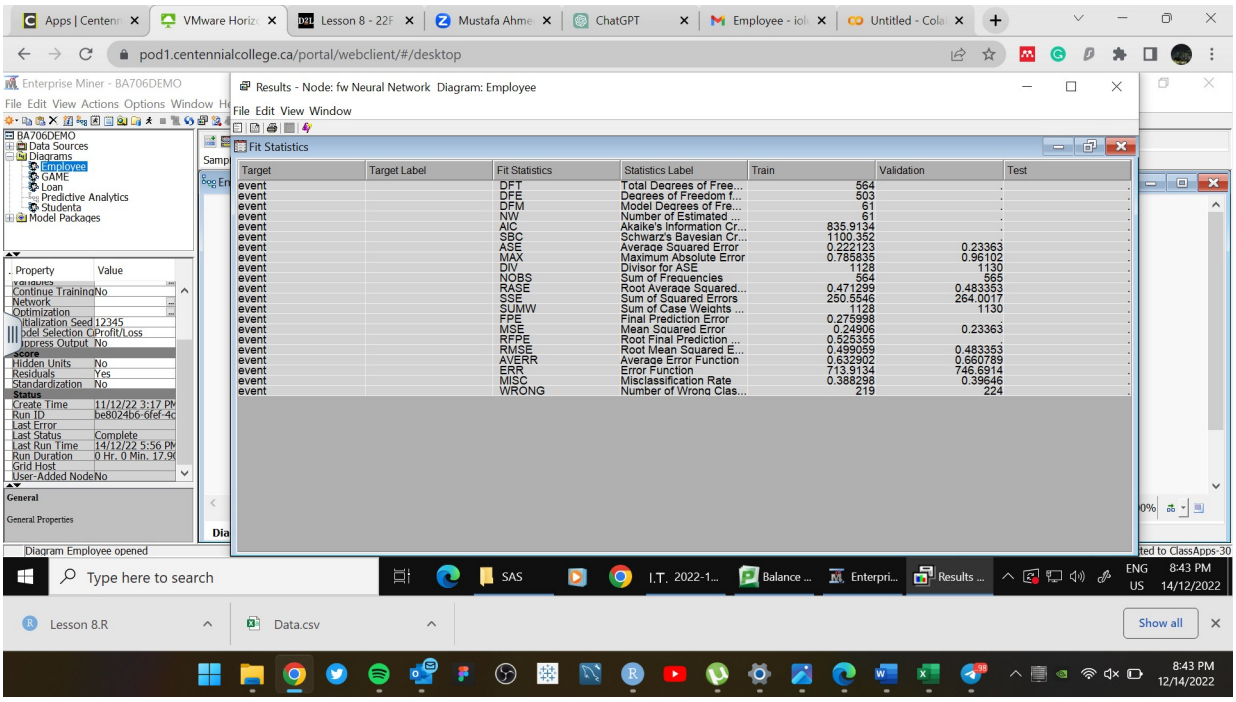
NEURAL NETWORK WITH 3 HIDDEN UNITS

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.

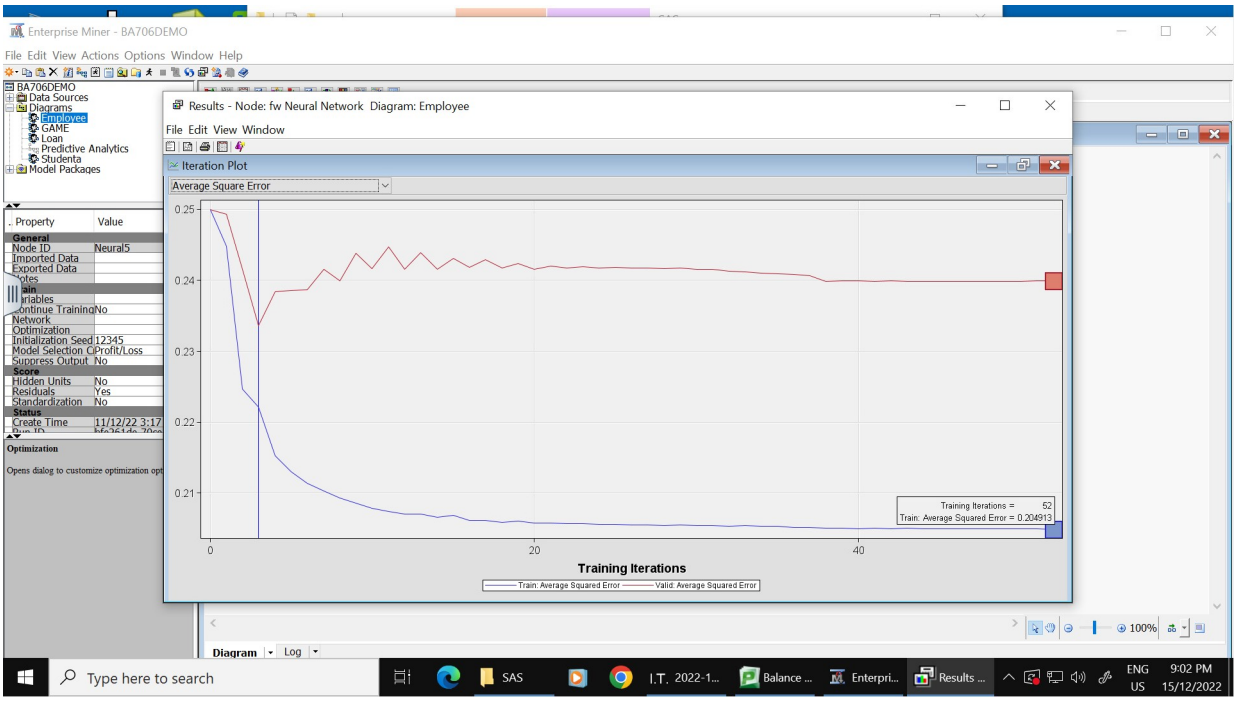

Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Employee
Optimization
. Property
Value
Training Technique
Default
Maximum Iterations
100
Maximum Time
10 Minutes
Nonlinear Options
Use Defaults
Yes
Absolute
-1.34078E154
Absolute Function
0
Absolute Function Times
1
Absolute Gradient
1.0E-5
Absolute Gradient Times
1
Absolute Parameter
1.0E-8
Absolute Parameter Times
1
Training Technique
Specifies the training technique. The following are valid selections: DEFAULT (the default depends on the number of weights that are applied during execution), Trust-Region, Levenberg-Marquardt, Quasi-Newton, Conjugate Gradient, DBLDOG, BackProp, RPROP (uses
OK
Cancel
Max Decision Tree
Max Decision Tree 3 maxB
Classification
File Import
Data Partition
full Regression
fw NN6
fw NN5
fw NN4
fw Neural Network
fw NN2
NN7
Model Comparison
cap and floor
StatExplore (2)
Neural Network cap...
Diagram
Log
Type here to search
SAS
I.T. 2022-12-0...
Balance for ...
Enterprise ...
ENG
US
7:42 PM
16/12/2022


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Employee
Network
. Property
Value
Architecture
Multilayer Perceptron
Direct Connection
No
Number of Hidden Units
3
Randomization Distribution
Normal
Randomization Center
0.0
Randomization Scale
0.1
Input Standardization
Standard Deviation
Hidden Layer Combination Function
Default
Hidden Layer Activation Function
Default
Hidden Bias
Yes
Target Layer Combination Function
Default
Target Layer Activation Function
Default
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
Optimization
Opens dialog to customize optimization options.
Diagram
Log
100%
Type here to search
SAS
I.T. 2022-12-0...
Balance for ...
Enterprise ...
ENG
US
6:17 PM
12/12/2022


Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
Results - Node: fw Neural Network Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
Iteration Plot
Average Square Error
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error
Output
User: 301259629
Date: December 15, 2022
Time: 21:00:51
* Training Output
Variable Summary
Measurement Frequency
Role Level Count
Node ID Neural5
Network
Optimization
Initialization Seed 12345
Model Selection Profit/Loss
Suppress Output No
Hidden Units No
Residuals Yes
Standardization No
Create Time 11/12/22 3:17
Optimization
Diagram
Log
100%
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG
US
9:01 PM
15/12/2022

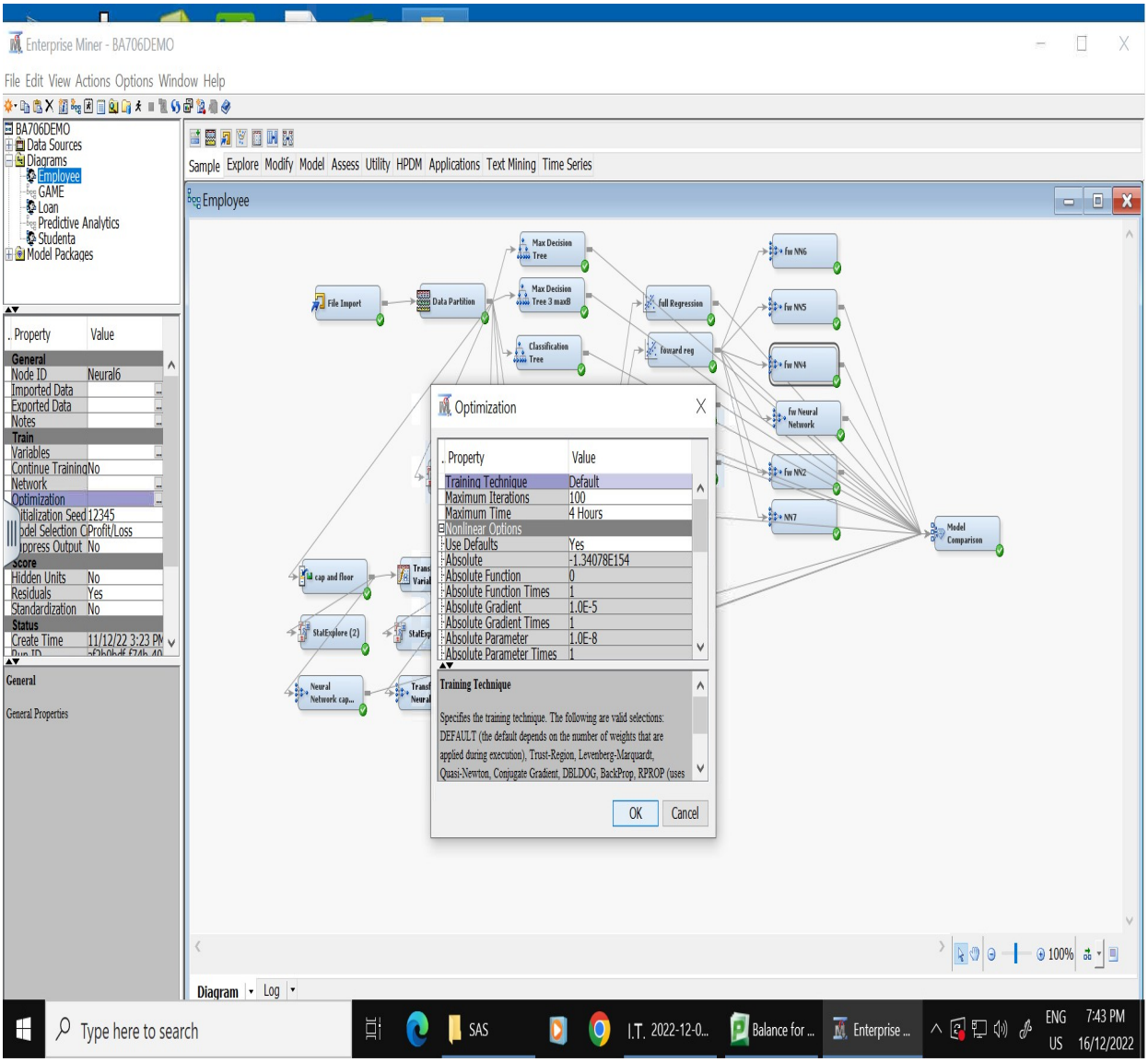


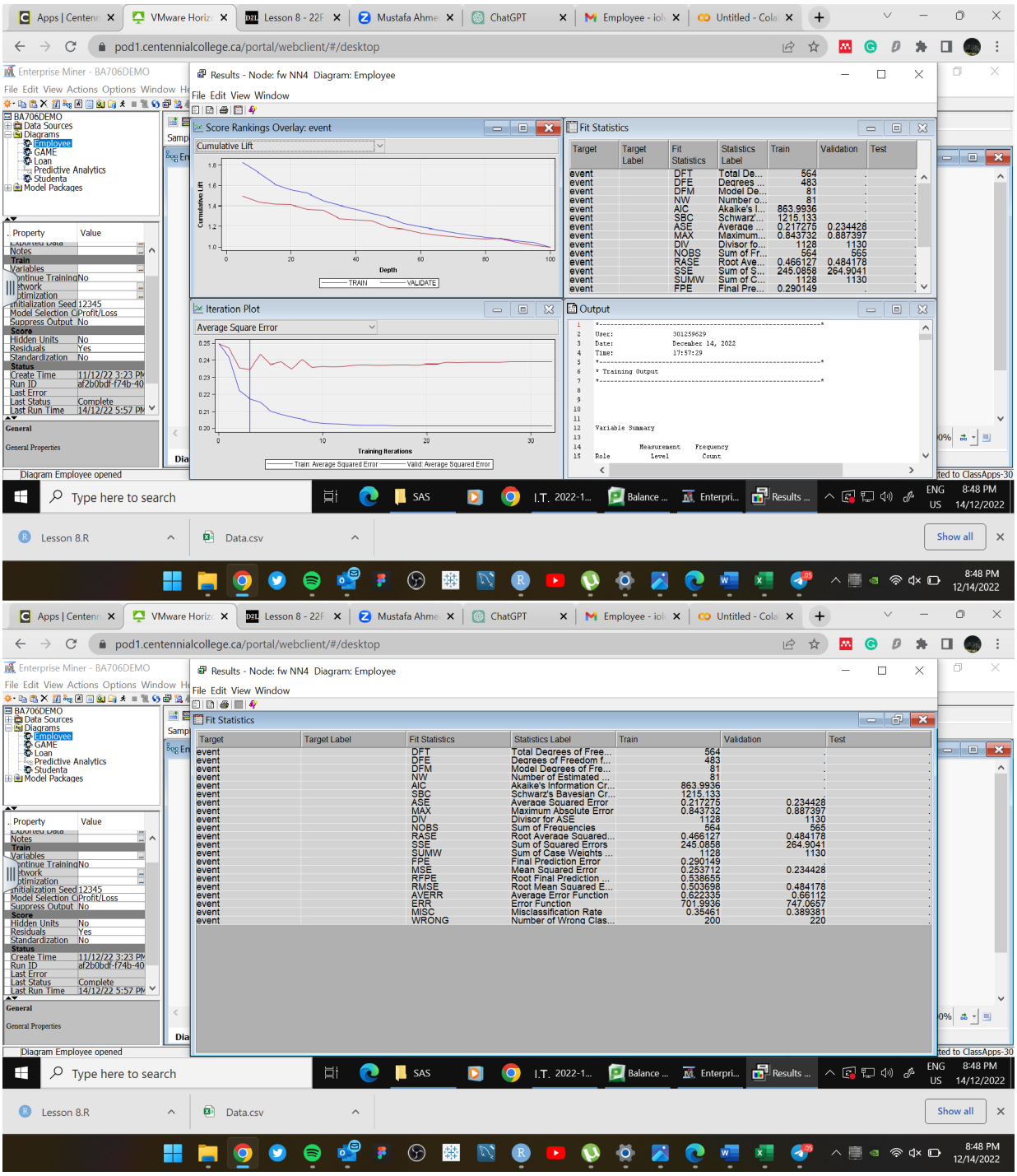
Here, we see that the average squared error of the training and validation data sets start off in the same direction of improvement till iteration 3 where the validation ASE starts to increase as a sign of overfitting. this occurs until the point of convergence at iteration 52.



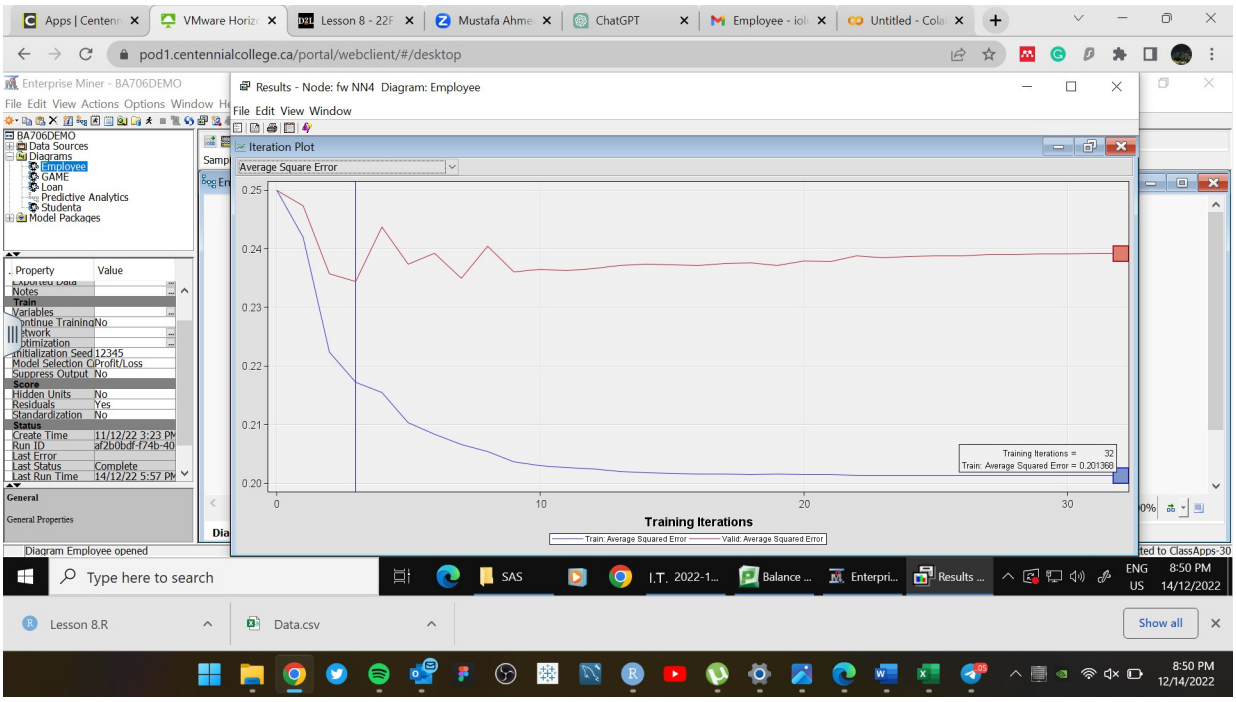
NEURAL NETWORK WITH 4 HIDDEN UNITS

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 os the two hidden units with an ASE of 0.234428 .





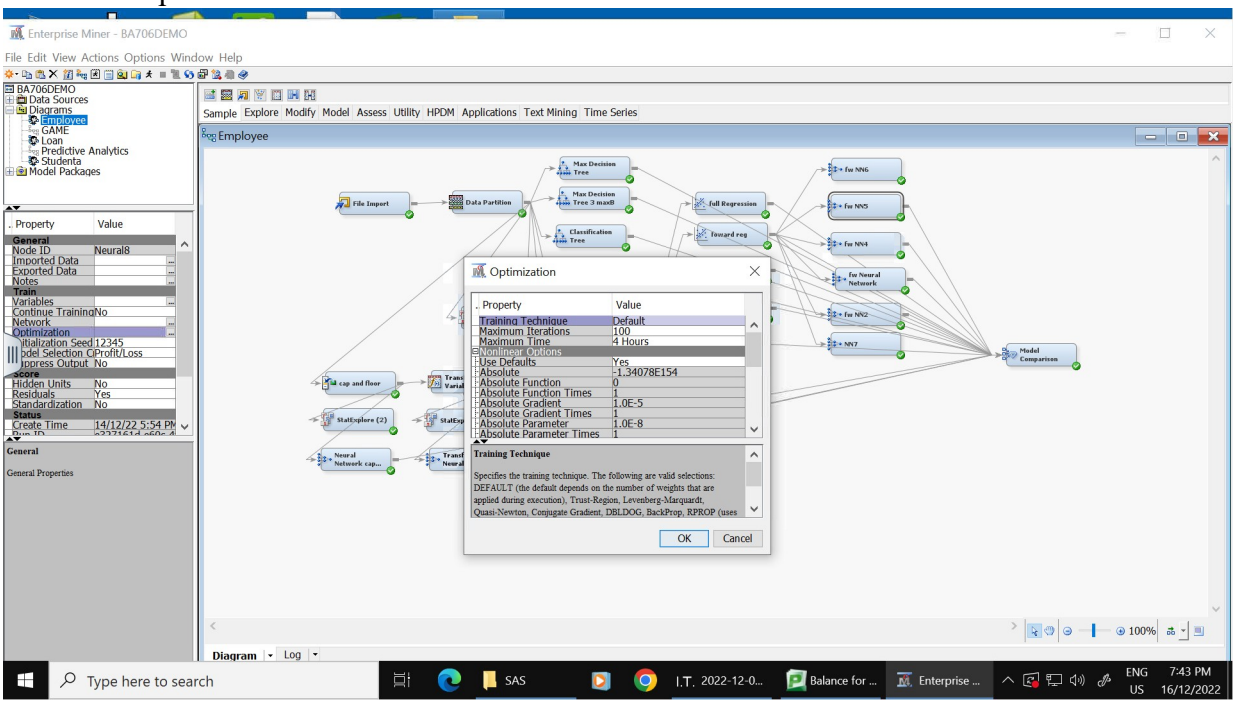
Here, we see that the average squared error of the training and validation data sets start off in the same direction of improvement and then the validation ASE starts to increase as a sign of overfitting. This occurs until the point of convergence at iteration 32.

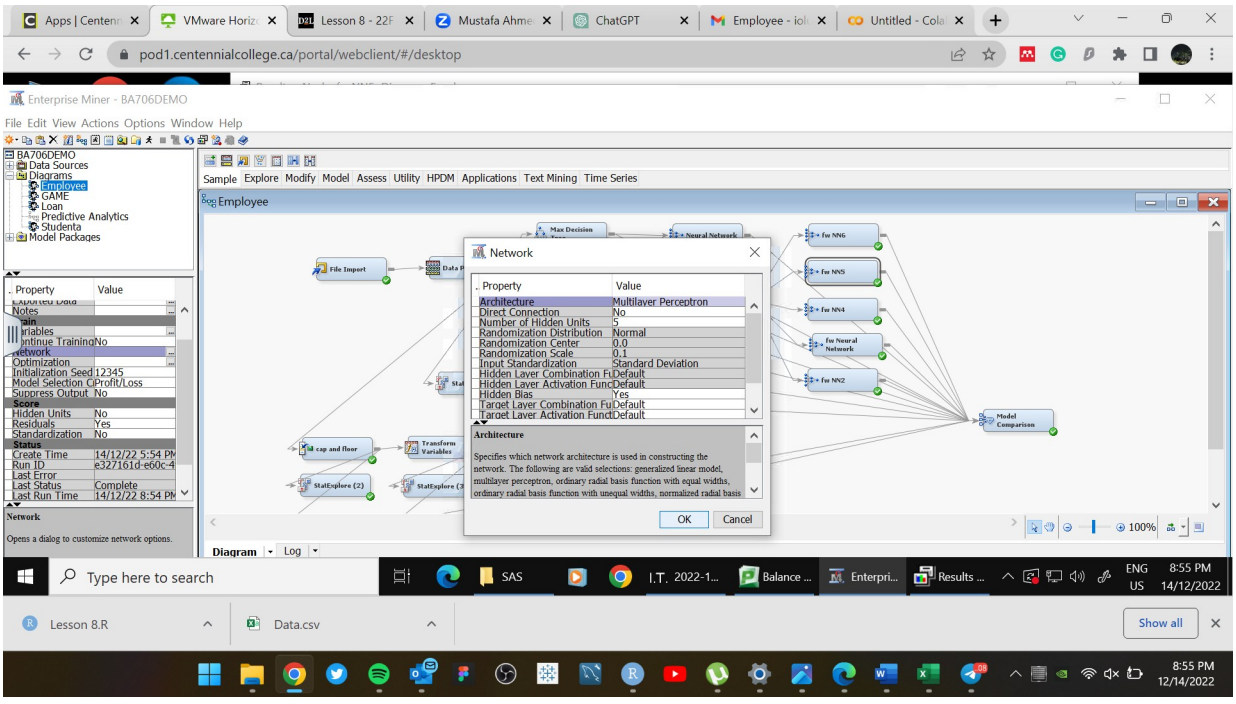


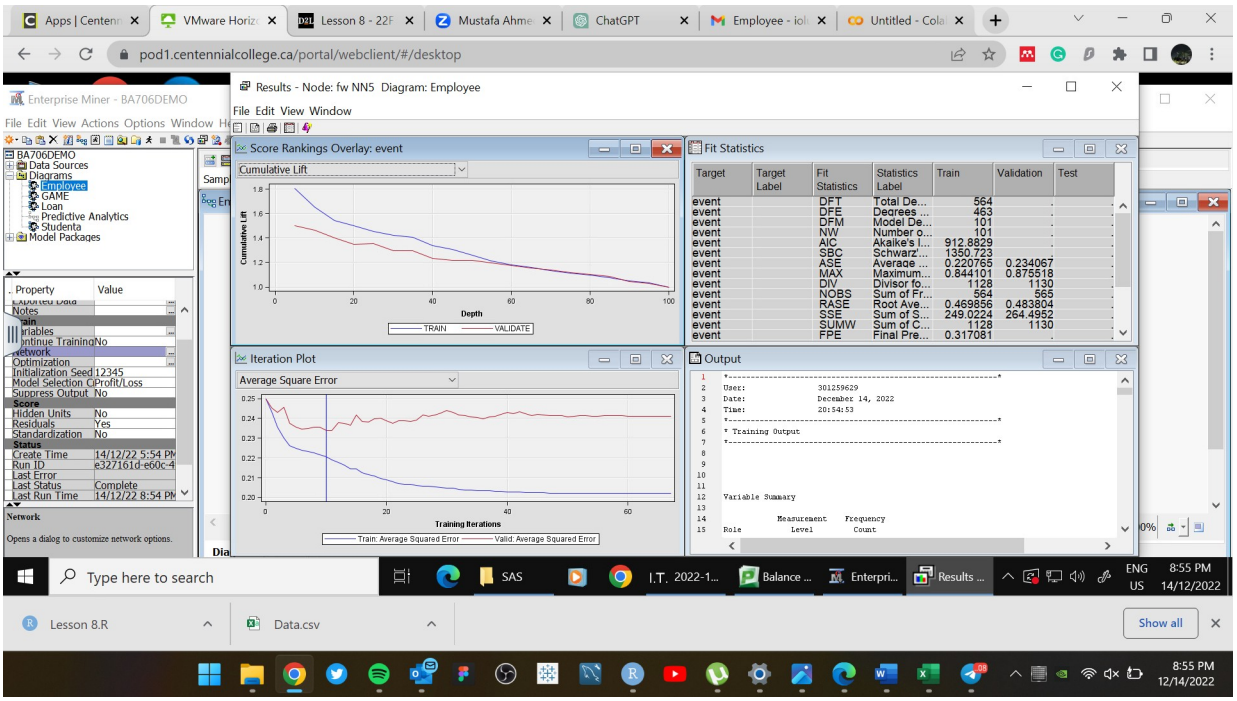
NEURAL NETWORK WITH 5 HIDDEN UNITS

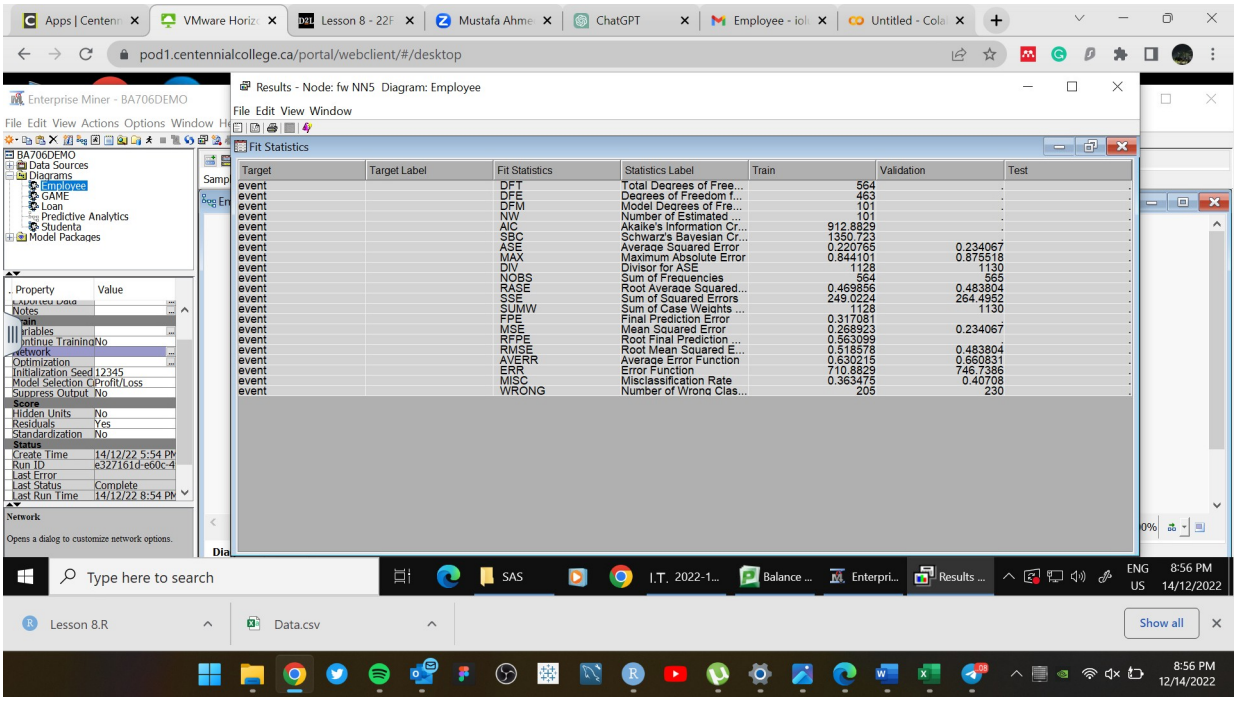
We further increased the number of hidden units to 5 and the outcome ASE of 0.234067 was an improvement on the two hidden unit neural network but not as good as the three hidden unit neural network. This proves no further need to run additional hidden units as the aim is to find

the leat complex model with the best result.

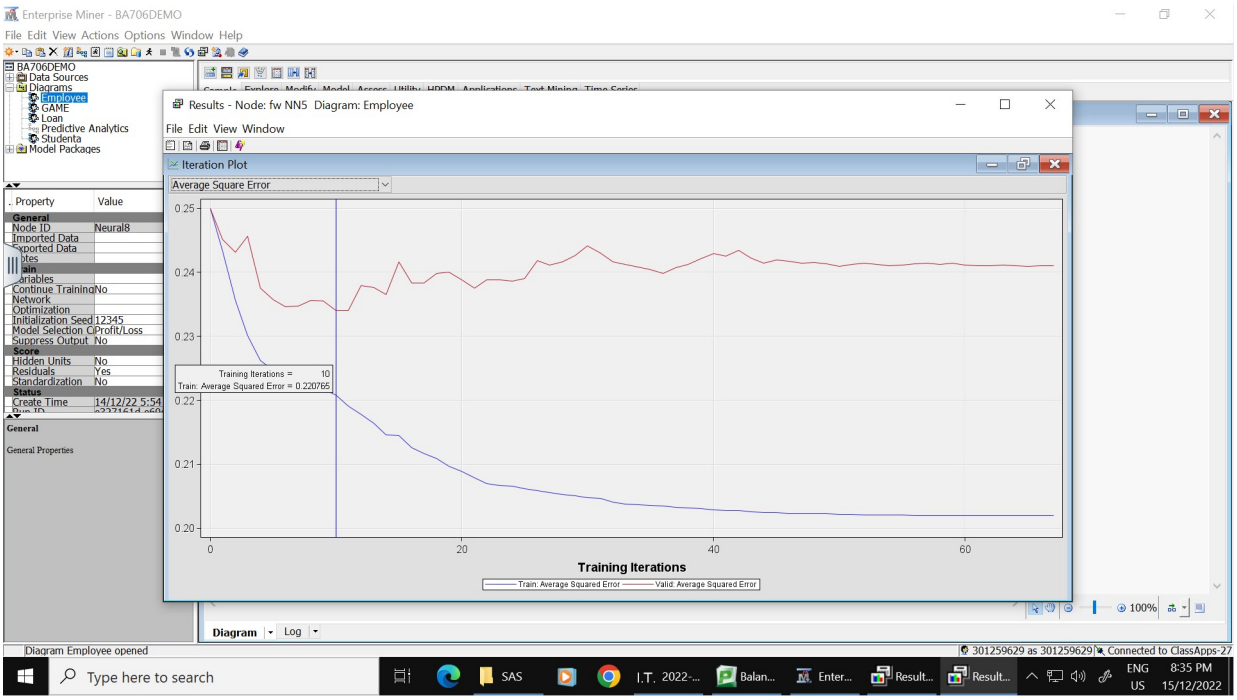


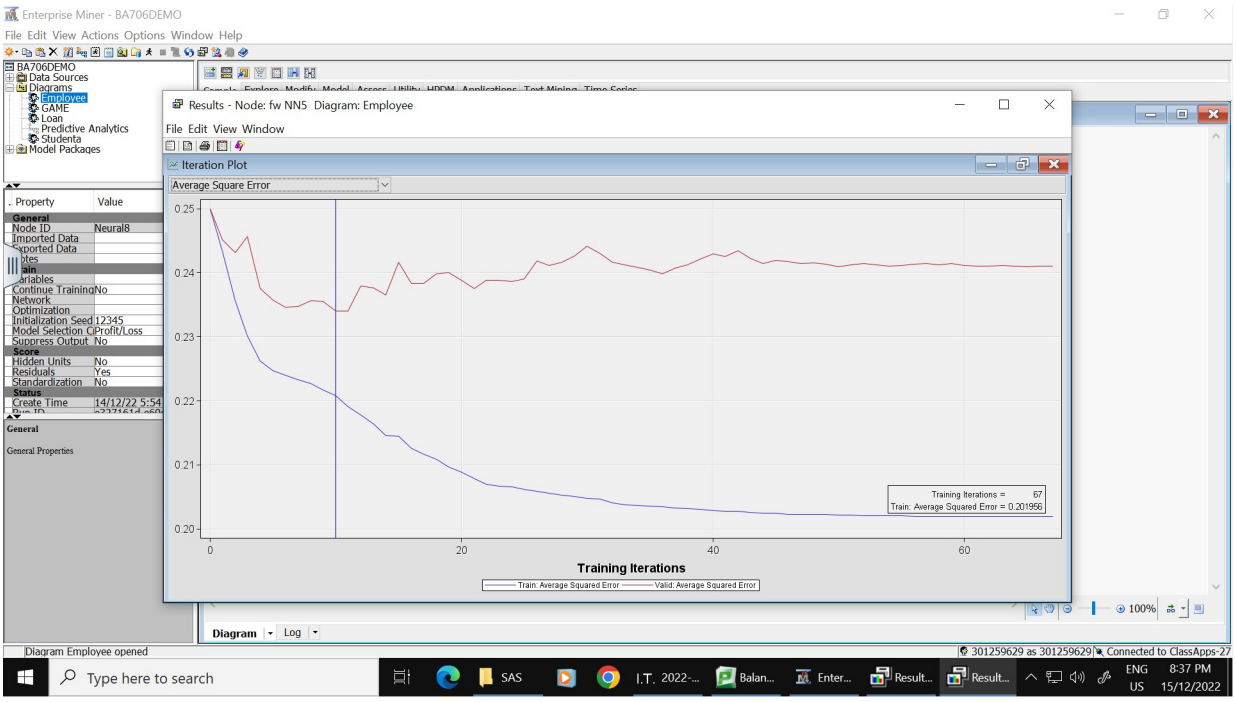



Results - Node: fw NN5 Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event DFT Total De... 564
event DFE Degrees ... 463
event DFM Model De... 101
event NW Number o... 101
event AIC Akaike's I... 912.8829
event SBC Schwarz'... 1350.723
event ASE Average ... 0.220765 0.234067
event MAX Maximum... 0.844101 0.875518
event DIV Divisor fo... 1128 1130
event NOBS Sum of Fr... 564 565
event RASE Root Ave... 0.469856 0.483804
event SSE Sum of S... 249.0224 264.4952
event SUMW Sum of C... 1128 1130
event FPE Final Pre... 0.317081
Iteration Plot
Average Square Error
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error
Output
User: 301259629
Date: December 14, 2022
Time: 20:54:53
* Training Output
Variable Summary
Measurement Frequency
Role Level Count
Enterprise Miner - BA706DEMO
File Edit View Actions Options Window
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property Value
Network
Opens a dialog to customize network options.
Type here to search
Lesson 8.R
Data.csv
Show all
8:55 PM
12/14/2022


Results - Node: fw NN5 Diagram: Employee
File Edit View Window
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
event DFT Total Degrees of Free... 564
event DFE Degrees of Freedom f... 463
event DFM Model Degrees of Fre... 101
event NW Number of Estimated ... 101
event AIC Akaike's Information Cr... 912.8829
event SBC Schwarz's Bayesian Cr... 1350.723
event ASE Average Squared Error 0.220765 0.234067
event MAX Maximum Absolute Error 0.844101 0.875518
event DIV Divisor for ASE 1128 1130
event NOBS Sum of Frequencies 564 565
event RASE Root Average Squared... 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 MSE Mean Squared Error 0.268923 0.234067
event RFPE Root Final Prediction ... 0.563099
event RMSE Root Mean Squared E... 0.518578 0.483804
event AVERR Average Error Function 0.630215 0.660831
event ERR Error Function 710.8829 746.7386
event MISC Misclassification Rate 0.363475 0.40708
event WRONG Number of Wrong Clas... 205 230
Enterprise Miner - BA706DEMO
File Edit View Actions Options Window
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property Value
Network
Opens a dialog to customize network options.
Type here to search
Lesson 8.R
Data.csv
Show all
8:56 PM
12/14/2022

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.

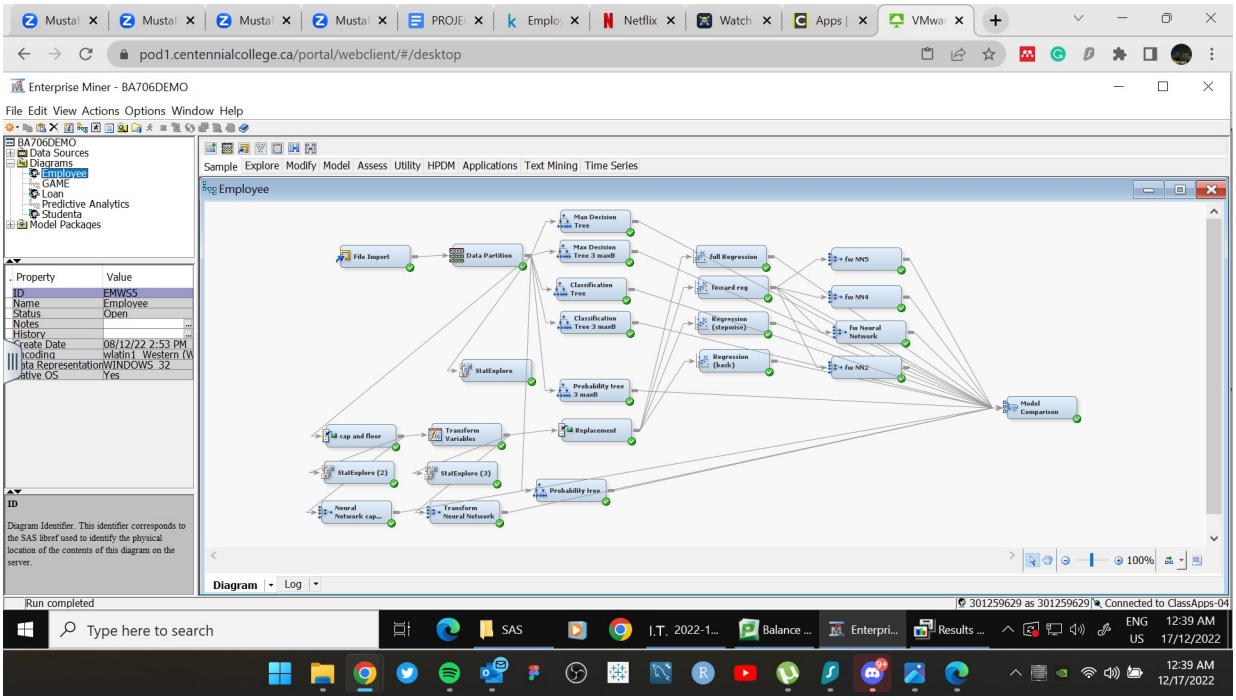



Enterprise Miner - BA706DEMO
File Edit View Actions Options Window Help
BA706DEMO
Data Sources
Diagrams
Employee
GAME
Loan
Predictive Analytics
Studenta
Model Packages
Property
Value
Results - Node: fw NN5 Diagram: Employee
File Edit View Window
Iteration Plot
Average Square Error
0.25
0.24
0.23
0.22
0.21
0.20
0
20
40
60
Training Iterations
Training Iterations = 67
Train: Average Squared Error = 0.201956
Train: Average Squared Error
Valid: Average Squared Error
Diagram
Log
100%
Diagram Employee opened
Connected to ClassApps-27
Type here to search
ENG
US
8:37 PM
15/12/2022

ASSESSMENT

MODEL COMPARISON

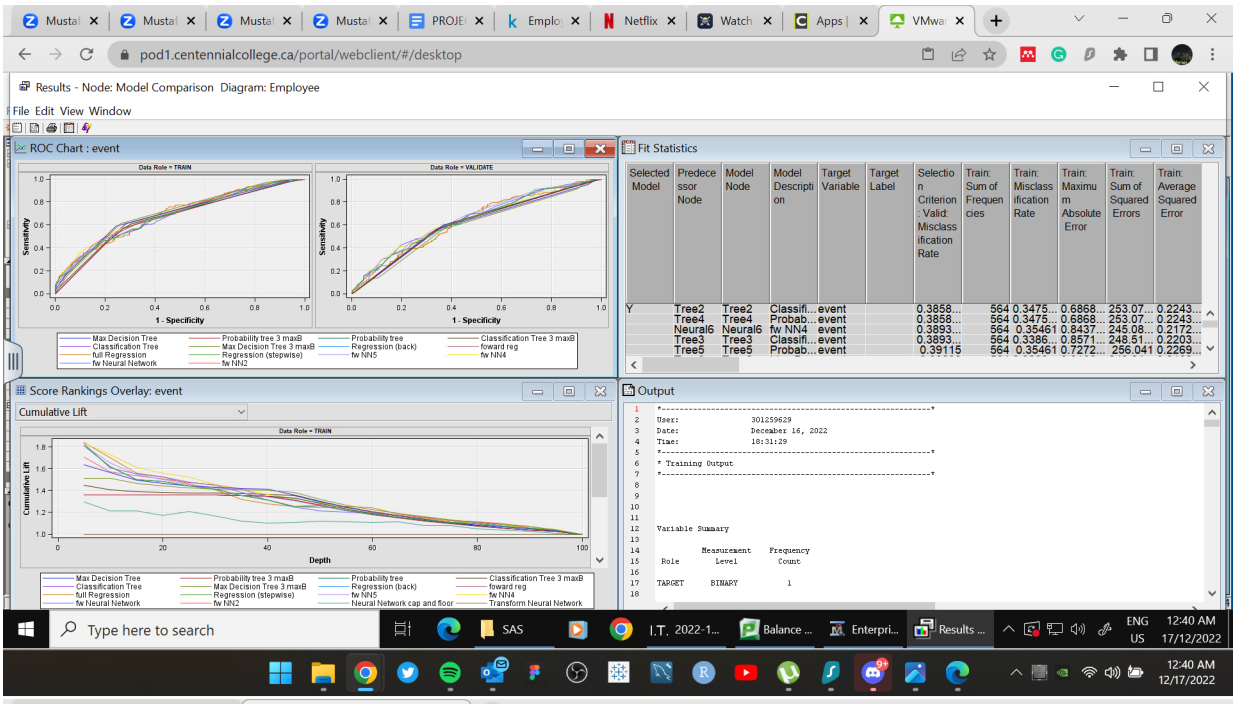
Finally, a model comparison node was introduced to the diagram workspace to which the different decision trees, regressions, and neural network models were connected. The results show a comparison of all models using several statistical tests. The below shows the final diagram of all the models we tested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

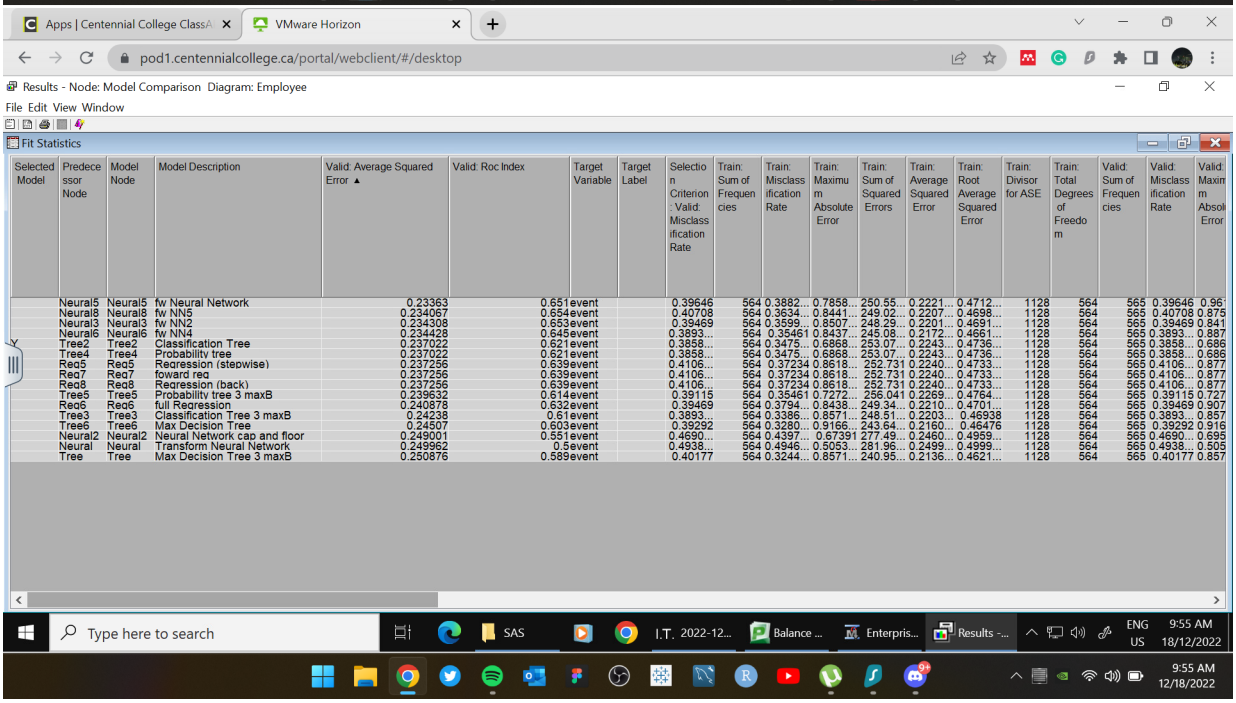


As stated earlier, we will be selecting the most fit model using the Average squared error. Below shows the summary of all models the Average squared error and ROC index. Using the ROC index as a selection criterion, the best model would be the Neural Network with 5 hidden units and ROC index of 0.654. 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
Neural Network 3 hidden units	**0.233630**	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


pod1.centennialcollege.ca/portal/webclient/#/desktop
Results - Node: Model Comparison Diagram: Employee
File Edit View Window
ROC Chart : event
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
Max Decision Tree
Classification Tree
full Regression
fw Neural Network
Probability tree 3 maxB
Max Decision Tree 3 maxB
Regression (stepwise)
fw NN2
Probability tree
Regression (back)
fw NN5
Classification Tree 3 maxB
foward reg
fw NN4
Fit Statistics
Score Rankings Overlay: event
Cumulative Lift
Depth
Output
Variable Summary
Measurement Frequency
Role Level Count
TARGET BINARY 1
Type here to search
SAS
I.T. 2022-1...
Balance ...
Enterpri...
Results ...
ENG US
12:40 AM 17/12/2022
12:40 AM 12/17/2022


Apps | Centennial College Class
VMware Horizon
pod1.centennialcollege.ca/portal/webclient/#/desktop
Results - Node: Model Comparison Diagram: Employee
File Edit View Window
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Valid: Average Squared Error
Valid: Roc Index
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Sum of Frequencies
Train: Misclassification Rate
Train: Maximum Absolute Error
Train: Sum of Squared Errors
Train: Average Squared Error
Train: Root Average Squared Error
Train: Divisor for ASE
Train: Total Degrees of Freedom
Valid: Sum of Frequencies
Valid: Misclassification Rate
Valid: Maximum Absolute Error
Neural5 Neural5 fw Neural Network 0.23363 0.651 event 0.39646 564 0.3882... 0.7858... 250.55... 0.2221... 0.4712... 1128 564 565 0.39646 0.96
Neural8 Neural8 fw NN5 0.234067 0.654 event 0.40708 564 0.3634... 0.8441... 249.02... 0.2207... 0.4698... 1128 564 565 0.40708 0.875
Neural3 Neural3 fw NN2 0.234308 0.653 event 0.39469 564 0.3599... 0.8507... 248.29... 0.2201... 0.4691... 1128 564 565 0.39469 0.841
Neural6 Neural6 fw NN4 0.234428 0.645 event 0.3893... 564 0.35461 0.8437... 245.08... 0.2172... 0.4661... 1128 564 565 0.3893... 0.887
Tree2 Tree2 Classification Tree 0.237022 0.621 event 0.3858... 564 0.3475... 0.6868... 253.07... 0.2243... 0.4736... 1128 564 565 0.3858... 0.686
Tree4 Tree4 Probability tree 0.237022 0.621 event 0.3858... 564 0.3475... 0.6868... 253.07... 0.2243... 0.4736... 1128 564 565 0.3858... 0.686
Reg5 Reg5 Regression (stepwise) 0.237256 0.639 event 0.4106... 564 0.37234 0.8618... 252.731 0.2240... 0.4733... 1128 564 565 0.4106... 0.877
Reg7 Reg7 foward reg 0.237256 0.639 event 0.4106... 564 0.37234 0.8618... 252.731 0.2240... 0.4733... 1128 564 565 0.4106... 0.877
Reg8 Reg8 Regression (back) 0.237256 0.639 event 0.4106... 564 0.37234 0.8618... 252.731 0.2240... 0.4733... 1128 564 565 0.4106... 0.877
Tree5 Tree5 Probability tree 3 maxB 0.239632 0.614 event 0.39115 564 0.35461 0.7272... 256.041 0.2269... 0.4764... 1128 564 565 0.39115 0.727
Reg6 Reg6 full Regression 0.240878 0.632 event 0.39469 564 0.3794... 0.8438... 249.34... 0.2210... 0.4701... 1128 564 565 0.39469 0.907
Tree3 Tree3 Classification Tree 3 maxB 0.24238 0.61 event 0.3893... 564 0.3386... 0.8571... 248.51... 0.2203... 0.46938 1128 564 565 0.3893... 0.857
Tree6 Tree6 Max Decision Tree 0.24507 0.603 event 0.39292 564 0.3280... 0.9166... 243.64... 0.2160... 0.46476 1128 564 565 0.39292 0.916
Neural2 Neural2 Neural Network cap and floor 0.249001 0.551 event 0.4690... 564 0.4397... 0.67391 277.49... 0.2460... 0.4959... 1128 564 565 0.4690... 0.695
Neural Neural Transform Neural Network 0.249962 0.5 event 0.4938... 564 0.4946... 0.5053... 281.96... 0.2499... 0.4999... 1128 564 565 0.4938... 0.505
Tree Tree Max Decision Tree 3 maxB 0.250876 0.589 event 0.40177 564 0.3244... 0.8571... 240.95... 0.2136... 0.4621... 1128 564 565 0.40177 0.857
Type here to search
SAS
I.T. 2022-12...
Balance ...
Enterpris...
Results -...
ENG US
9:55 AM 18/12/2022
9:55 AM 12/18/2022


pod1.centennialcollege.ca/portal/webclient/#/desktop
Results - Node: Model Comparison Diagram: Employee
File Edit View Window
ROC Chart : event
Data Role = VALIDATE
Sensitivity
1 - Specificity
Type here to search


pod1.centennialcollege.ca/portal/webclient/#/desktop
Results - Node: Model Comparison Diagram: Employee
File Edit View Window
Score Rankings Overlay: event
Cumulative Lift
Data Role = VALIDATE
Cumulative Lift
Depth
Type here to search

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

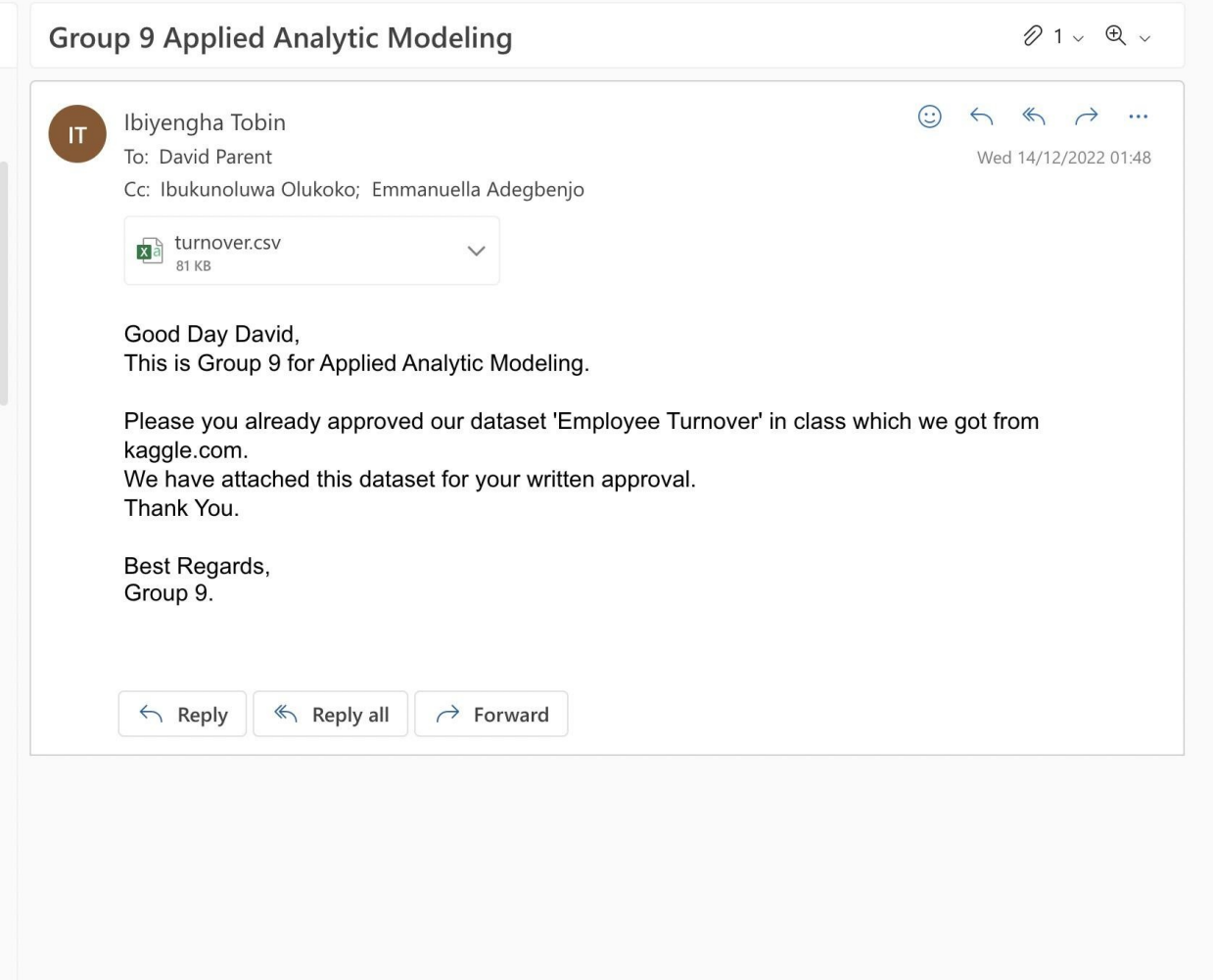
banks, consult, and state are 4.6, 4.,4 and 3.1 times respectively more likely to quit their jobs when compared with those in the transport industry. There might also be a need to study the transport industry to see why it seems more stable and able to retain its employees.

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

1

IT

Ibiyengha Tobin

To: David Parent

Cc: Ibukunoluwa Olukoko; Emmanuella Adegbenjo

Wed 14/12/2022 01:48

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