MKTG 746 Big Data & Predictive Analytics

Group Project

IBM Employee Attrition Prediction

Group members

Gurpreet Singh (301346925)

Sourav (301417782)

Pankaj (301418786)

Chandan Suhruth (301322272)

Abhishek Kulkarni (301369504)

1. Executive Summary

A predictive analysis was conducted to understand and determine how these factors relate to workforce attrition. The prediction model was run on SAS Enterprise Miner. After data importing and wrangling, decision trees, logistic regressions, and neural networks were run. The best model was then identified by model comparison.

2. Introduction

Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

Our role is to uncover the factors that lead to employee attrition through Exploratory Data Analysis and explore them by using various classification models to predict if an employee is likely to quit. This could greatly increase the HR's ability to intervene on time and remedy the situation to prevent attrition.

While this model can be routinely run to identify employees who are most likely to quit, the key driver of success would be the human element of reaching out to the employee, understanding the current situation of the employee, and taking action to remedy controllable factors that can prevent attrition of the employee.

3. File Import

3.1. Data Source

The dataset used for this project was the "IBM Employee Attrition Prediction," obtained from Kaggle via this link. This is a hypothetical dataset created by IBM data scientists. The dataset has dimensions of (23436 rows x 37 columns) and contains numeric and categorical data types describing each employee's background and characteristics. It is labeled (supervised learning) with whether they are still in the company or whether they have gone to work somewhere else. Machine Learning models can help to understand and determine how these factors relate to workforce attrition.

3.2. Data Dictionary

The dataset contained 22 interval, 10 Nominal, 1 binary, 7 rejected

1. Age	The age of the employee.
2. Attrition	The target variable indicating whether the employee has left the company
3. BusinessTravel	Frequency of business travel, such as Travel_Rarely, Travel_Frequently, or Non-Travel.
4. DailyRate	The daily rate of pay for the employee.
5. Department	The department in which the employee works.
6. DistanceFromHome	The distance of the employee's residence from the workplace.
7. Education	The level of education attained by the employee (e.g., 1

8. EducationField	The field of education or specialization of the employee.
9. EmployeeCount	The count of employees, typically 1 for all rows.
10. EmployeeNumber	A unique identifier for each employee.
11. Application ID	A unique identifier for each job application.
12. EnvironmentSatisfaction	Satisfaction level of the employee with their work environment (e.g., 1
13. Gender	Gender of the employee (Male or Female).
14. HourlyRate	The hourly rate of pay for the employee.
15. JobInvolvement	Level of involvement of the employee in their job (e.g., 1
16. JobLevel	The level of the employee's job.
17. JobRole	The role of the employee in the company.
18. JobSatisfaction	Satisfaction level of the employee with their job (e.g., 1
19. MaritalStatus	Marital status of the employee (Single, Married, Divorced).
20. MonthlyIncome	The monthly income of the employee.
21. MonthlyRate	The monthly rate of pay for the employee.
22. NumCompaniesWorked	The number of companies the employee has worked for previously.
23. OverTime	Whether the employee works overtime (Yes or No).
24. PercentSalaryHike	The percentage increase in salary for the employee.
25. PerformanceRating	The performance rating of the employee (e.g., 3
26. RelationshipSatisfaction	Satisfaction level of the employee with their relationships at work (e.g., 1
27. StockOptionLevel	The level of stock options granted to the employee.
28. TotalWorkingYears	Total number of years the employee has been working.
29. TrainingTimesLastYear	Number of times the employee was trained last year.

30. WorkLifeBalance	Satisfaction level of the employee with their work-life balance (e.g., 1
31. YearsAtCompany	Number of years the employee has worked at the company.
32. YearsInCurrentRole	Number of years the employee has been in their current role.
33. YearsSinceLastPromotion	Number of years since the employee's last promotion.
34. YearsWithCurrManager	Number of years the employee has been with their current manager.
35. Employee Source	The source from which the employee was recruited.

3.3. File Import

Before importing the dataset into SAS Enterprise Miner, it had been realized that there were some "NA" entries in the Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, Application ID, EnvironmentSatisfaction, and Gender respectively. All NA entries were replaced by blanks in Microsoft Excel. Then the CSV file was imported into SAS Enterprise Miner with the File Import node.

Under Edit Variables of the File Import node, the level of each variable was selected to the respective data type. Since the objective of this analysis was to predict the Attrition in the company, it was a binary variable.

3.4. Data Leakage

EmployeeNumber or Application ID: If these columns are directly or indirectly related to the target variable (Attrition), such as indicating whether an employee has previously left the company, including them in the model could lead to data leakage.

To summarize we rejected Application id and employee number,
Column AI, AJ, AK, AL, AM are empty columns so rejected them too.

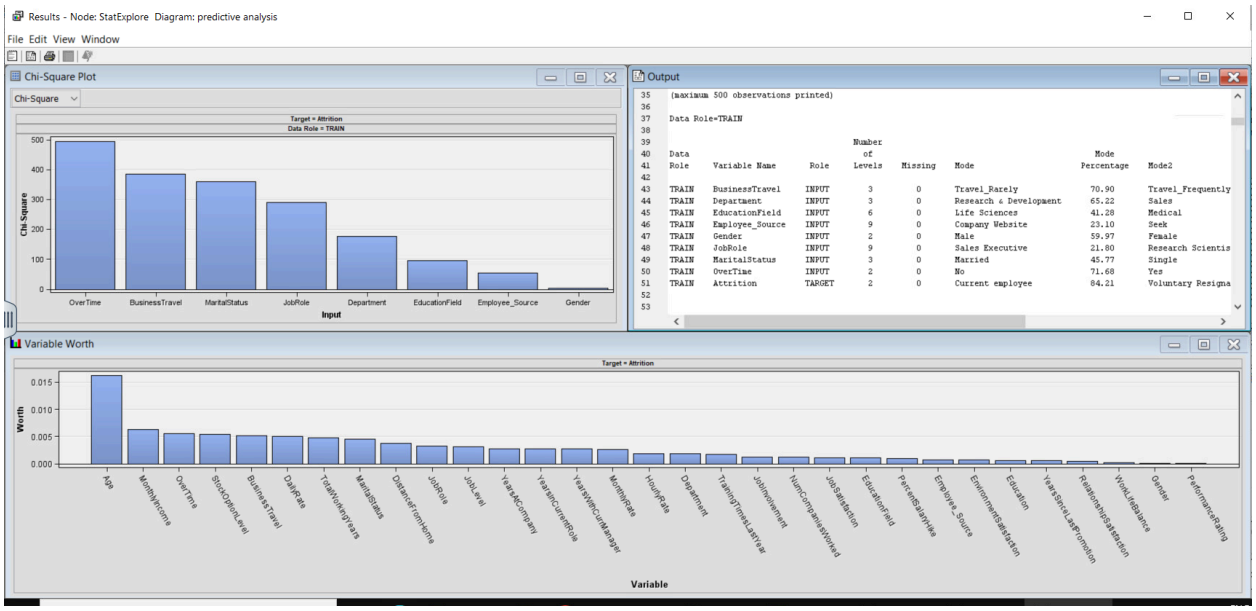
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AI	Rejected	Nominal	No		No	.	.
AJ	Rejected	Nominal	No		No	.	.
AK	Rejected	Nominal	No		No	.	.
AL	Rejected	Nominal	No		No	.	.
AM	Rejected	Nominal	No		No	.	.
Age	Input	Interval	No		No	.	.
Application_ID	Rejected	Nominal	No		No	.	.
Attrition	Target	Binary	No		No	.	.
BusinessTravel	Input	Nominal	No		No	.	.
DailyRate	Input	Interval	No		No	.	.
Department	Input	Nominal	No		No	.	.
DistanceFromHome	Input	Interval	No		No	.	.
Education	Input	Interval	No		No	.	.
EducationField	Input	Nominal	No		No	.	.
Employee_Source	Input	Nominal	No		No	.	.
EmployeeNumber	Rejected	Nominal	No		No	.	.
EnvironmentSatisfaction	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
HourlyRate	Input	Interval	No		No	.	.
JobInvolvement	Input	Interval	No		No	.	.
JobLevel	Input	Interval	No		No	.	.
JobRole	Input	Nominal	No		No	.	.
JobSatisfaction	Input	Interval	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
MonthlyIncome	Input	Interval	No		No	.	.
MonthlyRate	Input	Interval	No		No	.	.
NumCompaniesWorked	Input	Interval	No		No	.	.
OverTime	Input	Nominal	No		No	.	.
PercentSalaryHike	Input	Interval	No		No	.	.
PerformanceRating	Input	Interval	No		No	.	.
RelationshipSatisfaction	Input	Interval	No		No	.	.
StockOptionLevel	Input	Interval	No		No	.	.
TotalWorkingYears	Input	Interval	No		No	.	.
TrainingTimesLastYear	Input	Interval	No		No	.	.
WorkLifeBalance	Input	Interval	No		No	.	.
YearsAtCompany	Input	Interval	No		No	.	.
YearsInCurrentRole	Input	Interval	No		No	.	.
YearsSinceLastPromotion	Input	Interval	No		No	.	.
YearsWithCurrManager	Input	Interval	No		No	.	.

4. Data Wrangling

4.1. Data Filter

In order to figure out whether there were redundant or irrelevant variables, a StatExplore node was connected to the File Import node to further explore the variables.

STAT EXPLORE

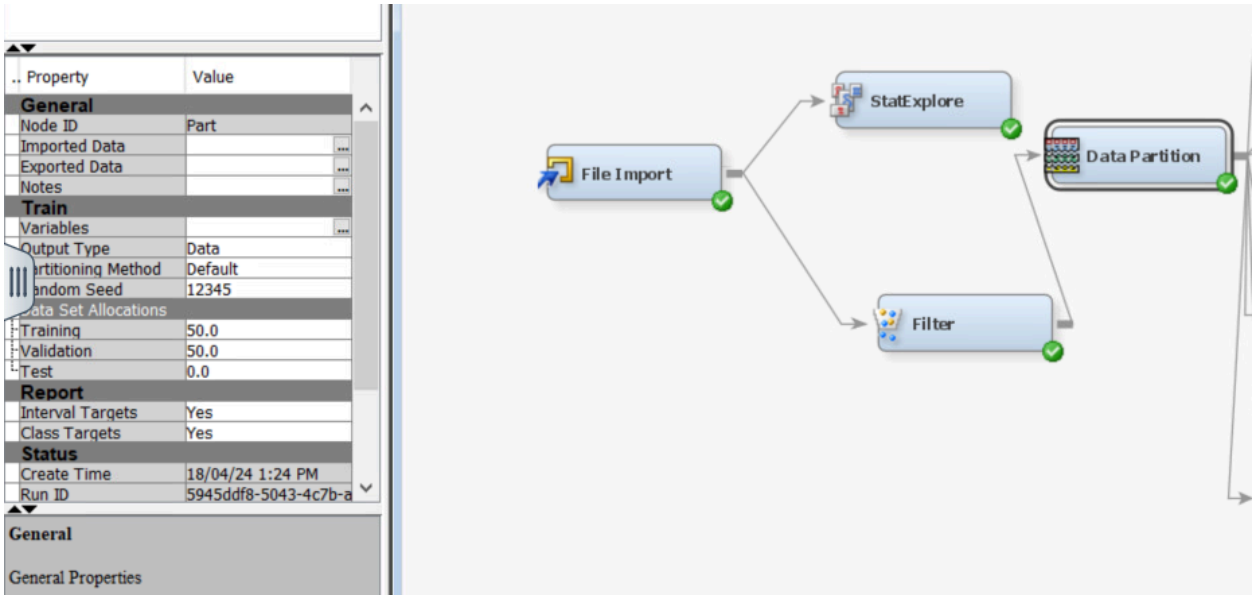


The first graph, shows the results of a statistical test that compares how likely each variable is related to the employee's decision to leave the company (Attrition). The taller the bar, the stronger the relationship between that variable and attrition. For example, the tallest bar for "Overtime" suggests that whether an employee works overtime is strongly associated with their likelihood to leave the company.

The second graph, "Variable Worth," ranks the variables based on their importance in predicting attrition. A higher bar means the variable is more important. Here, "Age" appears to be the most significant predictor of whether employees stay or leave, while other variables like "MonthlyIncome" and "JobSatisfaction" also play a role, but to a lesser extent.

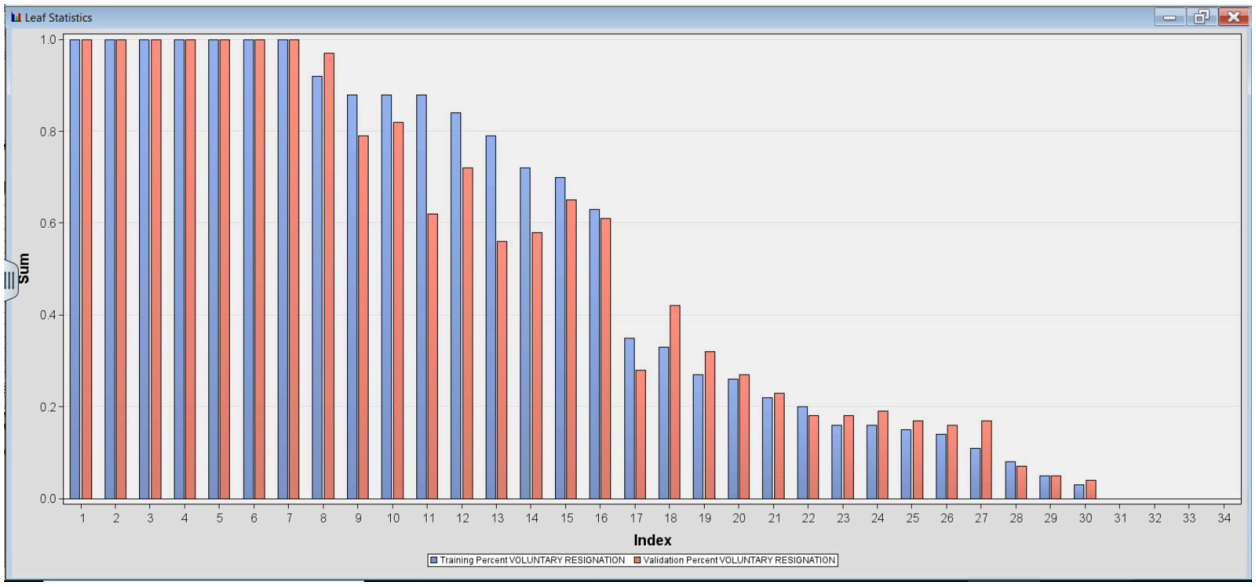
These graphs tell us what factors might influence an employee's decision to stay with or leave the company. For instance, those working overtime or of a certain age might be more likely to leave, which could guide how the company approaches work distribution, benefits, or career development programs to retain staff.

Data Partition



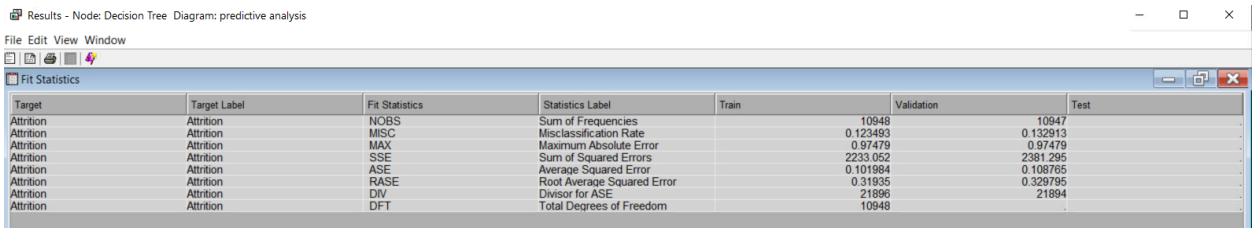
The Data Partition node has been set up to split the dataset into two equal parts: 50% for training the model and 50% for validating its predictions. This is done to assess the model's performance and ensure that it can make accurate predictions on new, unseen data. It's like a practice test for the model before the real exam, where the training set is for learning and the validation set is for testing how well it's learned.

DECISION TREE 1



Leaf Statistics:

This graph shows the comparison of percentages for voluntary resignation across different segments (leaves) of the data. The blue bars represent the training data, and the red bars represent the validation data. Ideally, these should be similar, showing that the model is consistent across both datasets. The values on the x-axis represent different groups or segments of employees identified by the model.

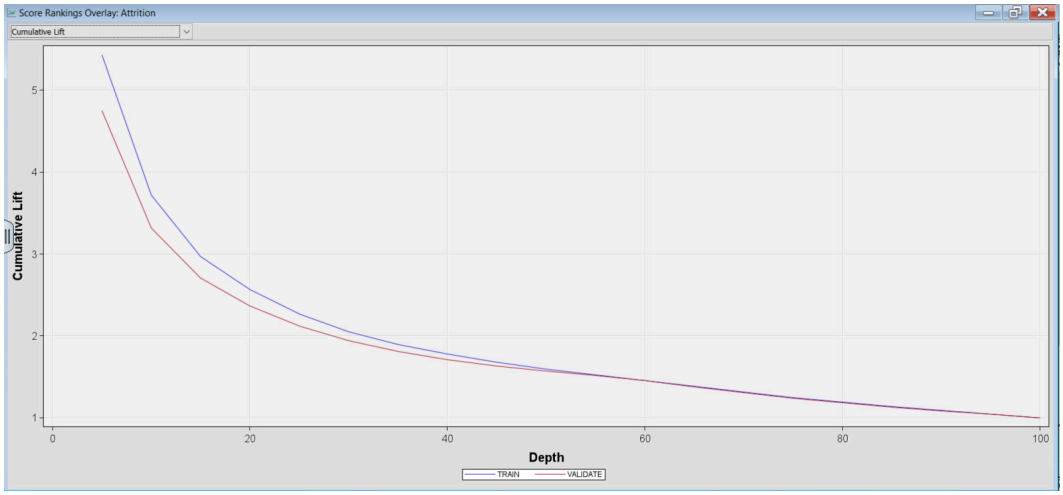


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	NOBS	Sum of Frequencies	10948	10947	
Attrition	Attrition	MISC	Misclassification Rate	0.123493	0.132913	
Attrition	Attrition	MAX	Maximum Absolute Error	0.97479	0.97479	
Attrition	Attrition	SSE	Sum of Squared Errors	2233.052	2381.295	
Attrition	Attrition	ASE	Average Squared Error	0.101984	0.108765	
Attrition	Attrition	RASE	Root Average Squared Error	0.31935	0.329795	
Attrition	Attrition	DIV	Divisor for ASE	21896	21894	
Attrition	Attrition	DFT	Total Degrees of Freedom	10948		

NORMAL DECISION TREE

Fit Statistics Table:

This table lists different statistical measures that tell us how well the model fits the data. For example, the Misclassification Rate (MISC) shows how often the model makes incorrect predictions. Lower values are better here.



Cumulative Lift Chart:
This plot illustrates how much better the model is at predicting attrition compared to random guessing. A lift value above 1 means the model is better than random, with higher values indicating stronger predictive power. The blue line represents the training set, and the red line represents the validation set.

Results - Node: Decision Tree Diagram: predictive analysis

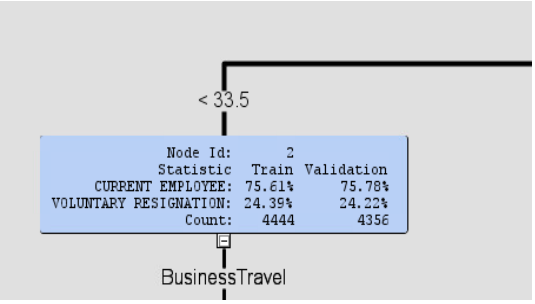
File Edit View Window

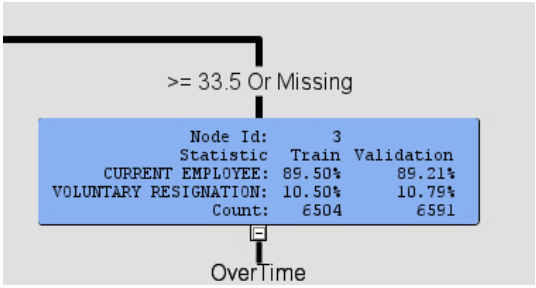
Output

	Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Validation to Training Importance
68	Age	Age	4	1.0000	1.0000	1.0000
69	DailyRate	DailyRate	5	0.7113	0.5578	0.7842
70	DistanceFromHome	DistanceFromHome	3	0.6955	0.5346	0.7687
71	BusinessTravel	BusinessTravel	3	0.5921	0.4260	0.7195
72	OverTime	OverTime	2	0.4629	0.4536	0.9800
73	JobRole	JobRole	2	0.4297	0.3836	0.8927
74	HourlyRate	HourlyRate	3	0.3545	0.1805	0.5092
75	MonthlyIncome	MonthlyIncome	2	0.2934	0.2666	0.9089
76	NumCompaniesWorked	NumCompaniesWorked	1	0.2666	0.1768	0.6633
77	Employee_Source		1	0.2664	0.2285	0.8577
78	Education	Education	1	0.2633	0.1646	0.6252
79	TotalWorkingYears	TotalWorkingYears	1	0.2180	0.2528	1.1593
80	YearsSinceLastPromotion	YearsSinceLastPromotion	1	0.2079	0.1511	0.7267
81	JobInvolvement	JobInvolvement	1	0.1900	0.1570	0.8263
82	Department	Department	1	0.1875	0.1477	0.7880
83	StockOptionLevel	StockOptionLevel	1	0.1586	0.1559	0.9829
84	MaritalStatus	MaritalStatus	1	0.1457	0.0000	0.0000

Variable Importance Table:
This table provides a detailed look at the importance of each variable used in the decision tree model. 'Age' again is at the top, suggesting it's the most decisive factor for predicting employee turnover, followed by 'DailyRate' and 'DistanceFromHome'.

DECISION TREE





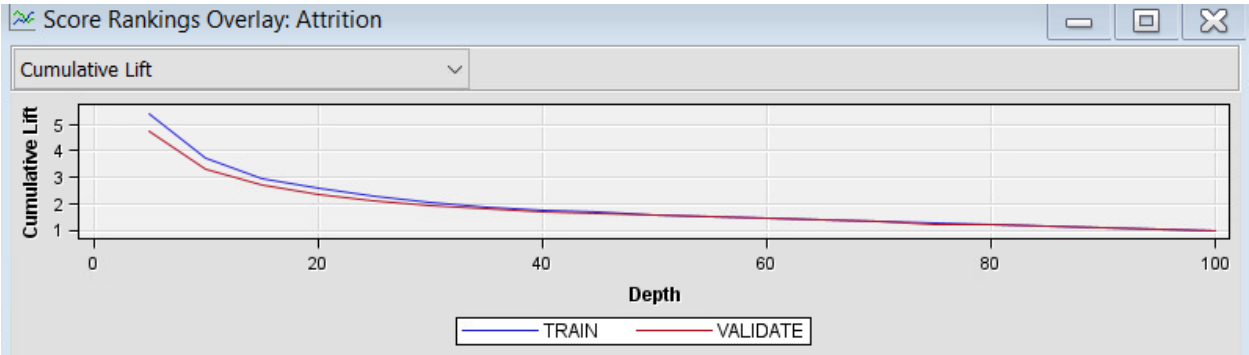
The decision tree model starts with the entire dataset and makes the first split based on the most influential variable, which, in this case, is age. This means that age was found to be the most significant factor in determining whether an employee would stay with the company or resign voluntarily. The dataset is then split into two groups based on this variable: those younger and those older than a certain age threshold.

For those younger than 33.5 years or where age data is missing, the next factor the model considers is business travel. This indicates that for younger employees, how often they travel for work influences their decision to stay with the company or leave.

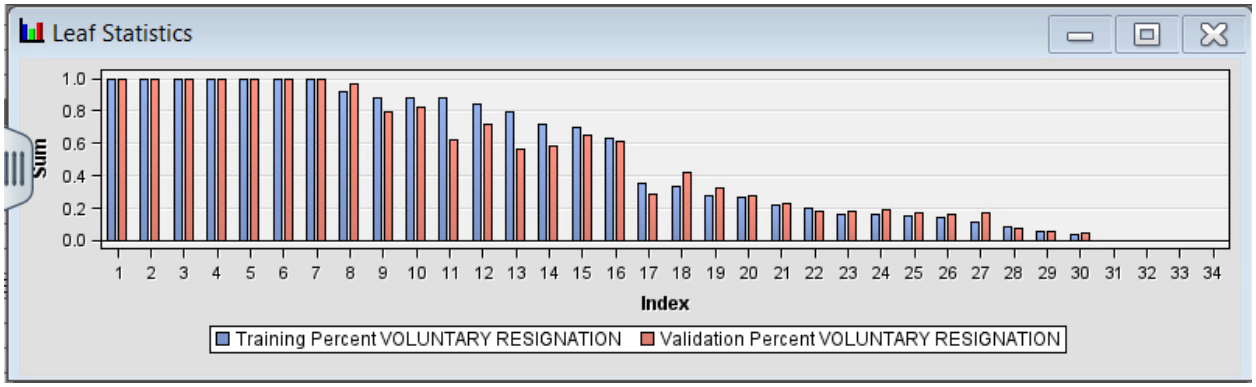
For those older than 33.5 years or where overtime information is missing, the model considers whether the employee works overtime. This implies that for the more senior employees, working overtime is a significant factor in their decision to possibly resign

Now, looking at the actual numbers, about 75.61% of younger employees in the training group (which is like a practice set for the model) and 75.78% of them in the validation group (like a final test) are still with the company. These similar numbers tell us that the model isn't just making lucky guesses but is actually learning the right patterns.

2 SPLIT ASE TREE



Cumulative Lift Chart:
This chart shows us how much better our model is at finding the employees who might leave, compared to if we just guessed without any information. The lift starts high, meaning the model is very good at spotting these employees early on when we look at the top-ranked predictions. As we move to the lower-ranked predictions, the lift decreases, which is normal. The blue line (training) and the red line (validation) are close together, which is good. It means the model is consistent and can be trusted to work well in real situations.



Leaf Statistics:
This bar chart is all about how the model's predictions stack up against the actual outcomes. Each pair of bars represents a group of employees. The blue bar shows the percentage of employees who actually left in our practice run (training), while the red bar is from a kind of test run (validation). We want the bars to be similar heights because it means our model's predictions are reliable. If the red bars are much higher or lower than the blue ones, we might need to check our model again.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	NOBS	Sum of Frequencies	10948	10947	.
Attrition	Attrition	MISC	Misclass	0.123493	0.132913	.
Attrition	Attrition	MAX	Maximum Absolute Error	0.97479	0.97479	.
Attrition	Attrition	SSE	Sum of Squared Errors	2233.052	2381.295	.
Attrition	Attrition	ASE	Average Squared Error	0.101984	0.108765	.
Attrition	Attrition	RASE	Root Average Squared Error	0.31935	0.329795	.
Attrition	Attrition	DIV	Divisor for ASE	21896	21894	.
Attrition	Attrition	DFT	Total Degrees of Freedom	10948	.	.

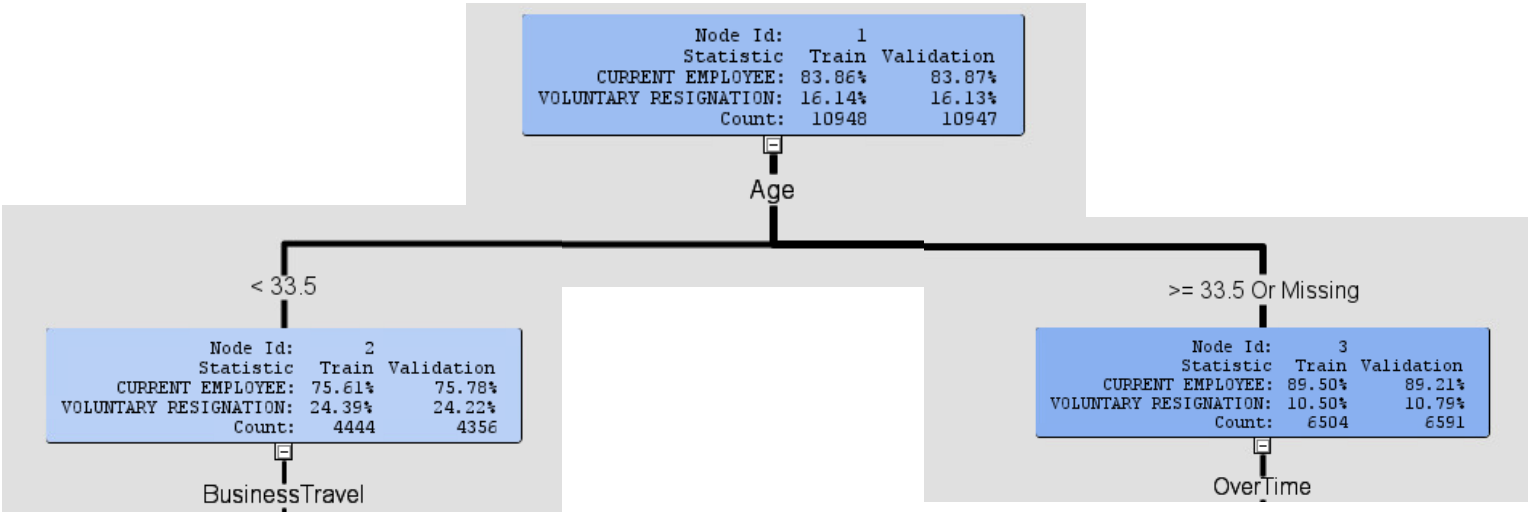
Statistics Label

In our model, an ASE of 0.101984 for training and 0.108765 for validation means we're pretty on point, but there's a slight difference between practice and the real game – a bit like a football team playing slightly differently on match day compared to training. Lower numbers are better, meaning our forecast is more accurate.

Output

	Variable Name	Label	Splitting Rules	Importance	Validation Importance
65					
66	Variable Name	Label	Splitting Rules	Importance	Validation Importance
67					
68	Age	Age	4	1.0000	1.0000
69	DailyRate	DailyRate	5	0.7113	0.5578
70	DistanceFromHome	DistanceFromHome	3	0.6955	0.5346
71	BusinessTravel	BusinessTravel	3	0.5921	0.4260
72	OverTime	OverTime	2	0.4629	0.4536
73	JobRole	JobRole	2	0.4297	0.3836
74	HourlyRate	HourlyRate	3	0.3545	0.1805
75	MonthlyIncome	MonthlyIncome	2	0.2934	0.2666
76	NumCompaniesWorked	NumCompaniesWorked	1	0.2666	0.1768
77	Employee_Source		1	0.2664	0.2285
78	Education	Education	1	0.2633	0.1646
79	TotalWorkingYears	TotalWorkingYears	1	0.2180	0.2528

Variable Importance:
This table tells us which things about employees (like their age, how much they get paid, etc.) are most important when predicting if they'll stay or leave. The ones with higher importance values have a bigger impact on the predictions. When making decisions in the business, these are the factors we might want to focus on if we're trying to keep people from leaving.



The decision tree starts by looking at the age of employees to predict if they are likely to stay with the company or leave. It uses age as the first deciding factor because, according to the analysis, it's a strong predictor of employee turnover.

The data is divided into two groups based on age, with the cutoff at 33.5 years. The 'less than or equal to 33.5 years' group is then further divided based on how often they travel for business. The 'greater than 33.5 years or missing data' group is further divided based on whether they work overtime. This splitting suggests that different factors are more or less important depending on the age group.

Younger Employees (Under 33.5 Years):
For employees younger than 33.5 years, or when the age data isn't available, the tree looks at how frequently they travel for business as the next key factor. This suggests that for this age group, the likelihood of staying with the company might be influenced by their travel commitments. Specifically, 75.61% of these younger employees in the training data (which is our model's learning phase) and 75.78% in the validation data (which tests the model's predictions) are still with the company. So, if business travel is causing younger employees to leave, the company might consider reevaluating its travel policies or offering support for work-life balance to retain them.

Older Employees (33.5 Years and Over):
For employees older than 33.5 years, or if their overtime data isn't recorded, the decision tree next considers the impact of overtime work. This indicates that for this more experienced group, whether or not they do overtime could play a significant role in their decision to stay or leave. We find that 89.50% of employees in this segment are

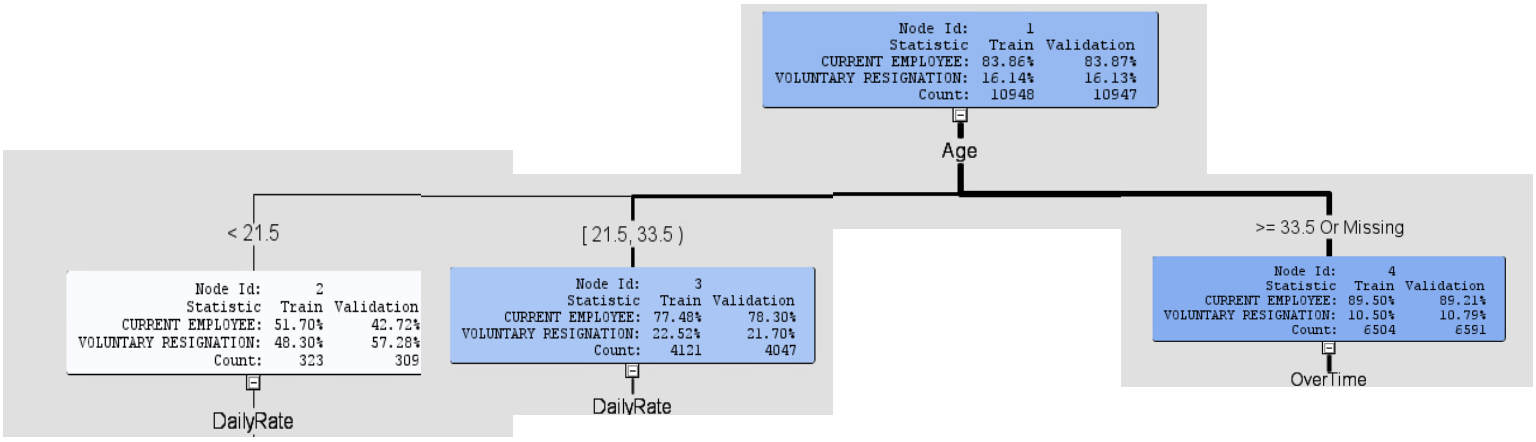
still with the company in the training data, and 89.21% in the validation data. If overtime is a push factor, the business might look into managing workload better or compensating overtime work in ways that could increase retention.

3 SPLIT ASE DECISION TREE

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	NOBS	Sum of Frequencies	10948	10947	.
Attrition	Attrition	MISC	Misclassification Rate	0.080654	0.097196	.
Attrition	Attrition	MAX	Maximum Absolute Error	0.989429	1	.
Attrition	Attrition	SSE	Sum of Squared Errors	1520.541	1804.499	.
Attrition	Attrition	ASE	Average Squared Error	0.069444	0.08242	.
Attrition	Attrition	RASE	Root Average Squared Error	0.263522	0.287088	.
Attrition	Attrition	DIV	Divisor for ASE	21896	21894	.
Attrition	Attrition	DFT	Total Degrees of Freedom	10948	.	.

This 3-split ASE tree has an ASE of 0.069444 for the training set and 0.08242 for the validation set, which means it's fairly accurate but not perfect. If the ASE is lower than our other trees for the validation set (the real test of how good our model is), then it's the sharper tool for predicting if an employee might leave.



The decision tree begins by evaluating the entire set of employees and first considers age as the most predictive factor of whether an employee will stay or leave the company. The tree splits the data into two groups: one for employees younger than 33.5 years or with missing age data, and another for those 33.5 years or older.

For employees younger than 33.5 years or with missing age data, the daily rate of pay is the next variable considered. The implication here is that among the younger or

less-experienced employees, the daily rate seems to influence their decision to stay or leave. Specifically, those with a daily rate less than $21.5 are grouped separately, suggesting that pay rate matters more for younger employees in this income bracket when considering staying with the company or resigning.

For the group of employees who are 33.5 years or older, or where overtime data is missing, the model next looks at whether the employee works overtime. This suggests that for older or more experienced employees, overtime is a significant factor in their decision to stay or resign. Employees who do not work overtime or where this data is missing are more likely to remain with the company (with the training group showing 90.50% are current employees), indicating that a better work-life balance could be essential for retaining more experienced staff.

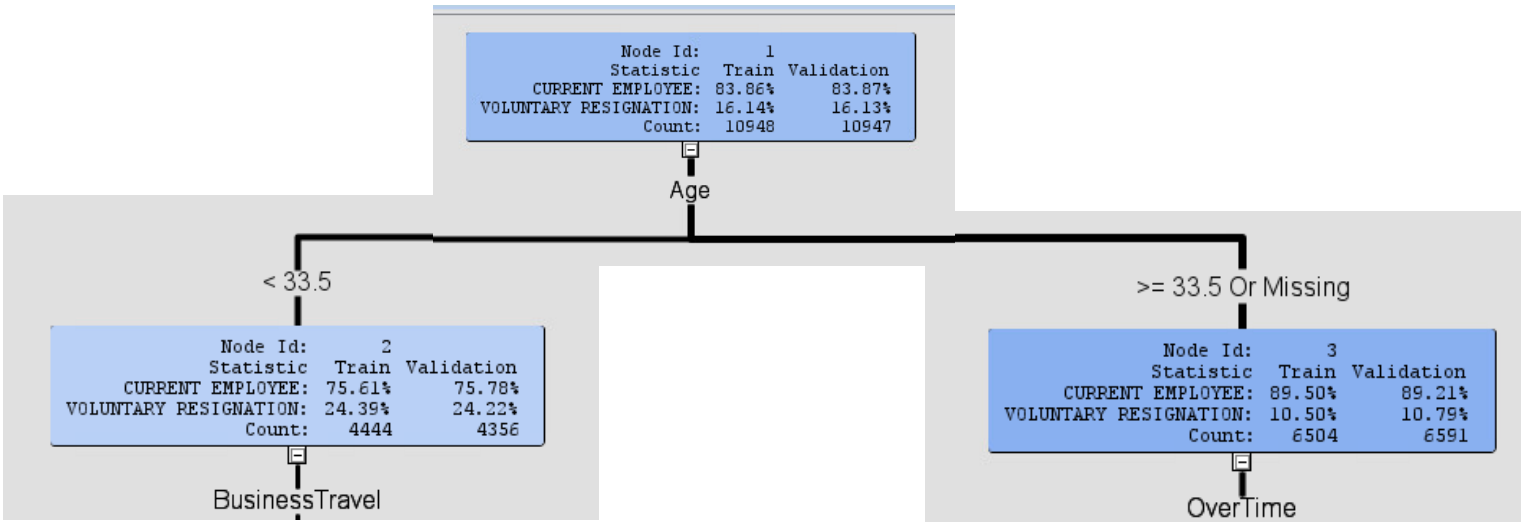
The percentages shown for each node, like "Current Employee: 83.86% Train, 83.87% Validation," tell us the proportion of the group under consideration that are still employed. The consistency between training and validation percentages indicates that the model's rules apply well to unseen data, not just the data on which it was trained. The "Count" in each leaf shows the number of observations (employees) that fall into each category, providing a sense of the sample size that the percentage is based on.

2 SPLIT MISCLASSIFICATION

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	NOBS	Sum of Frequencies	10948	10947	.
Attrition	Attrition	MISC	Misclassification Rate	0.123493	0.132913	.
Attrition	Attrition	MAX	Maximum Absolute Error	0.97479	0.97479	.
Attrition	Attrition	SSE	Sum of Squared Errors	2233.052	2381.295	.
Attrition	Attrition	ASE	Average Squared Error	0.101984	0.108765	.
Attrition	Attrition	RASE	Root Average Squared Error	0.31935	0.329795	.
Attrition	Attrition	DIV	Divisor for ASE	21896	21894	.
Attrition	Attrition	DFT	Total Degrees of Freedom	10948	.	.

In this particular model, we see an ASE of 0.101984 for the training set and 0.108765 for the validation set. In business terms, this means our model is fairly accurate in predicting whether an employee will stay with the company or leave. It's not perfect—no model is—but it's giving us a reliable indication of employee attrition based on the data it learned from.



The decision tree starts by examining the age of employees. The first split divides employees into two groups: those younger than 33.5 years and those 33.5 years or older, including any missing age data.

For employees younger than 33.5 years, the decision tree then considers business travel as a significant factor. Here, the percentages—75.61% for the training group and 75.78% for the validation group—indicate that the majority of these younger employees are still with the company. The close match between these percentages suggests that the model is similarly effective at predicting outcomes for both groups.

On the other side, for employees 33.5 years or older, overtime appears as a key factor influencing their stay or departure from the company. In this branch, a higher percentage of employees—89.50% for the training group and 89.21% for the validation group—are currently employed, which is higher compared to the younger employees. This indicates that age, combined with overtime, might be a stronger predictor for employee retention than business travel for younger employees.

The "Count" value in each leaf indicates the number of cases (or employees) that fall into each category. So, for instance, the leaf under the "Business Travel" split tells us that there are 4,444 younger employees considered in the training data and 4,356 in the validation data, and among these, about 75% are still employed.

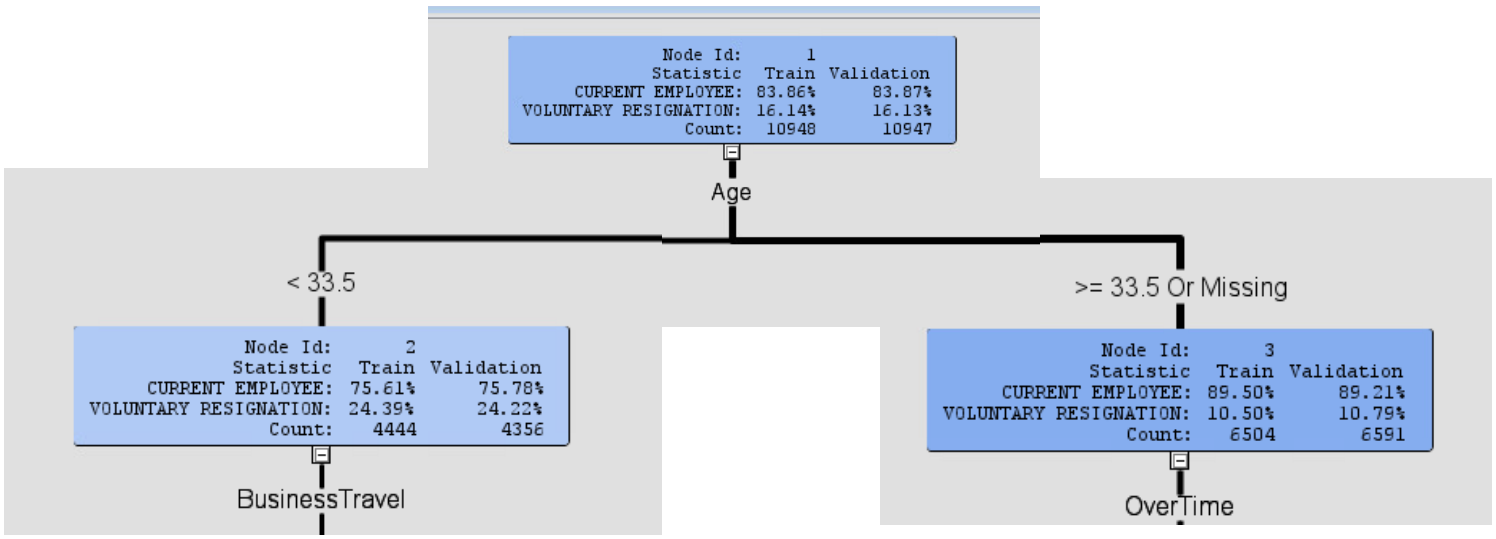
MAXIMAL TREE:

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	NOBS	Sum of Freque...	10948	10947	.
Attrition	Attrition	MISC	Misclassificatio...	0.123493	0.133004	.
Attrition	Attrition	MAX	Maximum Abs...	0.994083	0.994083	.
Attrition	Attrition	SSE	Sum of Square...	2137.722	2299.749	.
Attrition	Attrition	ASE	Average Squar...	0.097631	0.10504	.
Attrition	Attrition	RASE	Root Average ...	0.312459	0.324099	.
Attrition	Attrition	DIV	Divisor for ASE	21896	21894	.
Attrition	Attrition	DFT	Total Degrees ...	10948	.	.

Training ASE (0.097631):
This value shows how well the model predicts the training data, which is the data it has already seen. A value of approximately 0.098 suggests that, on average, the squared difference between the predicted outcomes and the actual outcomes is fairly low for the training set, meaning the model is quite accurate on the data it was trained on.

Validation ASE (0.10504):
This is the average of the squared differences between the predicted and actual outcomes for the validation set, which is new data the model hasn't seen before. A value of approximately 0.105 indicates that the model is also performing reasonably well on unseen data, although the error is slightly higher than in the training set, which is expected.



In the first split of the decision tree, age is the initial factor used to predict whether employees might leave or stay with the company. The split at "Age < 33.5" suggests that age is a strong indicator of attrition. Specifically, the model finds that:

For employees younger than 33.5 years or when the age data is not available, about 75.61% are currently employed (as seen in the training data) and about 75.78% in the validation data, indicating consistency across both sets of data. Conversely, about 24.39% of these younger employees are identified as voluntarily resigning in the training data, and 24.22% in the validation data.

For employees 33.5 years or older or when overtime data is missing, the retention is higher, with 89.50% being current employees in the training data and 89.21% in the validation data. The voluntary resignation rates are lower at 10.50% for the training data and 10.79% for the validation data.

The "Count" number represents the number of records in each node of the tree. For example, for employees under 33.5 years old, there are 4,444 cases in the training set and 4,356 in the validation set.

BEST DECISION TREE

Normal Decision Tree:

- Training ASE: 0.101384
- Validation ASE: 0.108765

2. 2 Split ASE Decision Tree:
- Training ASE: 0.101384
- Validation ASE: 0.108765
- This tree has identical performance to the Normal Decision Tree, suggesting no improvement.

3. 3 Split ASE Decision Tree:
- Training ASE: 0.069444
- Validation ASE: 0.08242
- This tree shows improvement over the first two trees, with lower error rates on both training and validation sets.

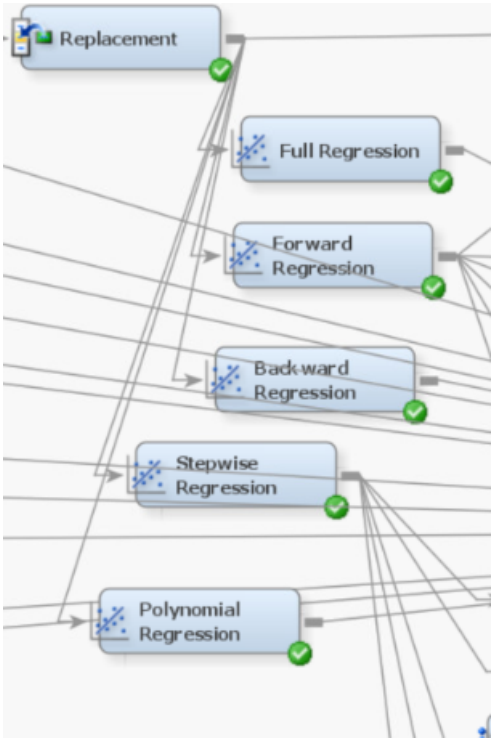
4. 4 Split ASE Decision Tree:
- Training ASE: 0.048436
- Validation ASE: 0.059434

- This tree demonstrates the best performance so far, with the lowest ASE values, indicating the most accurate predictions.

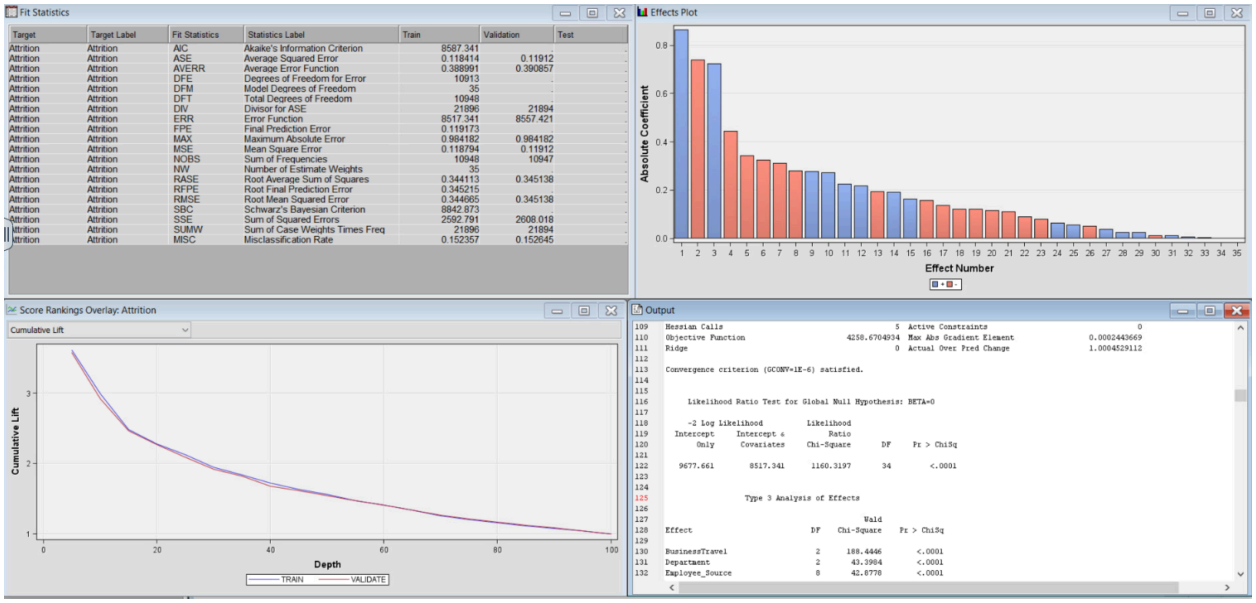
5. 2 Split Misclassification Decision Tree:
    - Training ASE: 0.101984
    - Validation ASE: 0.108765
    - The performance is very similar to that of the first Normal Decision Tree and the 2 Split ASE Tree, which suggests that splitting by misclassification hasn't provided a benefit in terms of ASE.
6. Maximal Tree:
    - Training ASE: 0.097631
    - Validation ASE: 0.10504
    - This tree has slightly better ASE values than the Normal and 2 Split Misclassification Trees but doesn't outperform the 3 Split or 4 Split ASE Trees.

In conclusion, the 4 Split ASE Decision Tree is the best among those listed because it has the lowest ASE values for both training and validation. This suggests that the model with four splits is the most effective at making accurate predictions while balancing the fit to both the training data and unseen validation data. In business terms, this model would likely be the most reliable for predicting outcomes based on the data it was trained on and is expected to generalize better to new, similar data.

REGRESSION



FULL REGRESSION



Fit Statistics:
This section provides various measures of how well the regression model fits the data. Notably, the ASE (Average Squared Error) is given as:

- Train ASE: 0.118941
- Validation ASE: 0.119813

The ASE for the training and validation is similar, suggesting the model fits both datasets reasonably well. However, the test ASE is significantly higher, which could indicate the model doesn't perform as well on new, unseen data, or there could be a significant difference in the distribution of the test data.

Score Rankings Overlay:
Attrition (Cumulative Lift Chart): This graph shows the cumulative lift, a measure of the model's effectiveness in ranking cases according to their probability of experiencing an event (in this case, attrition). The chart indicates how much better the model is at predicting cases than random chance. Ideally, you want the lift to start high and stay above the baseline (lift of 1), which would mean the model has good predictive power.

Effects Plot:
This bar chart shows the absolute coefficients for the effects (independent variables) included in the model. Larger values indicate a stronger relationship with the dependent variable. The coloring likely indicates the significance of these variables, with red possibly indicating less significance than blue. The variables are numbered, and without

the variable names, it’s not possible to determine which variables these numbers correspond to.

Output (Likelihood Ratio Test):
This section provides the likelihood ratio test results, which help determine if the model as a whole fits significantly better than a model with no predictors. The Chi-square values and associated p-values (<0.0001) indicate that variables such as BusinessTravel, Department, and Employee_Source are highly significant in predicting attrition.

FORWARD REGRESSION:

Cumulative Lift Chart:
This plot is showing the cumulative lift which assesses the model's ability to predict events more accurately than random chance. The lift values for both training and validation datasets start higher than the baseline and decrease gradually. This indicates that the model is useful in ranking predictions, especially at lower depths (meaning it ranks the highest risk cases well).

Fit Statistics:
The ASE (Average Squared Error) values given are:

- Train ASE: 0.118605
- Validation ASE: 0.119139

These values indicate that the model's performance on training and validation data is consistent, but its performance on test data is substantially worse, suggesting that the model may not generalize well to new data.

Effects Plot:
It shows the absolute coefficients for the variables included in the model after each step of the forward regression process. The larger bars represent more substantial effects on the dependent variable, attrition. The color difference likely denotes the significance or the scale of the coefficients, with blue possibly indicating stronger effects or variables that entered the model earlier in the process.

Output (Likelihood Ratio Test):
This section provides the likelihood ratio test results, which help determine if the model as a whole fits significantly better than a model with no predictors. The Chi-square values and associated p-values (<0.0001) indicate that variables such as

BusinessTravel, Department, and Employee_Source are highly significant in predicting attrition.

BACKWARD REGRESSION

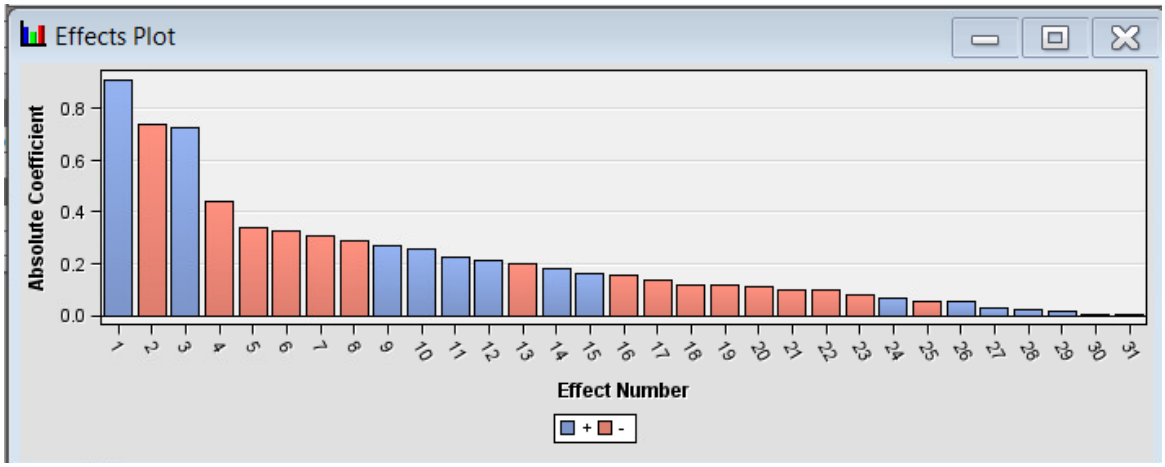
Cumulative Lift Chart:
This plot is showing the cumulative lift which assesses the model's ability to predict events more accurately than random chance. The lift values for both training and validation datasets start higher than the baseline and decrease gradually. This indicates that the model is useful in ranking predictions, especially at lower depths (meaning it ranks the highest risk cases well).

Fit Statistics:
The ASE (Average Squared Error) values given are:

- Train ASE: 0.118505
- Validation ASE: 0.119139

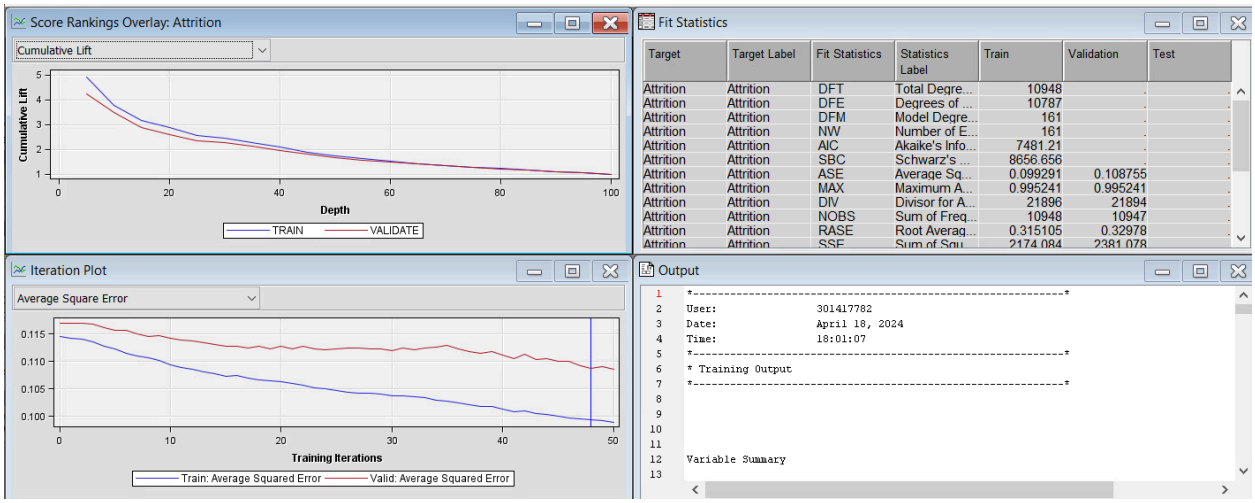
These values indicate that the model's performance on training and validation data is consistent, but its performance on test data is substantially worse, suggesting that the model may not generalize well to new data.

Effects Plot:
It shows the absolute coefficients for the variables included in the model after each step of the forward regression process. The larger bars represent more substantial effects on the dependent variable, attrition. The color difference likely denotes the significance or the scale of the coefficients, with blue possibly indicating stronger effects or variables that entered the model earlier in the process.



Output (Likelihood Ratio Test):

This section provides the likelihood ratio test results, which help determine if the model as a whole fits significantly better than a model with no predictors. The Chi-square values and associated p-values (<0.0001) indicate that variables such as BusinessTravel, Overtime, Repage, department, employment source are highly significant in predicting attrition.



In SAS Miner, an ASE (Average Squared Error) line at iteration 48 indicates the model's performance in terms of how well it fits the data. A lower ASE generally suggests better model accuracy, meaning the model's predictions are closer to the actual values. The iteration number signifies how many times the model has adjusted its parameters to minimize error during training. Therefore, at iteration 48, the model has undergone several adjustments and likely achieved a reasonable level of accuracy, but further iterations may be needed for refinement.

Fit Statistics:

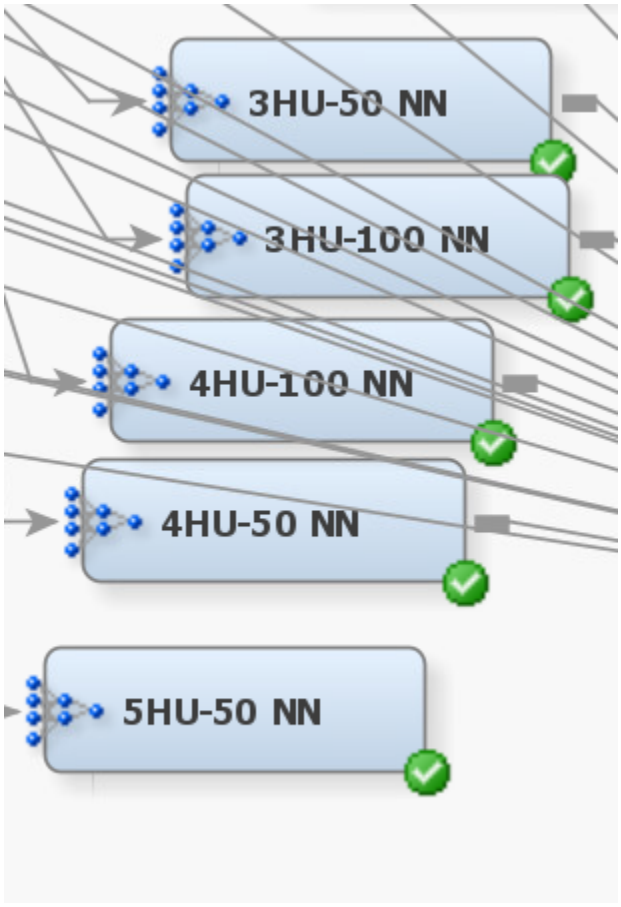
The ASE (Average Squared Error) values given are:

- Train ASE: 0.099291
- Validation ASE: 0.108755

These values indicate that the model's performance on training and validation data is consistent, but its performance on test data is substantially worse, suggesting that the model may not generalize well to new data.

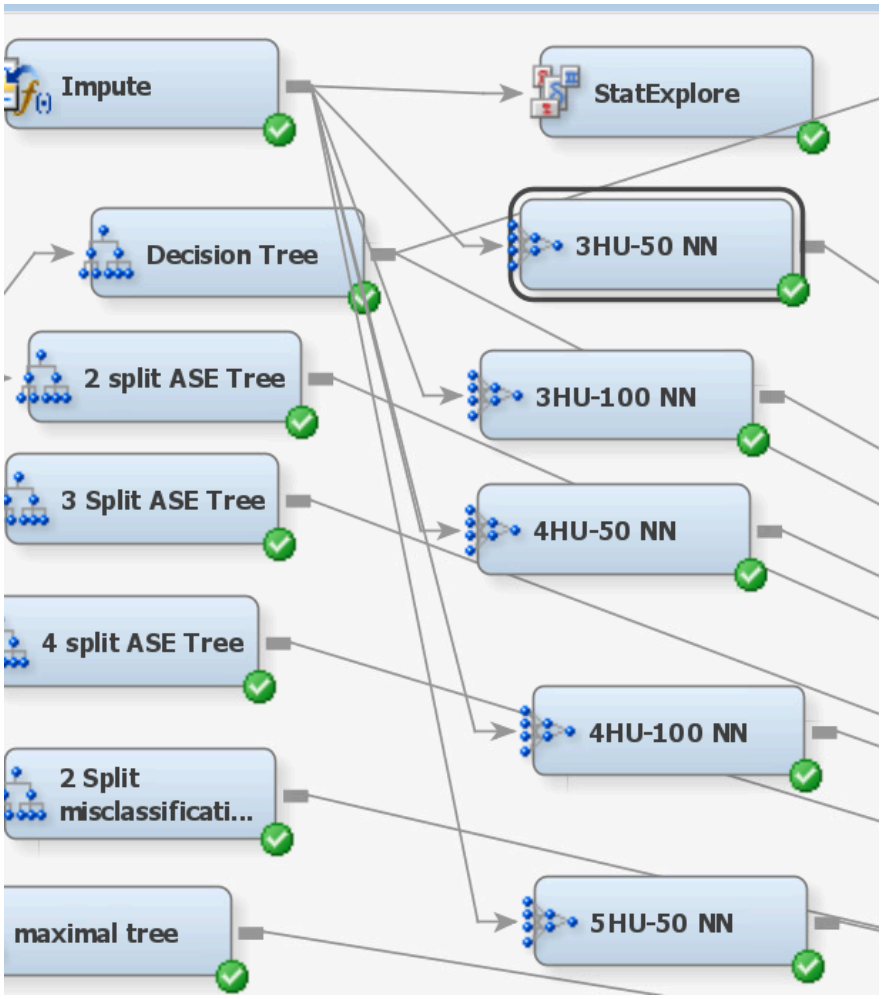
NEURAL NETWORK

Neural networks are computational models inspired by the human brain that are used to recognize patterns and make decisions. They consist of layers of interconnected nodes, similar to neurons, that process input data and learn to perform tasks by adjusting the strength of their connections based on the patterns they observe, much like learning from experience

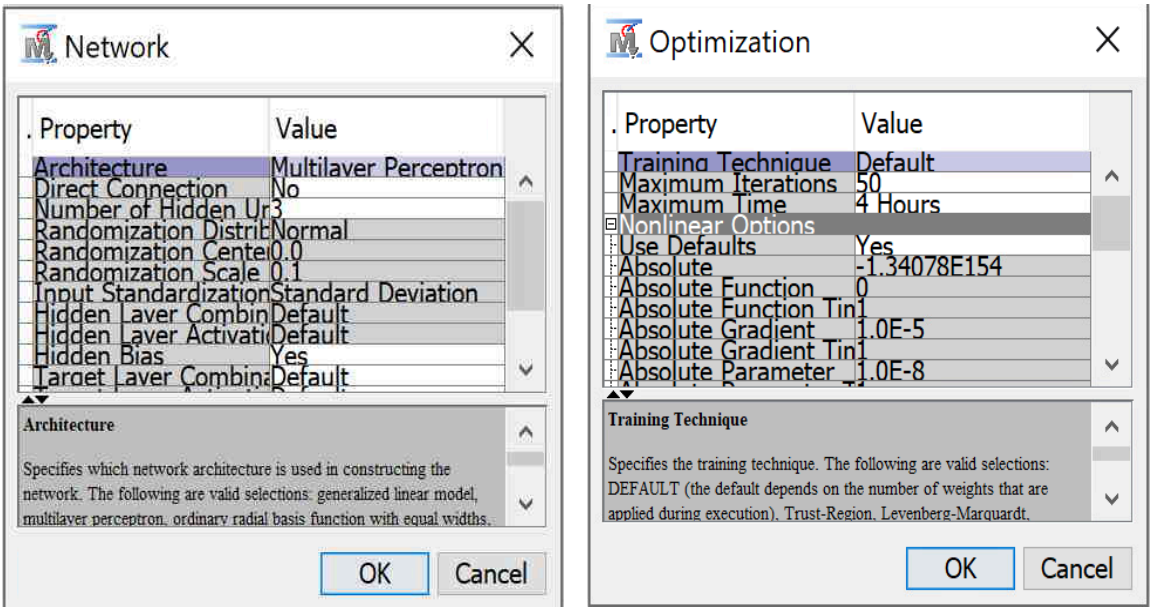


Neural Network connected from Impute node:

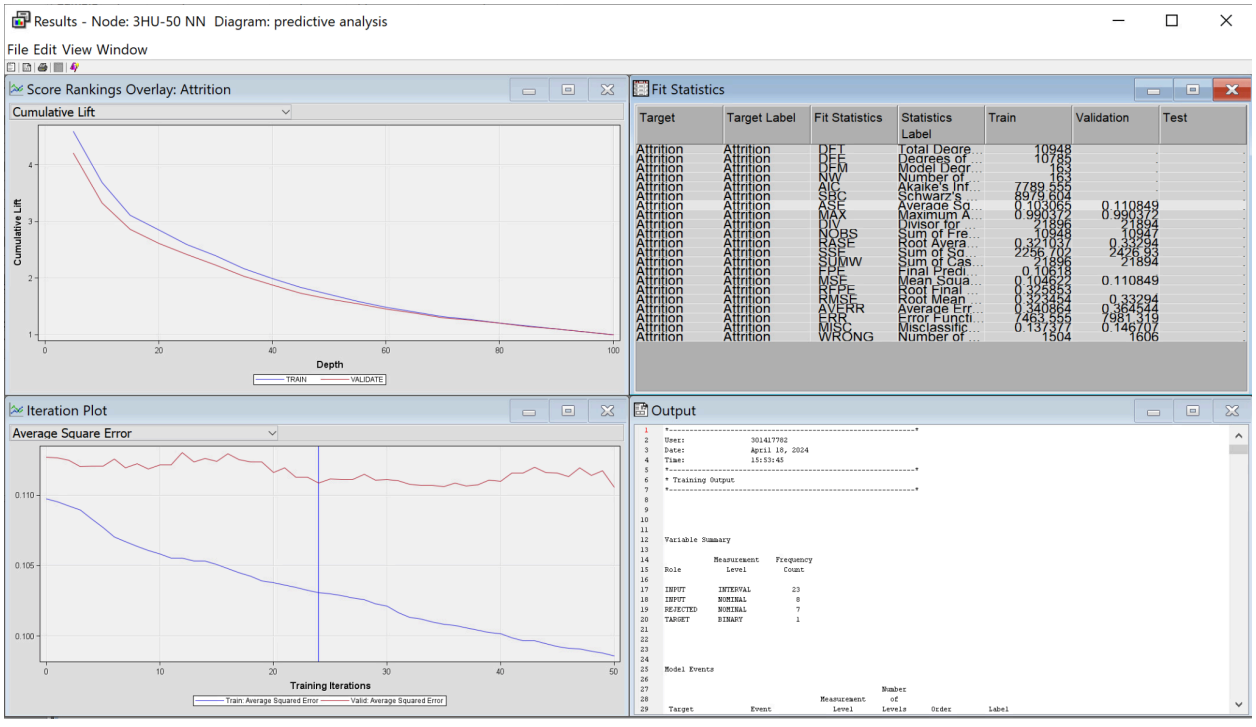
There are 5 neural networks that are connected from the impute node, these neural networks are connected in order to get the best model that works.



Neural Network 1:

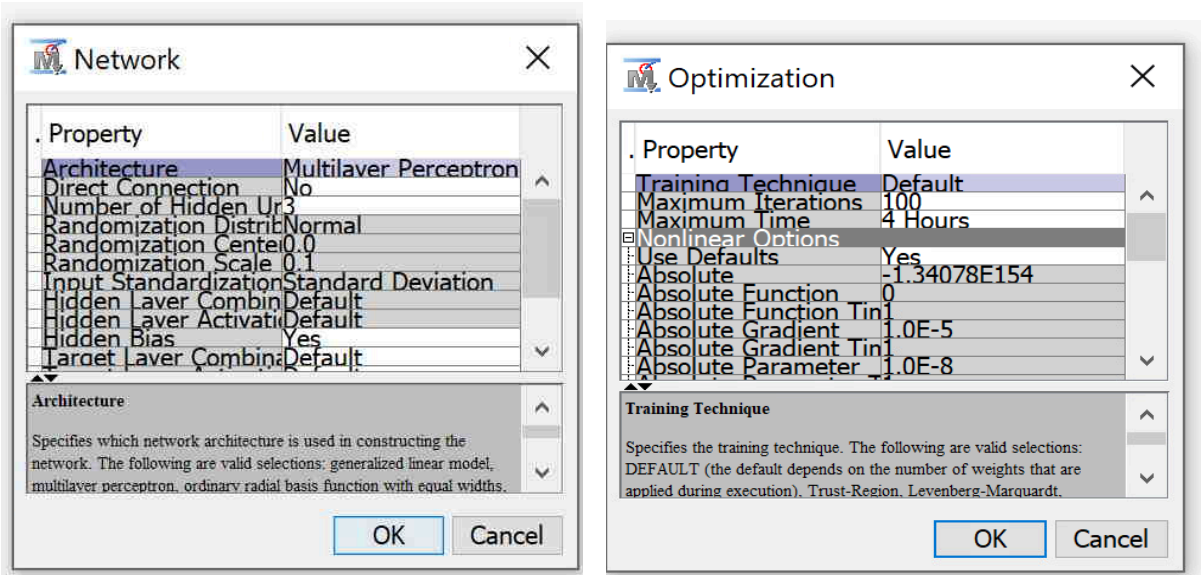


Here are the images for the neural network with **3 hidden** units and **50 iterations.**

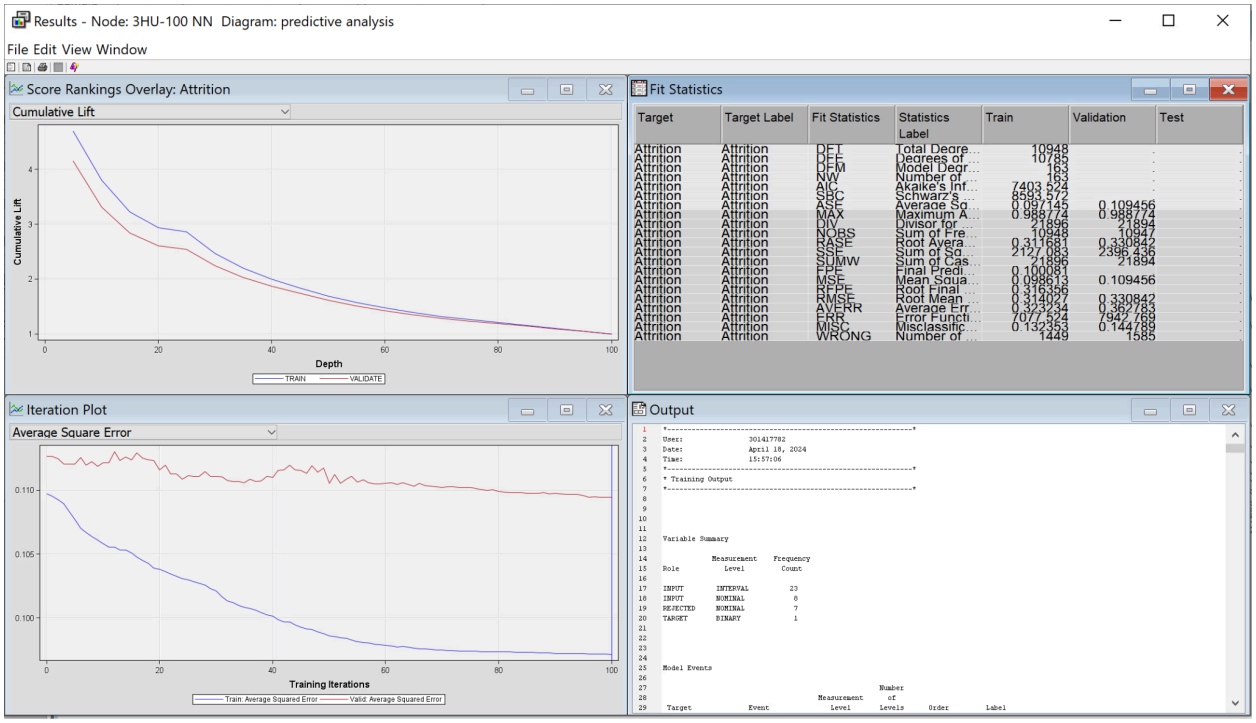


Here is the result of the first neural network with 3 hidden units and 50 iterations, which has given an average squared error of **0.110849.**

Neural Network 2:

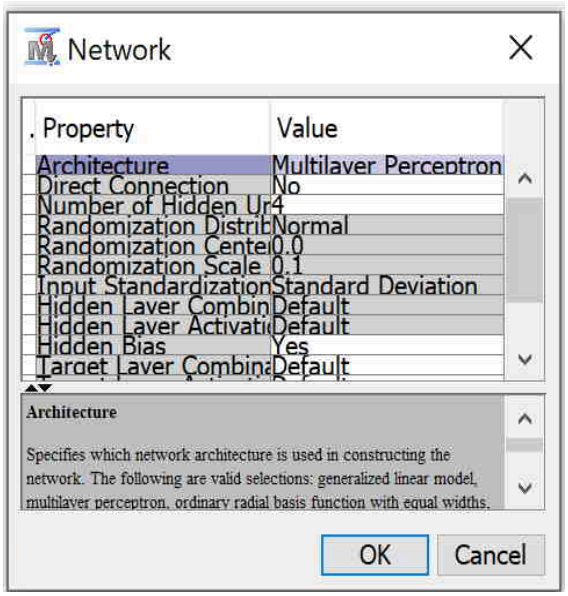


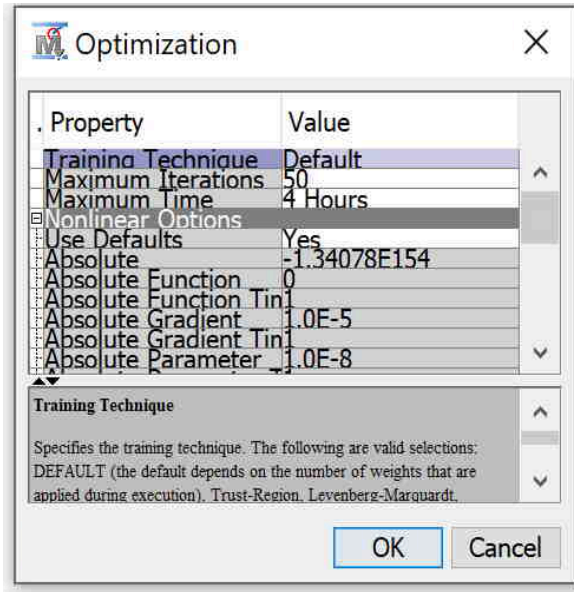
Here are the images of neural network 2 with **3 hidden units** and **100 iterations**.



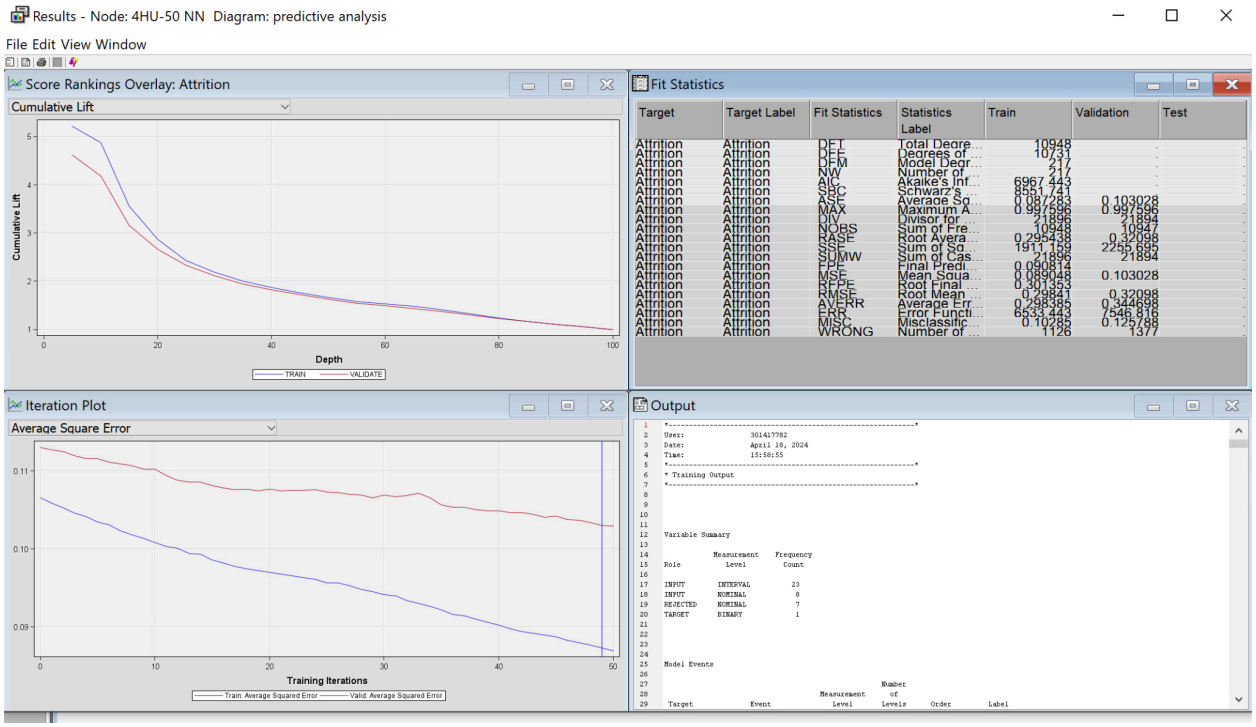
Here is the result of neural network 2 with 3 hidden units and 100 iterations, which has an average squared error of **0.109456.** In order to increase the accuracy of the model another 50 iterations are added and we can see with the results that the error has reduced compared to the 1st neural model.

Neural Network 3:



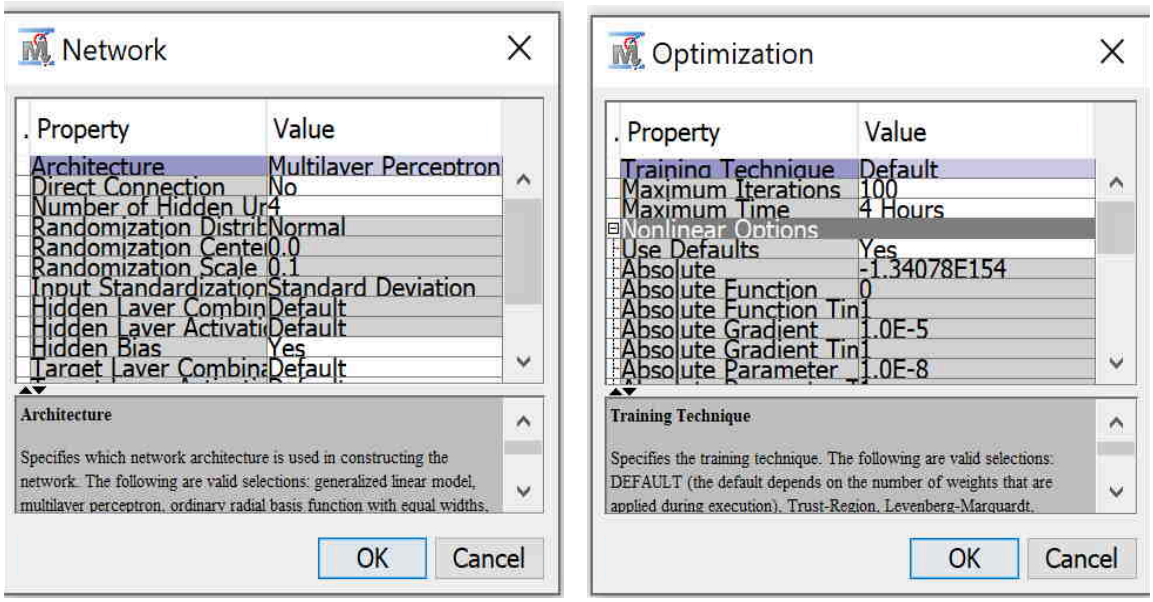


Here are the images for the neural network 3 with **4 hidden units** and **50 iterations**.

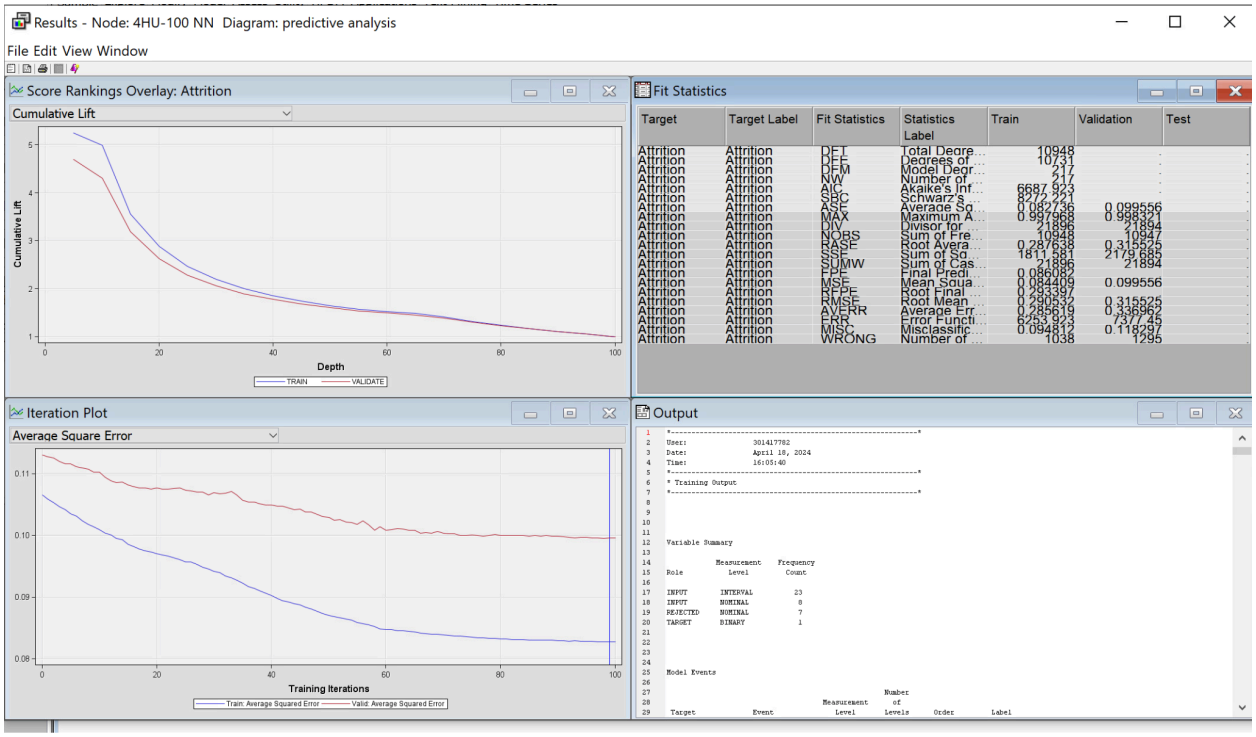


Here are the results of neural network 3 with 4 hidden units and 50 iterations, which gave us an average squared error of **0.103028.**

Neural Network 4:

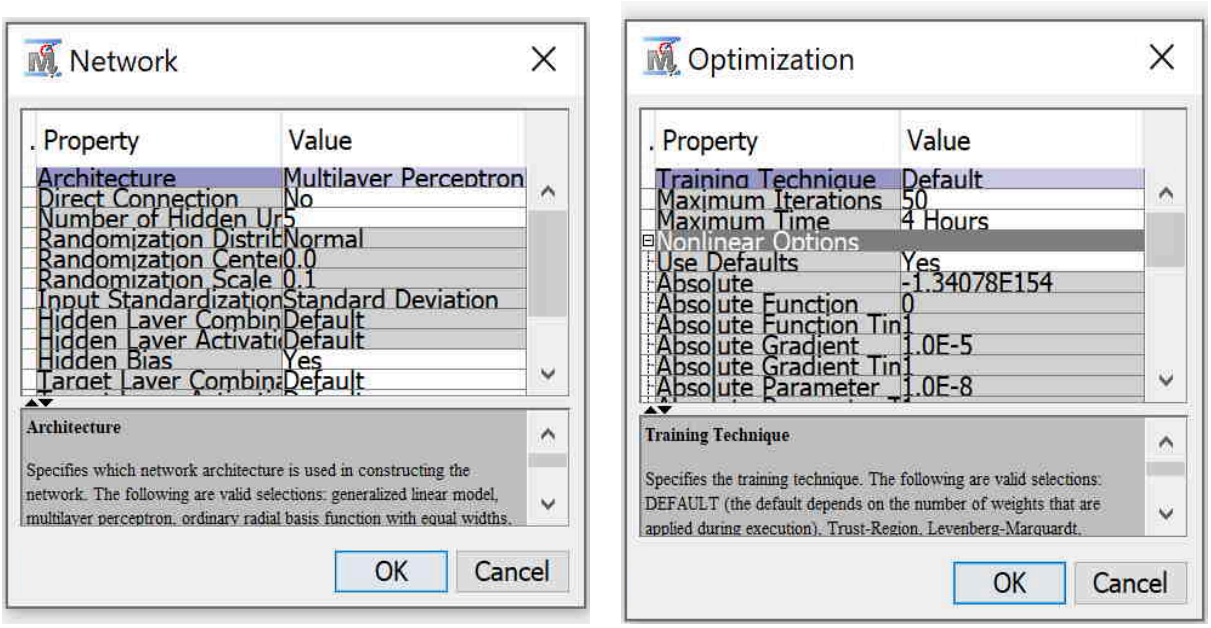


Here are the images for the neural network 4 with **4 hidden units** and **100 iterations**.

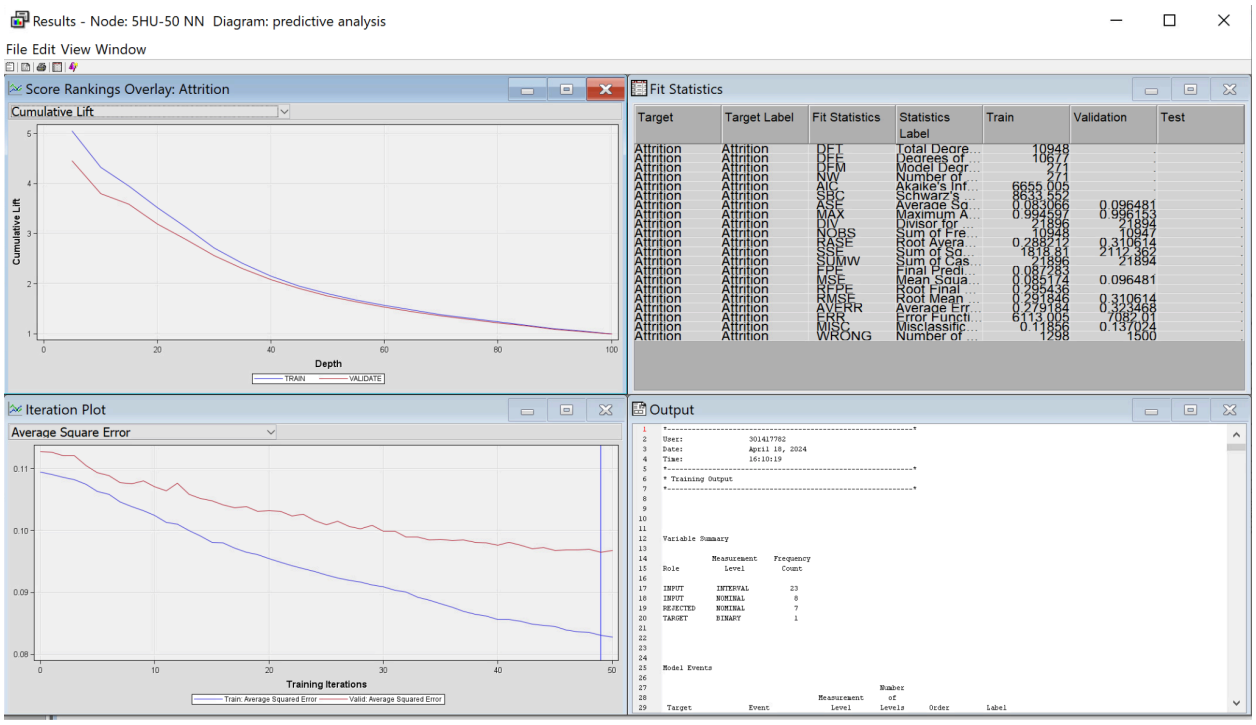


Here is the result of neural network 4 with 4 hidden units and 100 iterations, which has an average squared error of **0.099556.** In order to increase the accuracy of the model another 50 iterations are added and we can see with the results that the error has reduced compared to the 3rd neural model.

Neural Network 5:

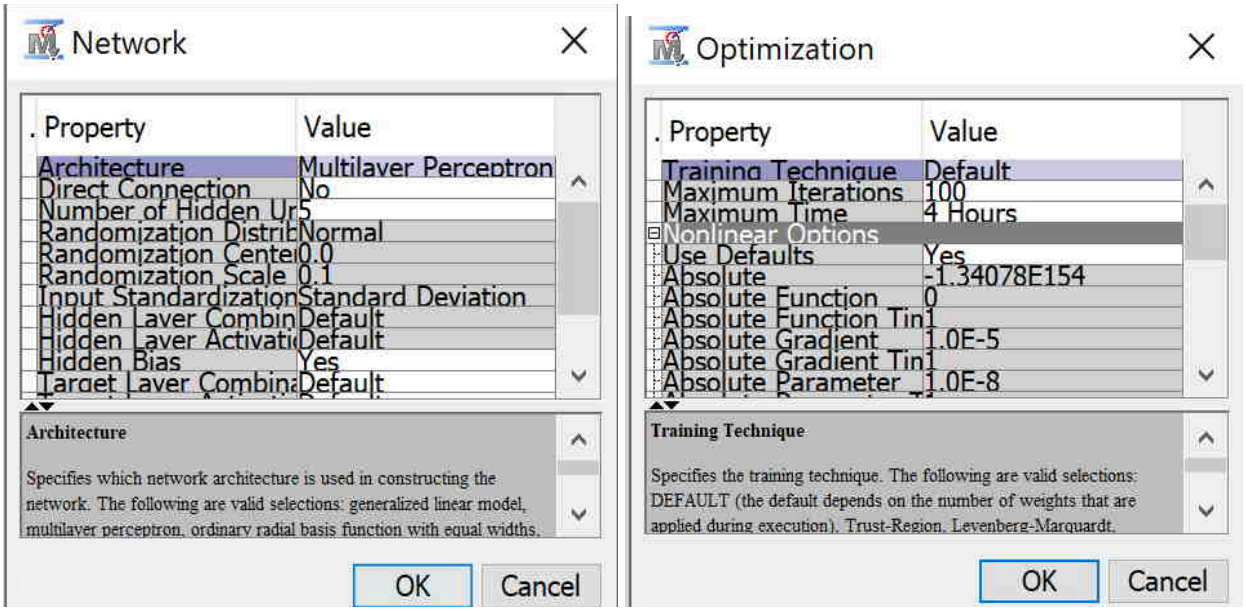


Here are the images for the neural network 5 with **5 hidden units** and **50 iterations**.

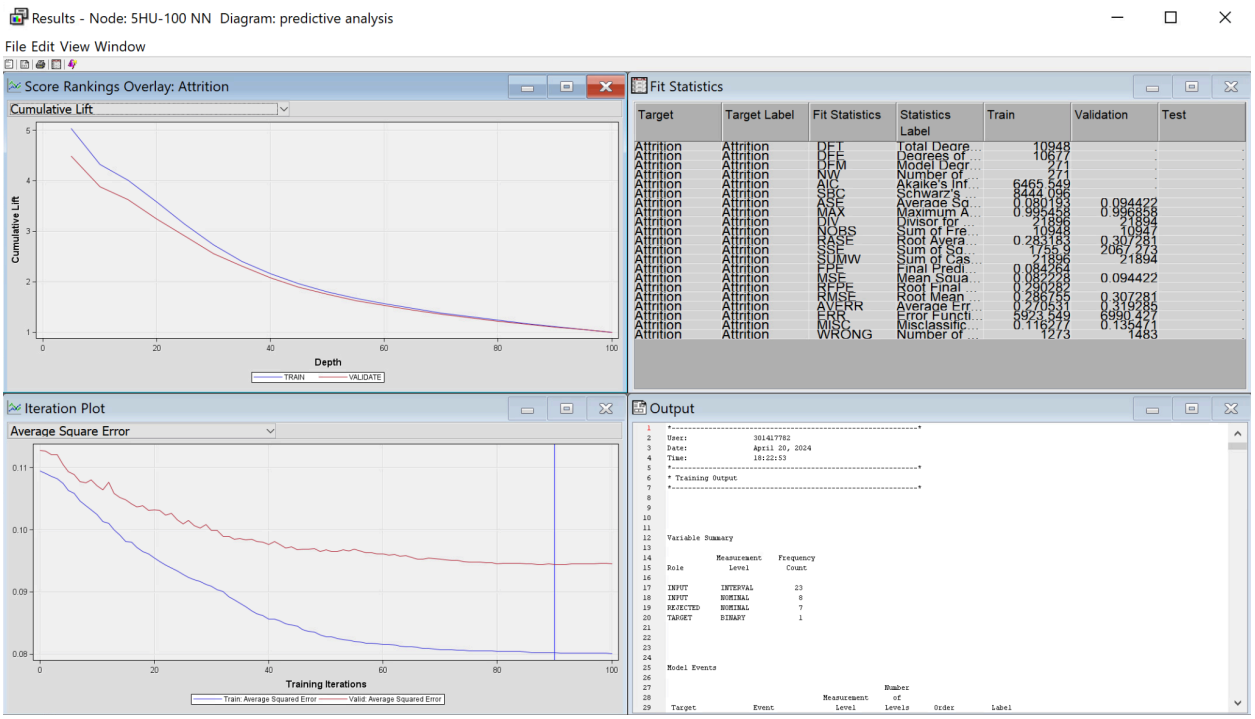


Here is the result of the 5th neural network with 5 hidden units and 50 iterations, which has given an average squared error of **0.096481.**

Neural Network 6:

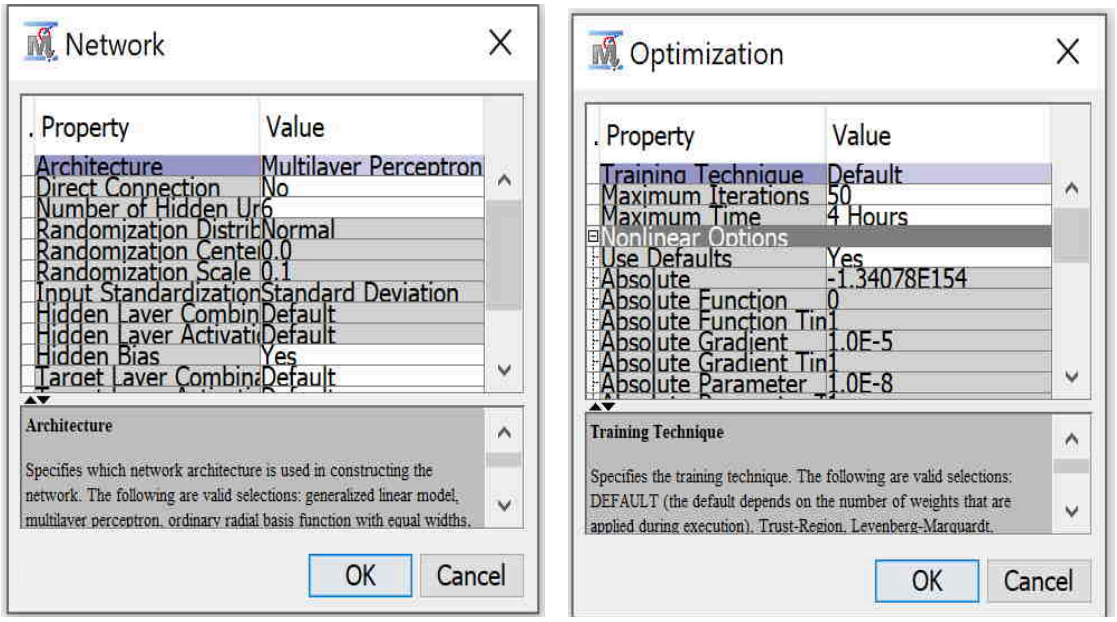


Here are the images for the neural network 6 with **5 hidden units** and **100 iterations**.

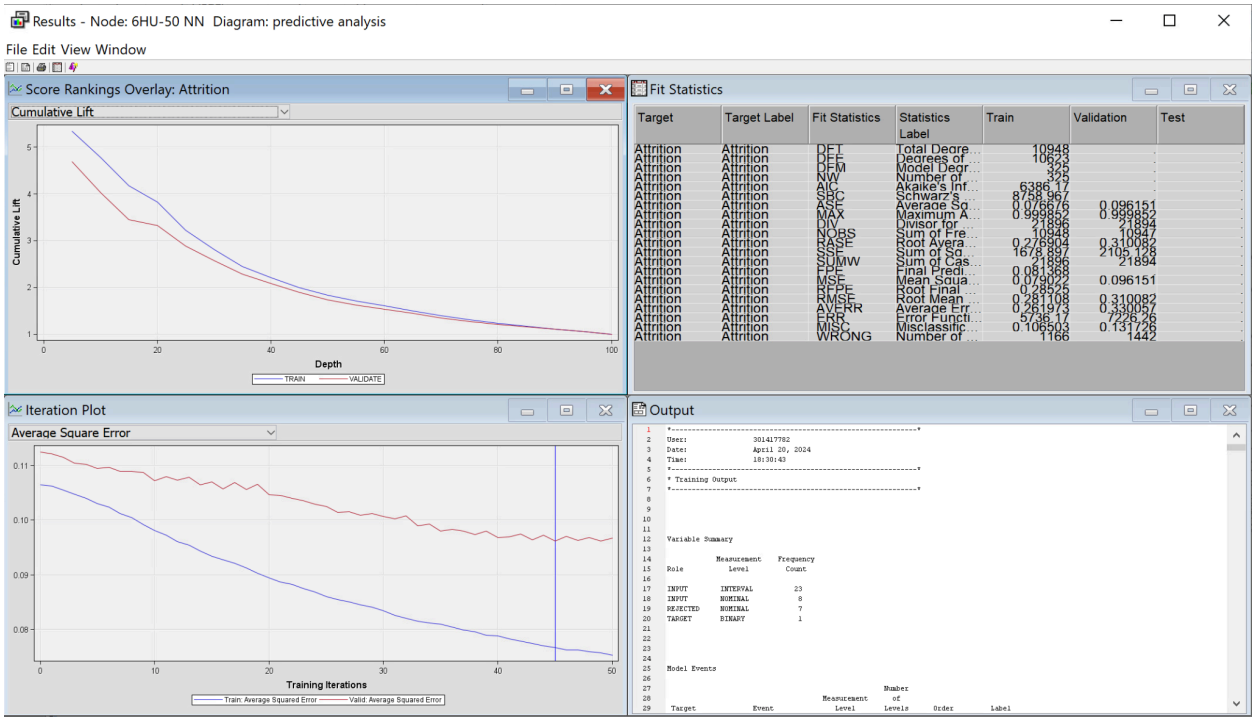


Here is the result of the 6th neural network with 5 hidden units and 100 iterations, which has given an average squared error of **0.094422.**

Neural Network 7:



Here are the images for the neural network 7 with **6 hidden units** and **50 iterations**.



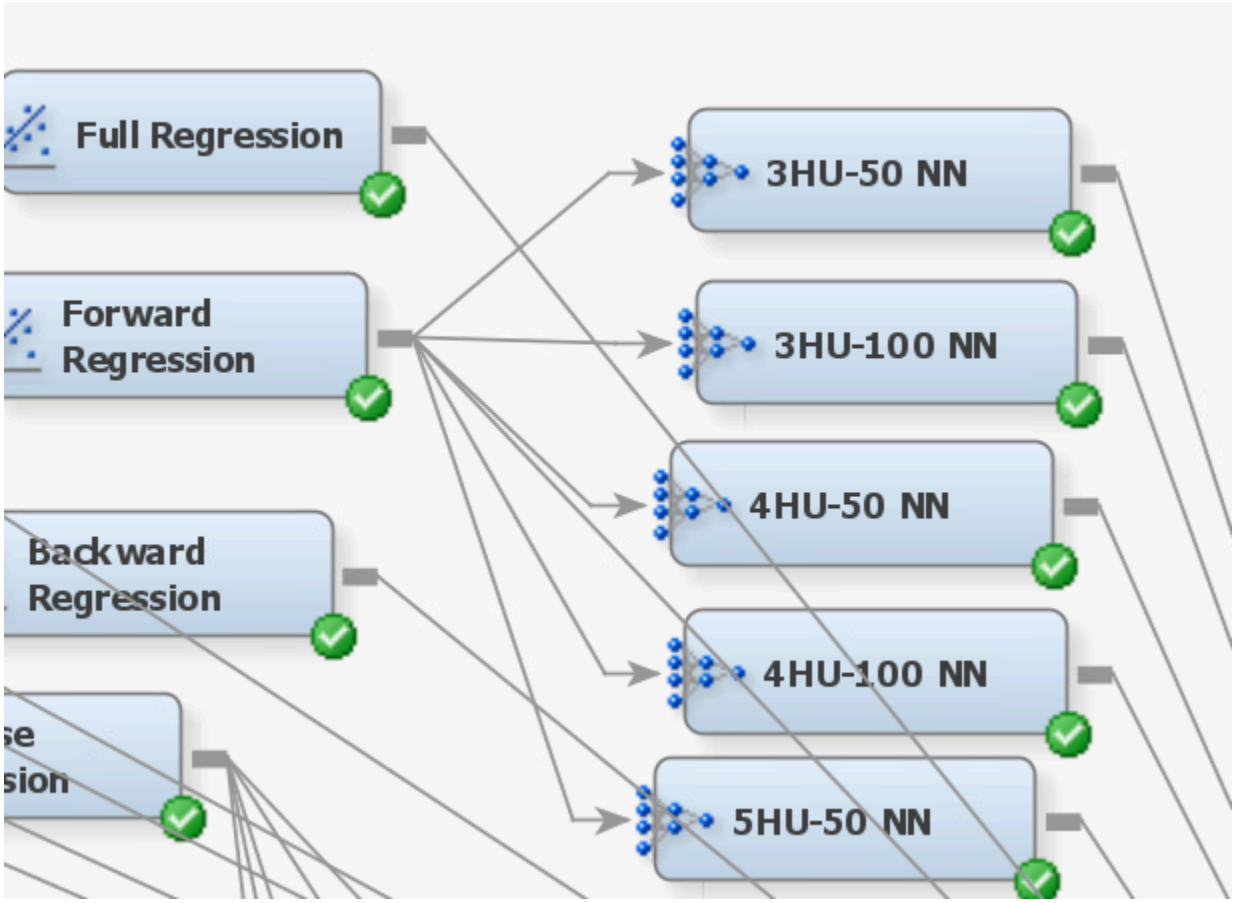
Here is the result of the 7th neural network with 6 hidden units and 50 iterations, which has given an average squared error of **0.096151.**

Here is the summary of the neural networks connected to the impute node.

Neural Network	Hidden Units	Iterations	ASE
3HU-50NN	3	50	0.110849
3HU-100NN	3	100	0.109456
4HU-50NN	4	50	0.103028
4HU-100NN	4	100	0.099556
5HU-50NN	5	50	0.096481
5HU-100NN	**5**	**100**	**0.94422**
6HU-50NN	6	50	0.96151

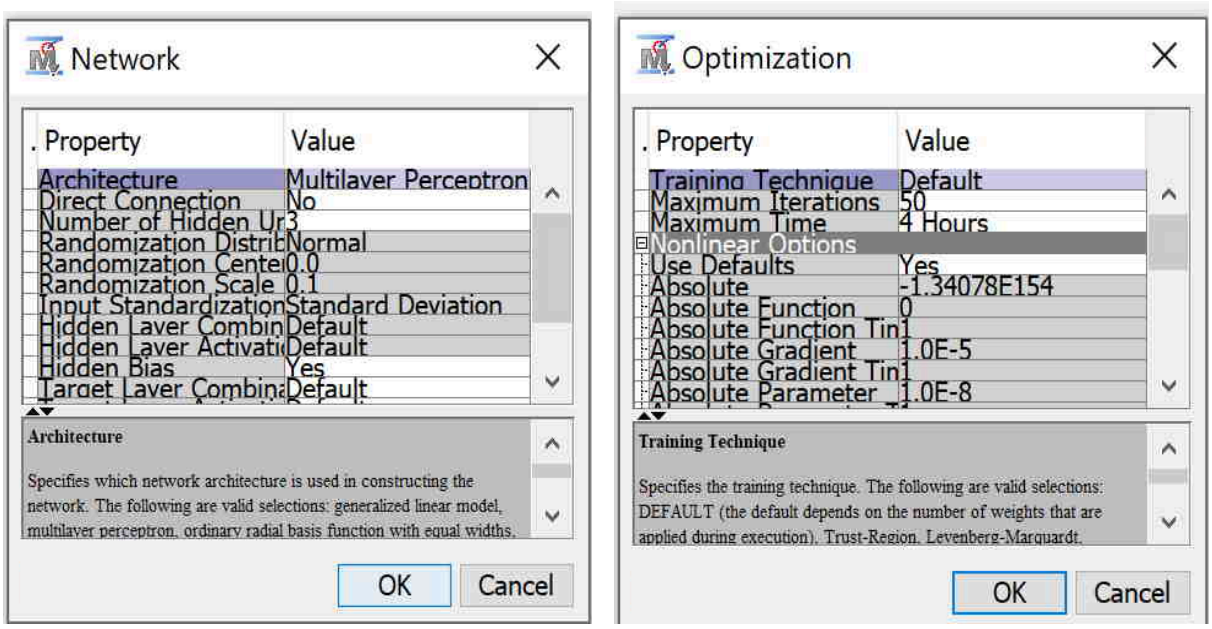
After performing 7 neural models, we arrived at our best model which is with 5 hidden units and 100 iterations. Another model was performed to get a more accurate result bu the ASE was higher than before. Hence, we stopped our neural model as we go the best model with an ASE of 0.94422.

Neural Networks connected from Forward regression:

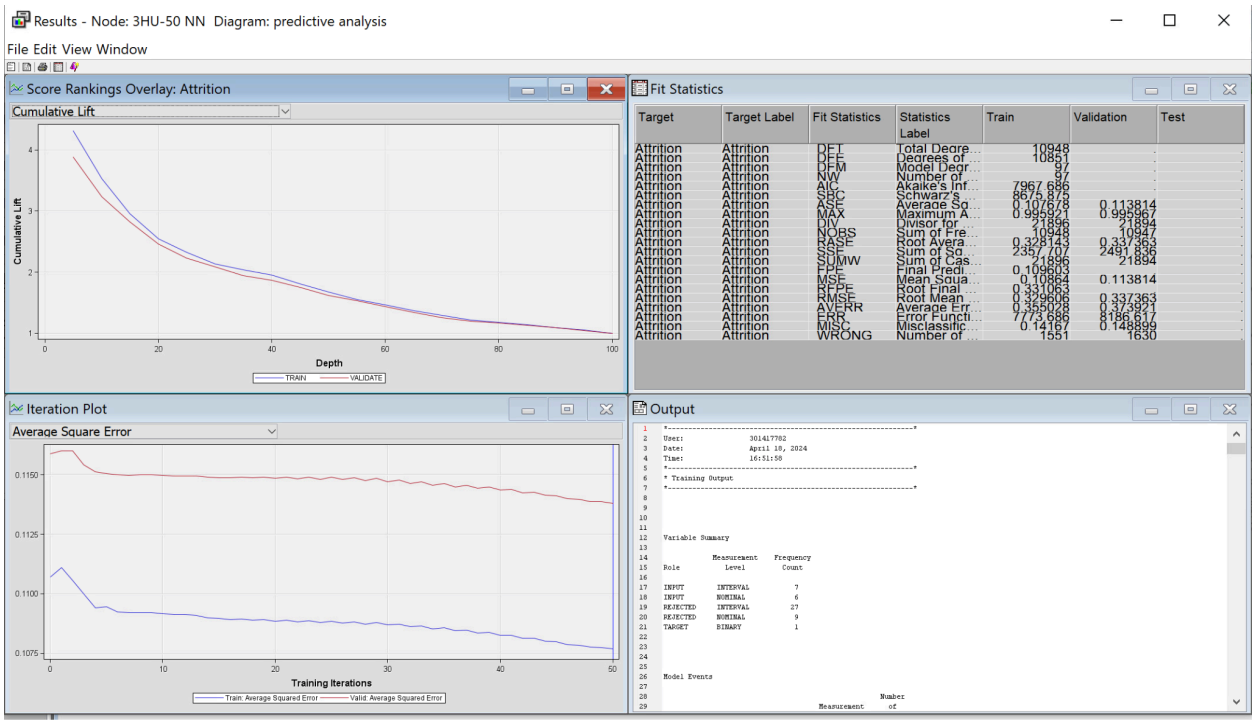


Here are the neural networks connected from the forward regression

Neural network 1:

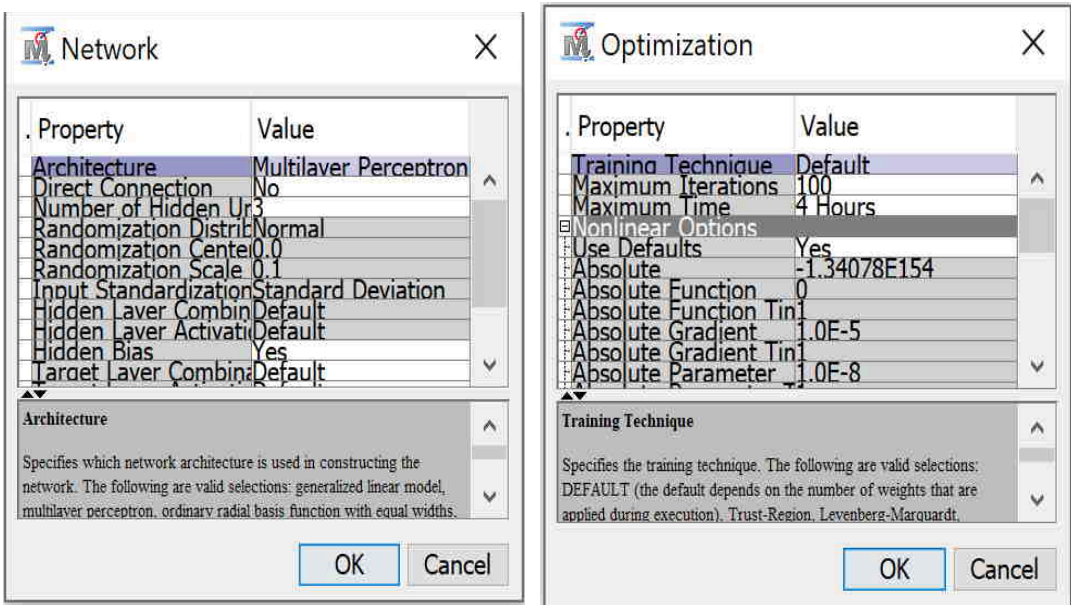


Here is the setting of the 1st neural network with **3 hidden units** and **50 iterations.**

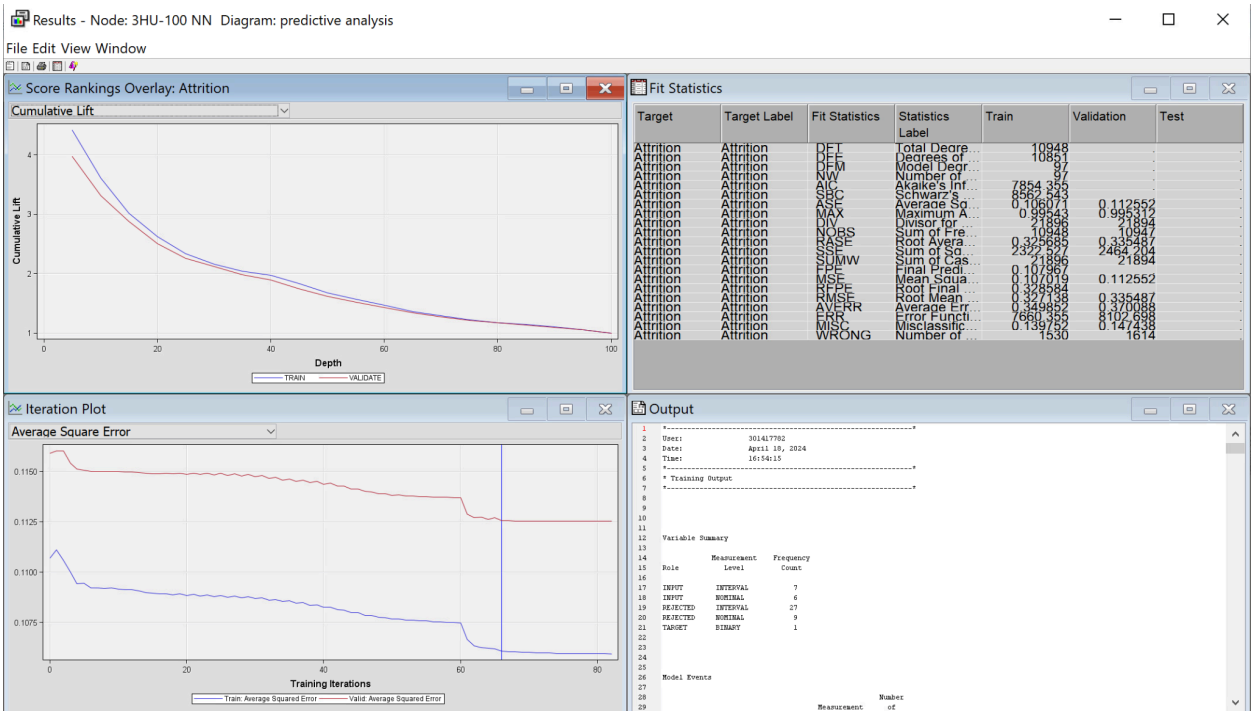


The first neural network with a configuration of 3 hidden units at 50 iterations had an average squared error of **0.113814**

Neural Network 2:

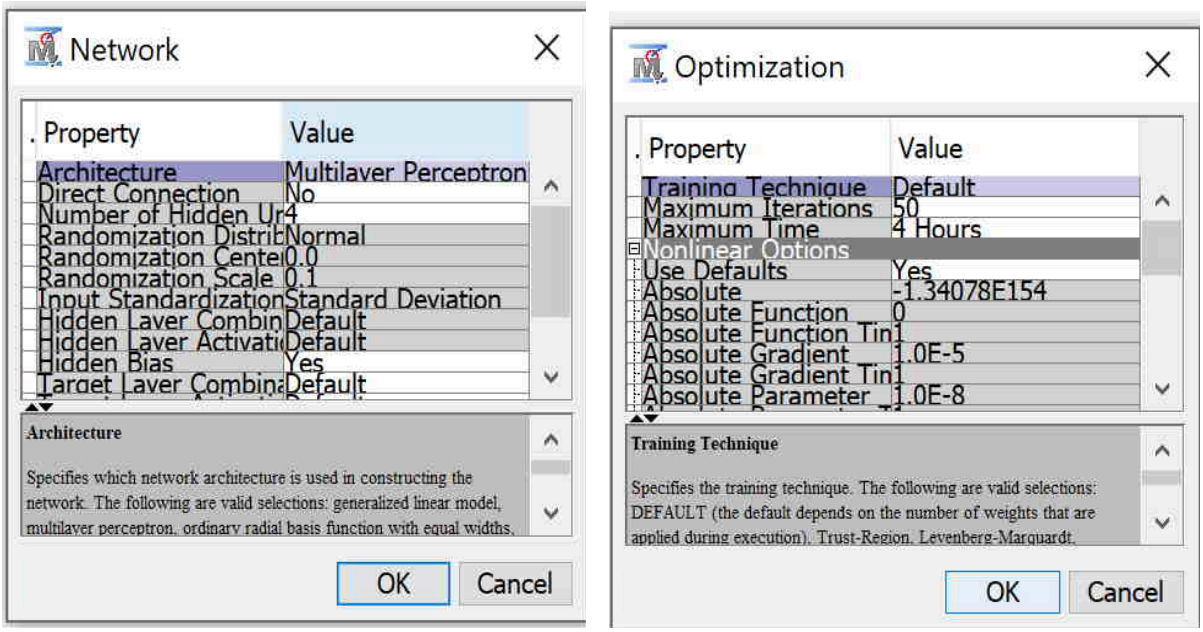


Here is the setting of the second neural network with **3 hidden units** and **100 iterations.**

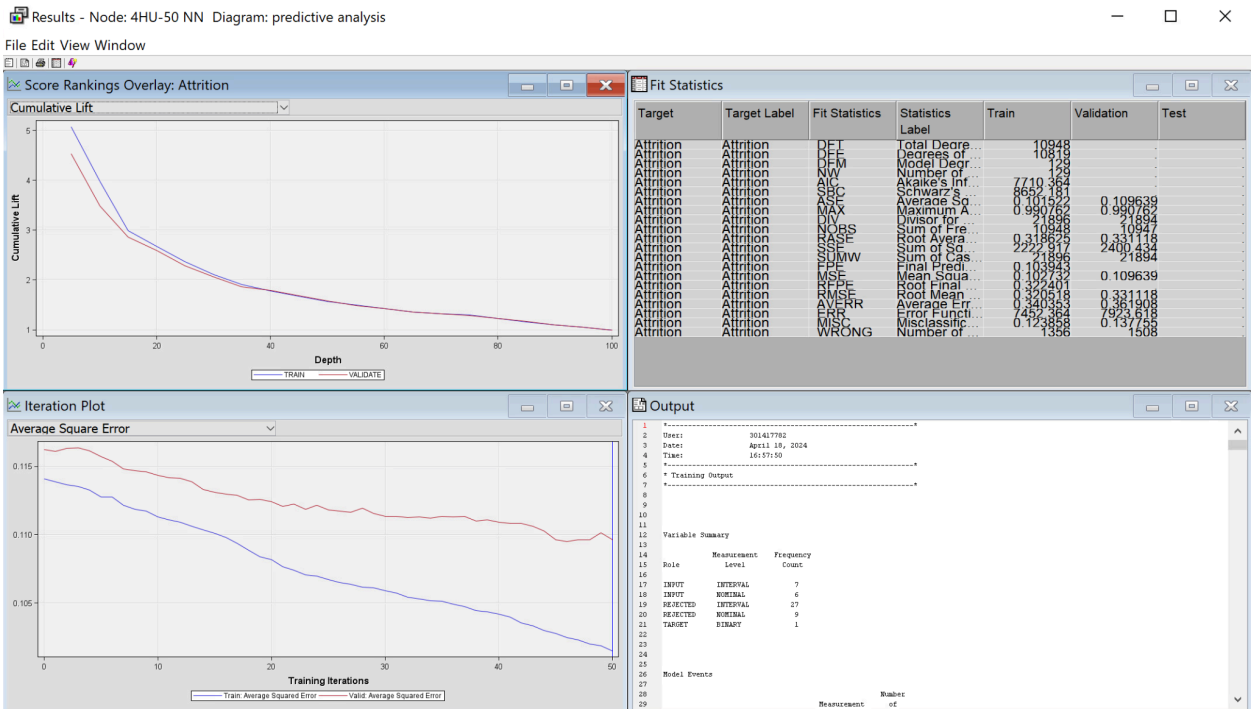


The second neural network with a configuration of 3 hidden units at 100 iterations had an average squared error of **0.112552.**

Neural Network 3:

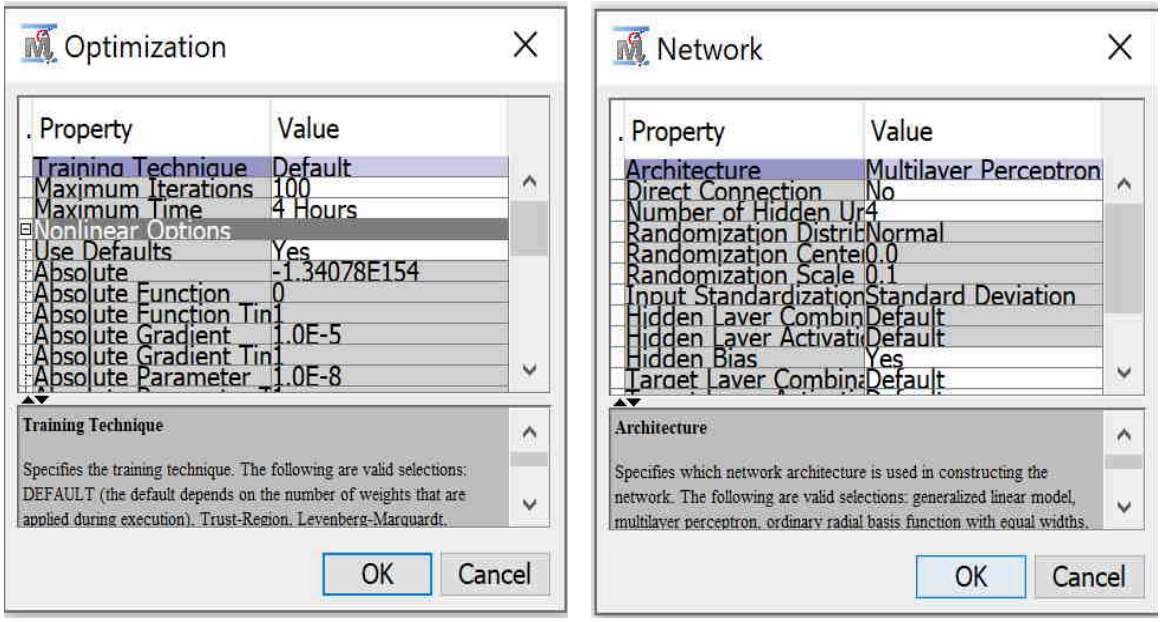


Here is the setting of the third neural network with **4 hidden units** and **50 iterations.**

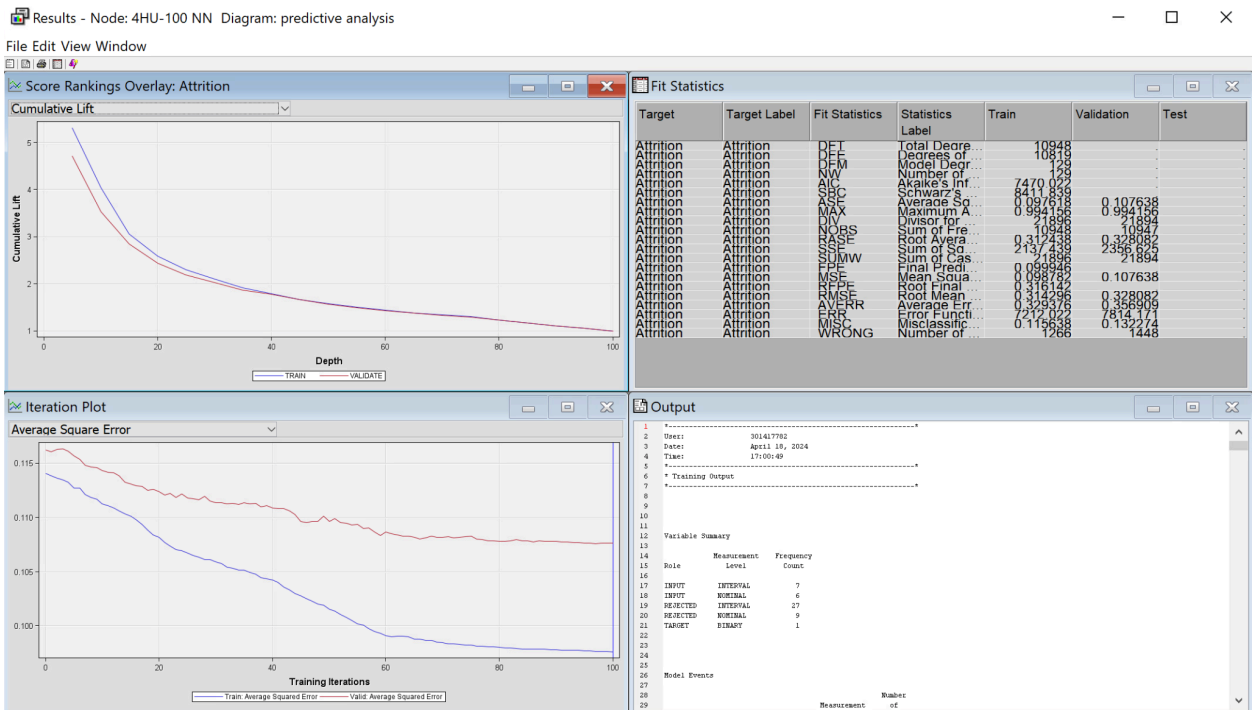


The third neural network with a configuration of 4 hidden units at 50 iterations had an average squared error of **0.109639.**

Neural Network 4:

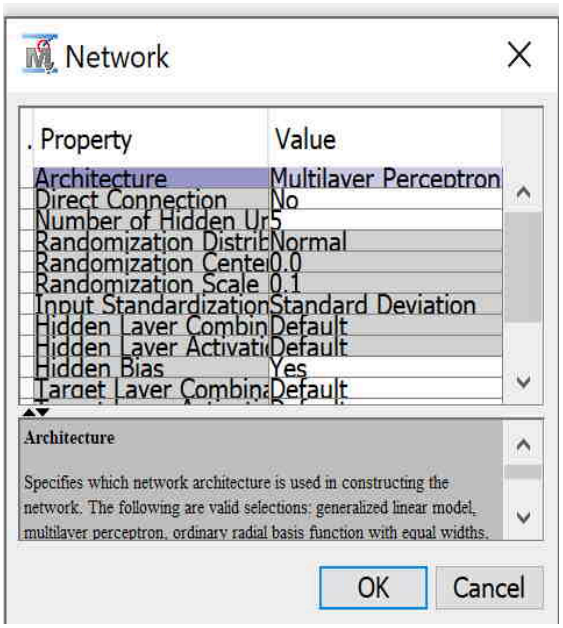


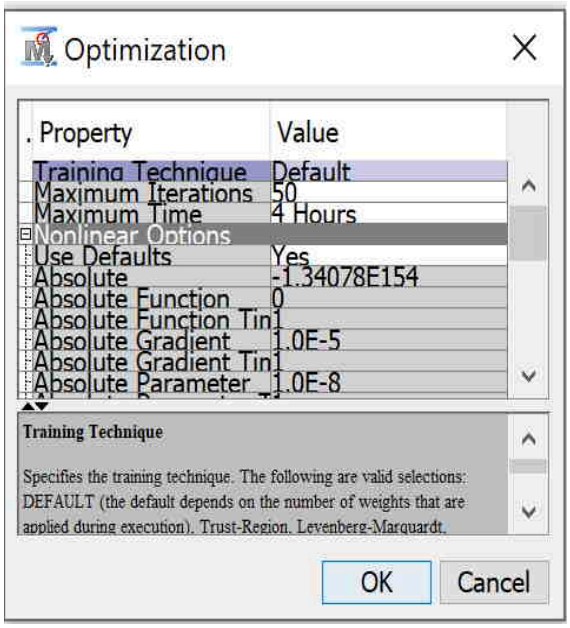
Here is the setting of the fourth neural network with **4 hidden units** and **100 iterations.**



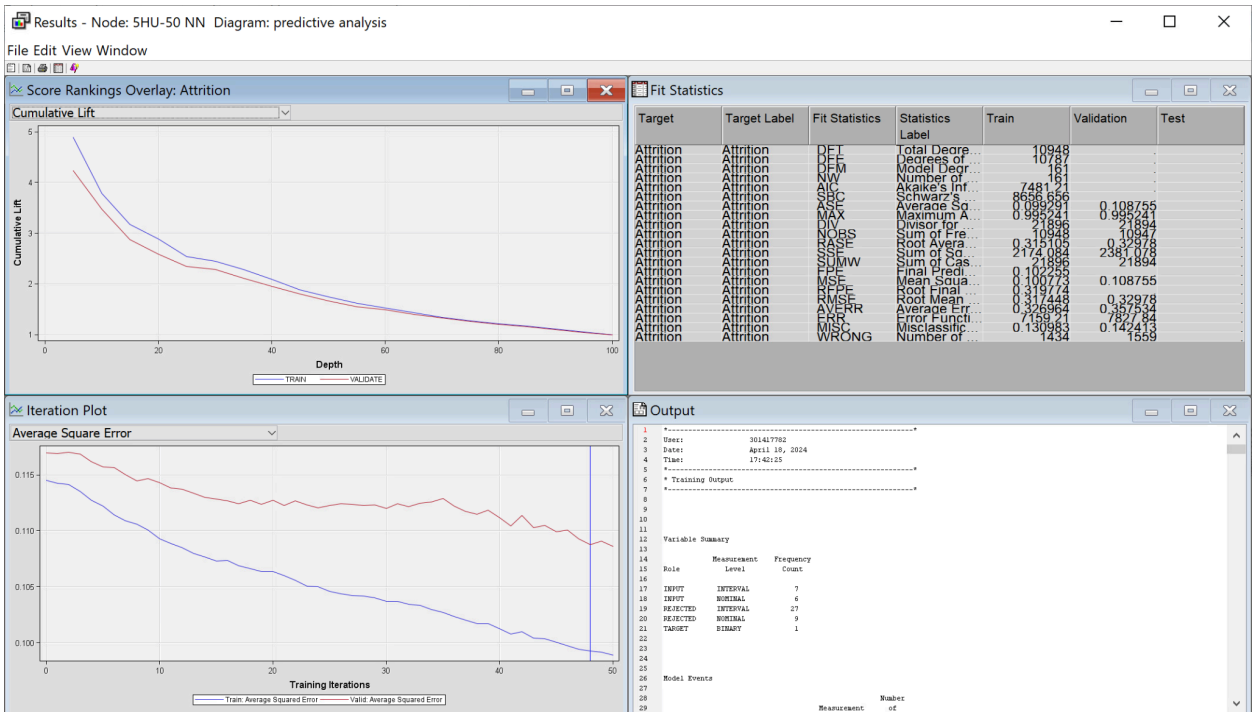
The fourth neural network with a configuration of 4 hidden units at 100 iterations had an average squared error of **0.107638.**

Neural Network 5:



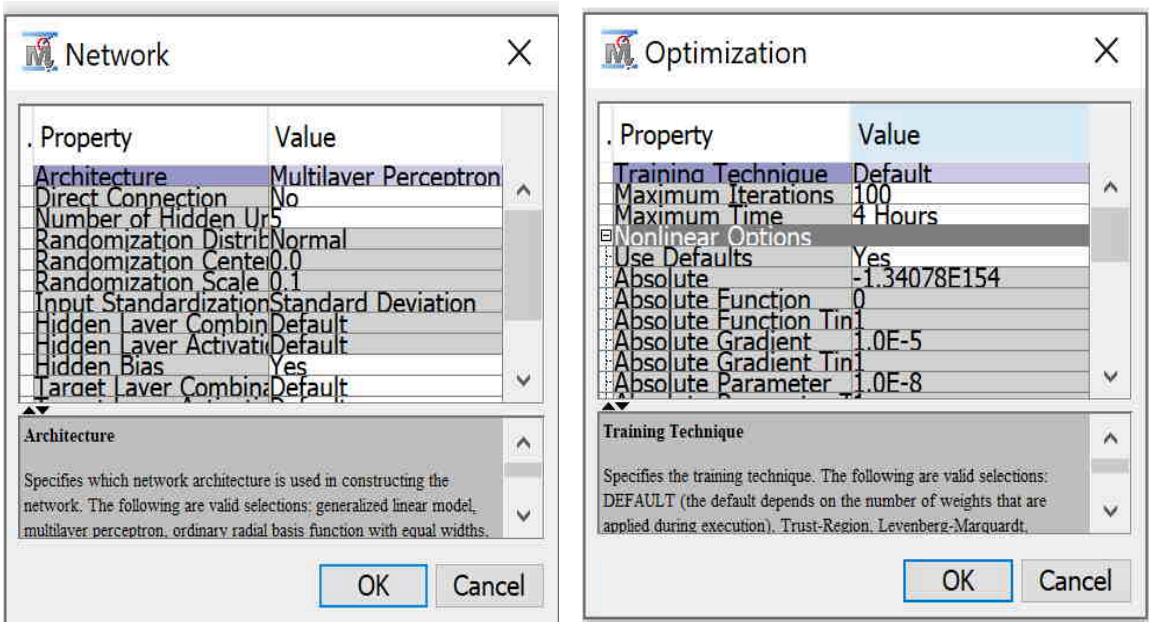


Here is the setting of the fifth neural network with **5 hidden units** and **50 iterations.**

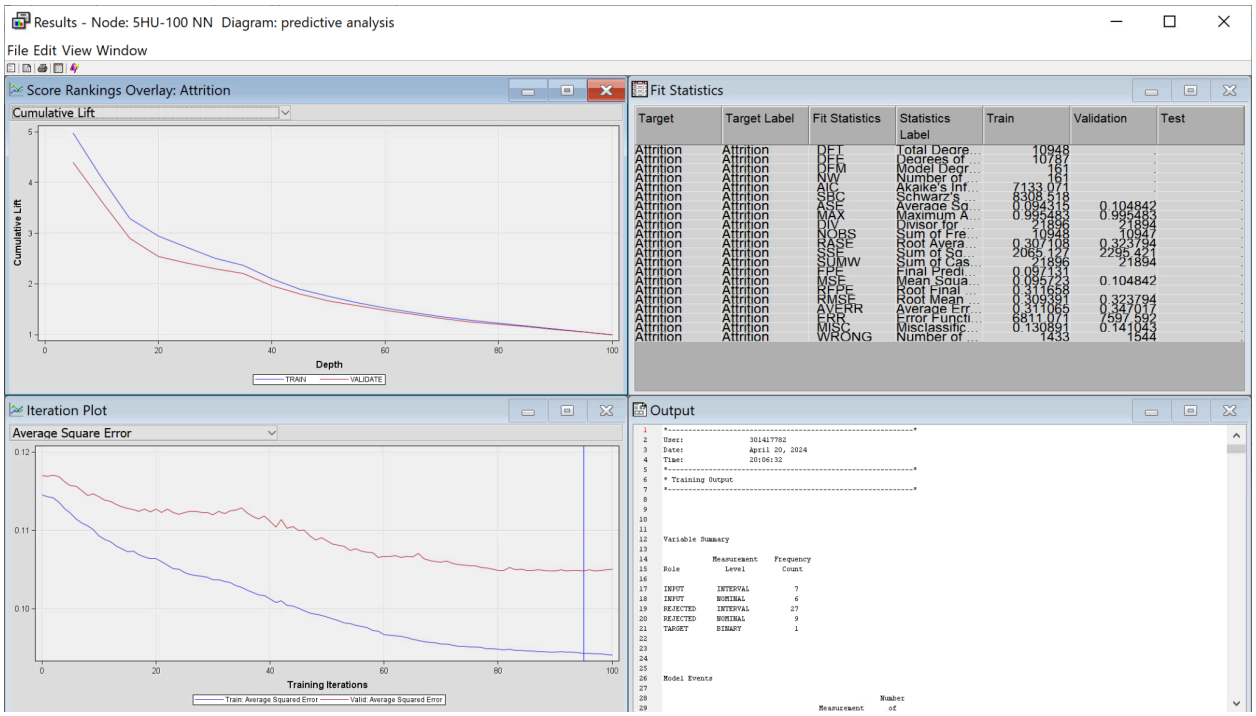


The fifth neural network with a configuration of 5 hidden units at 50 iterations had an average squared error of **0.108755**

Neural Network 6:



Here is the setting of the Sixth neural network with **5 hidden units** and **100 iterations.**



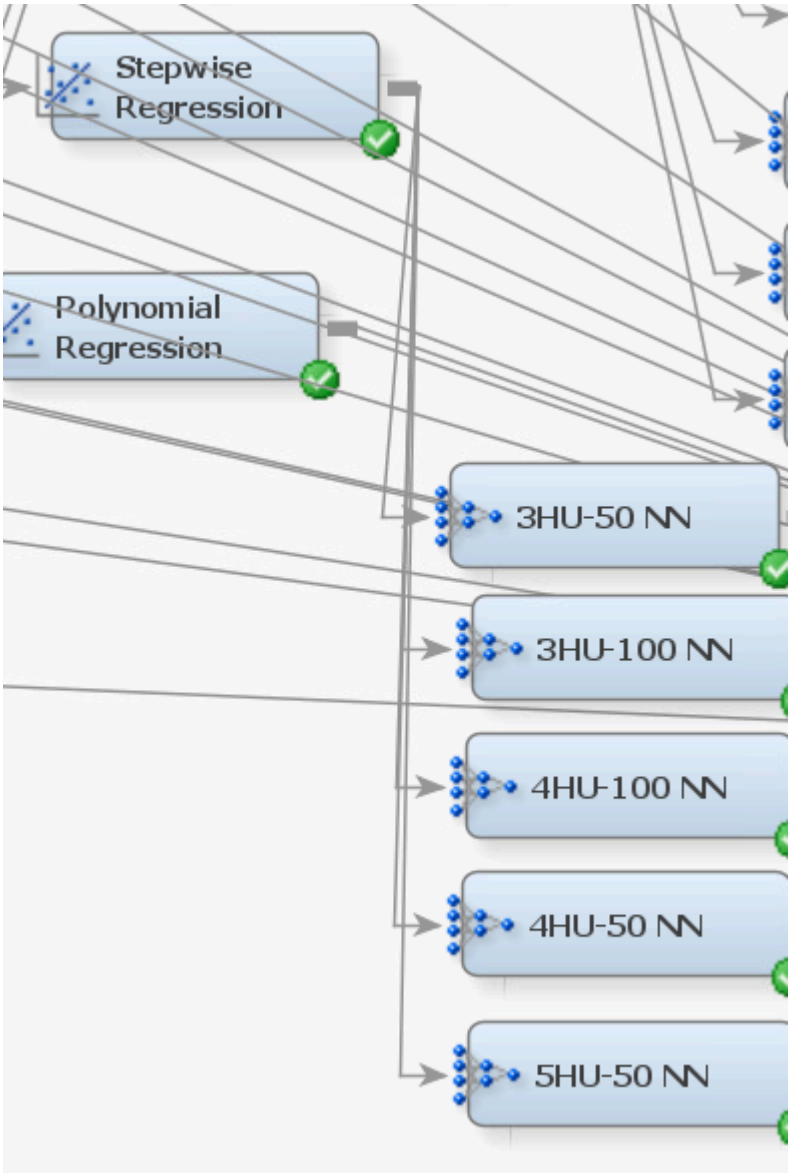
The Sixth neural network with a configuration of 5 hidden units at 100 iterations had an average squared error of **0.104842**

Here is the summary of the neural networks connected to the forward regression

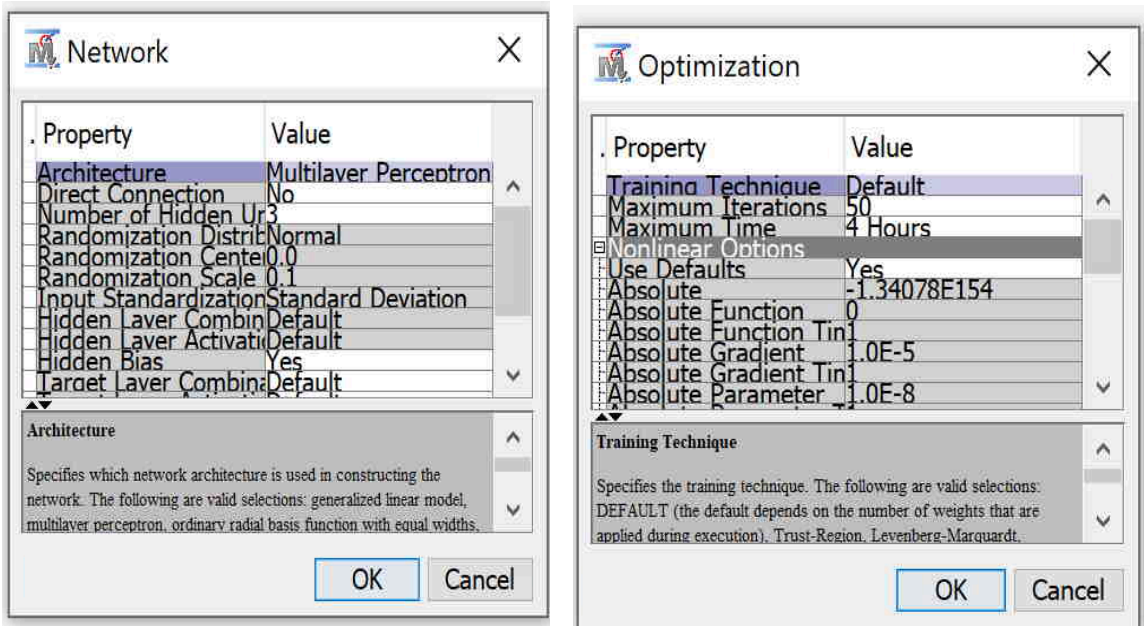
Neural Network	Hidden Units	Iterations	ASE
3HU-50NN	3	50	0.113814
3HU-100NN	3	100	0.112552.
4HU-50NN	4	50	0.109639.
4HU-100NN	4	100	0.107638.
5HU-50NN	5	50	0.108755
5HU-100NN	**5**	**100**	**0.104842**
6HU-50NN	6	50	0.10643

After performing 7 neural models, we arrived at our best model which is with 5 hidden units and 100 iterations. Another model was performed to get a more accurate result but the ASE was higher than before. Hence, we stopped our neural model as we got the best model with an ASE of 0.104842

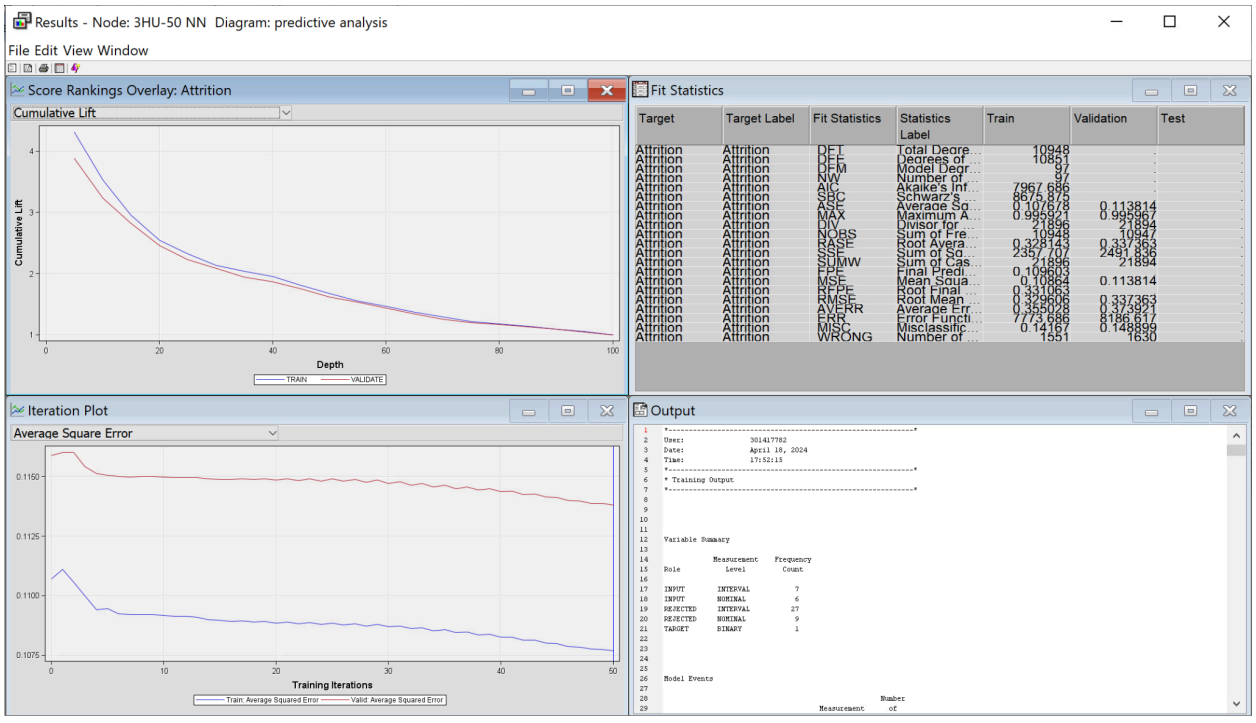
Neural Network connected from the stepwise regression:



Neural Network 1:

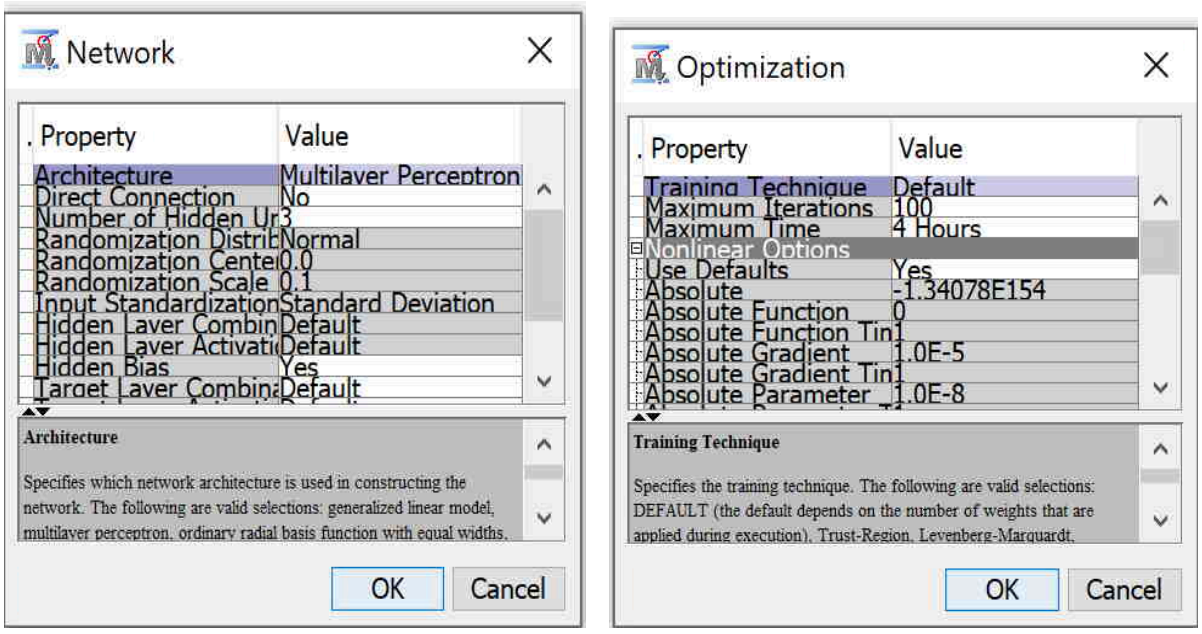


Here is the setting of the first neural network with **3 hidden units** and **50 iterations.**

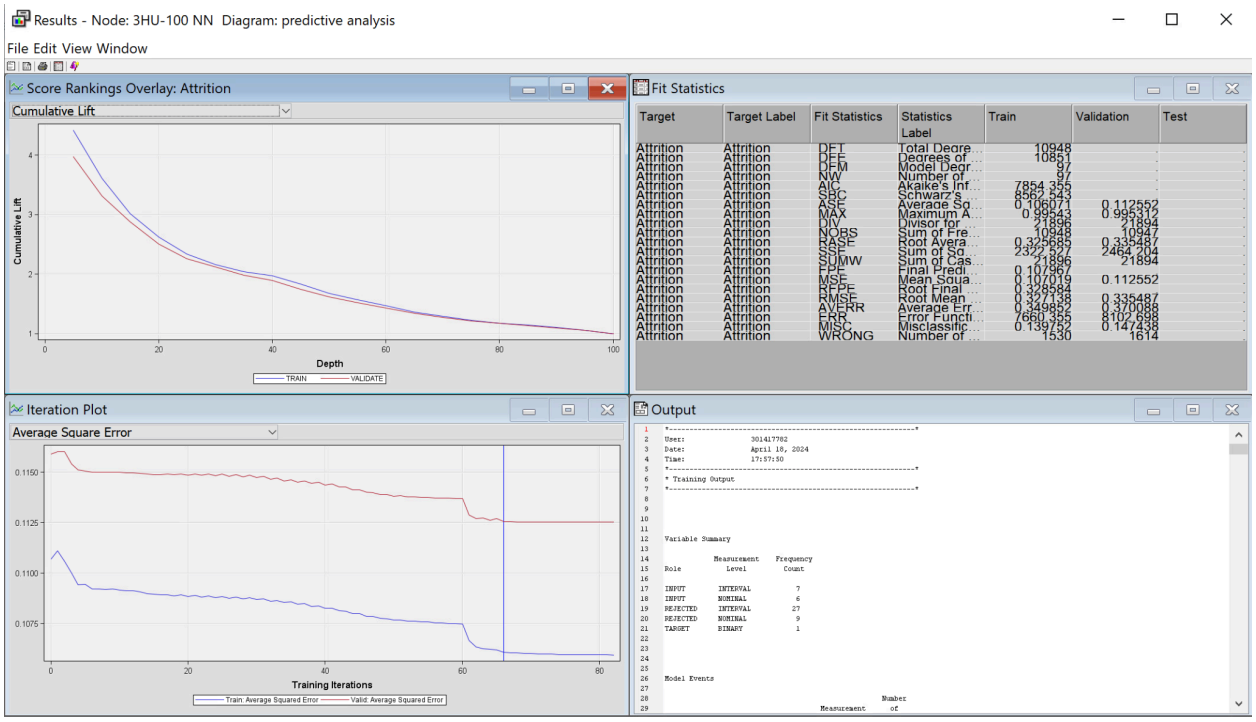


The first neural network with a configuration of 3 hidden units at 50 iterations had an average squared error of **0.113814**

Neural Network 2:

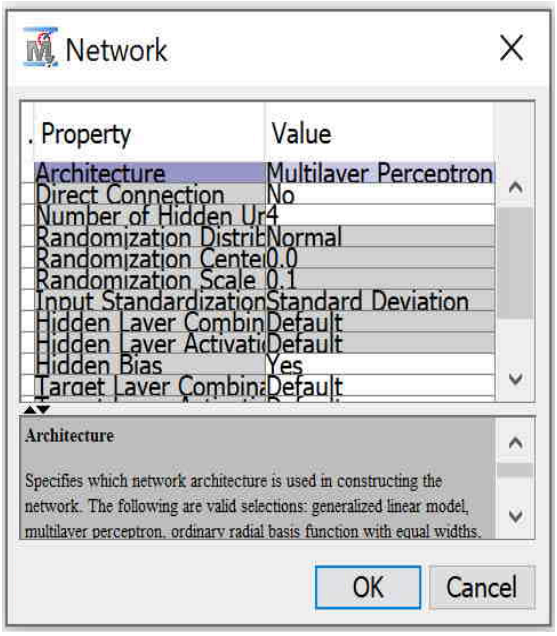


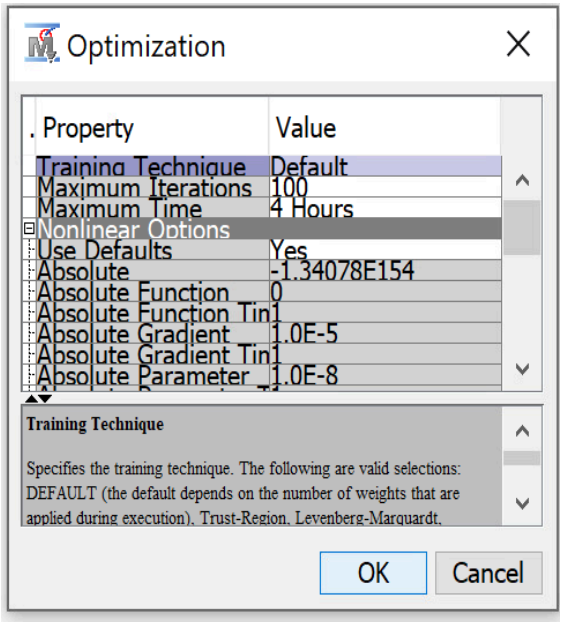
Here is the setting of the second neural network with **3 hidden units** and **100 iterations.**



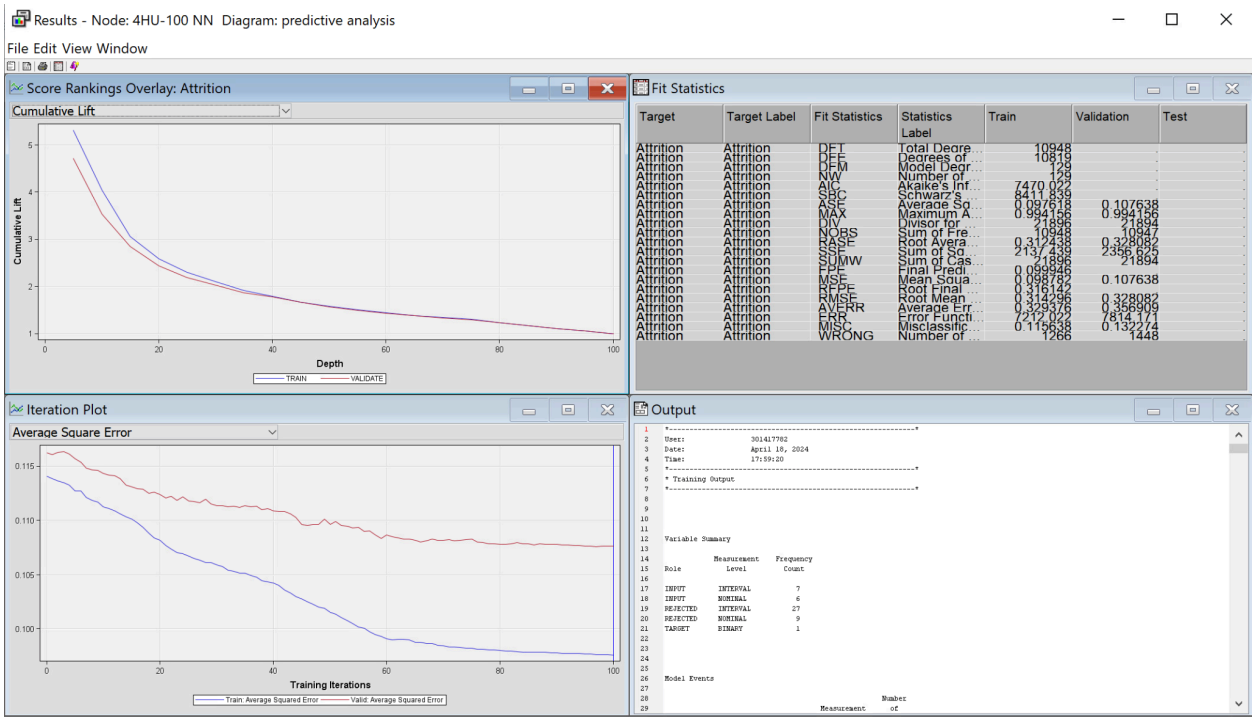
The second neural network with a configuration of 3 hidden units at 100 iterations had an average squared error of **0.112552**

Neural Network 3:



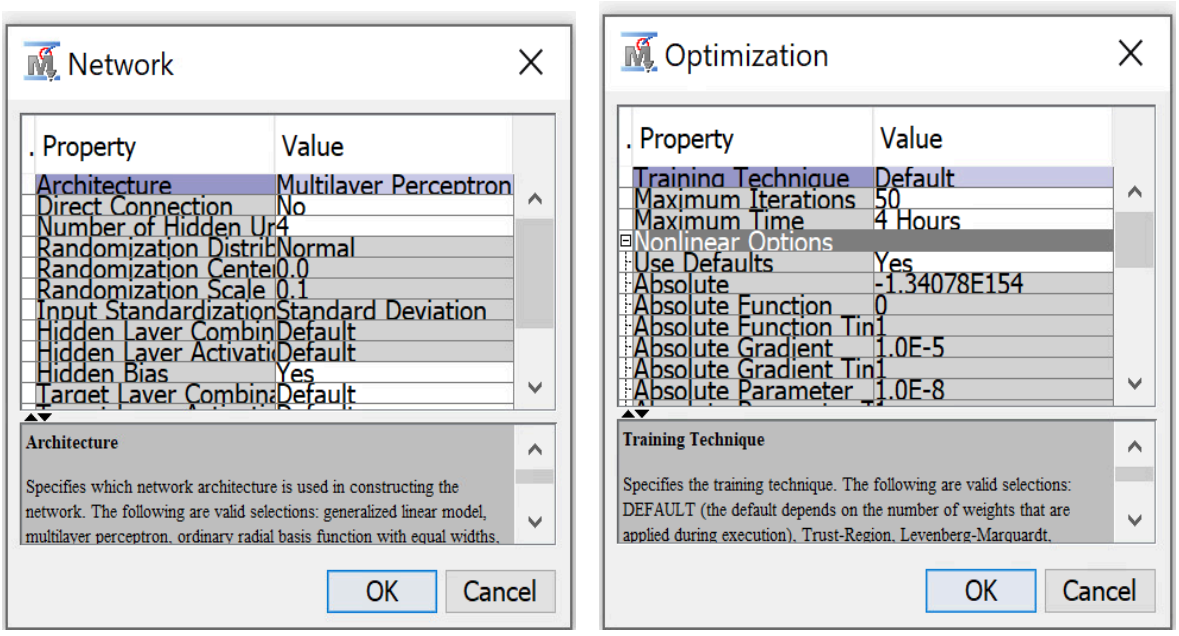


Here is the setting of the third neural network with **4 hidden units** and **100 iterations.**

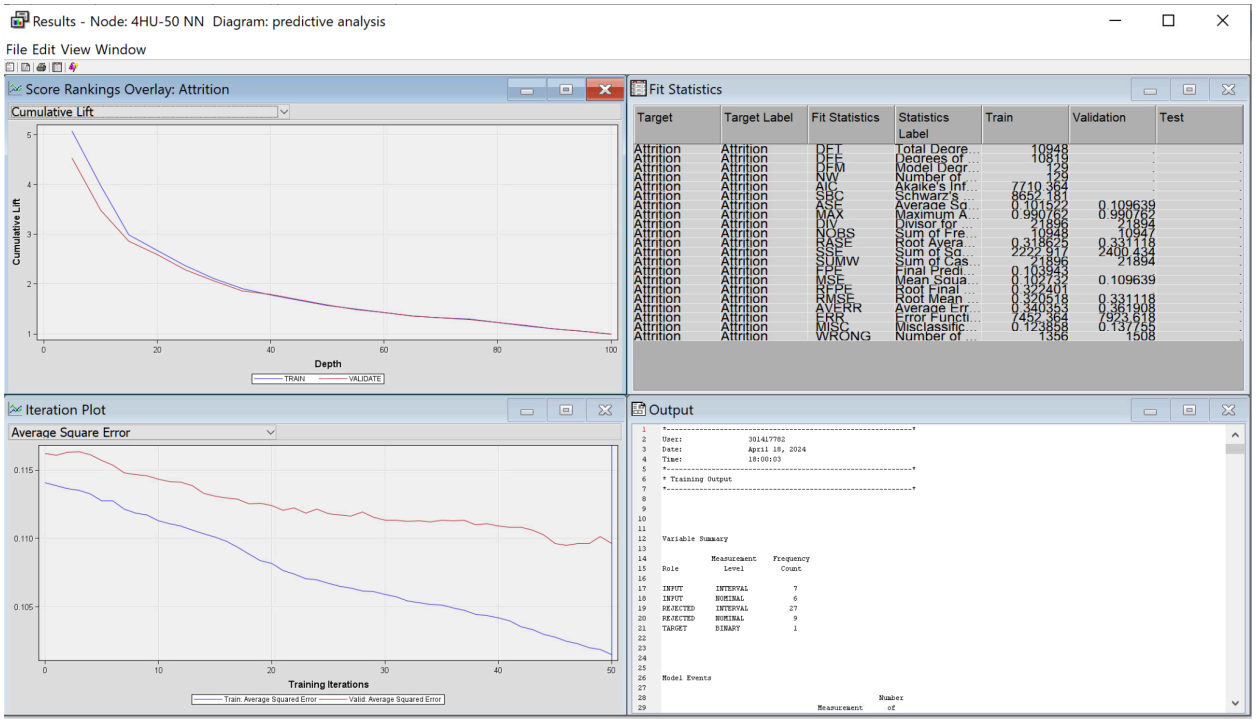


The third neural network with a configuration of 4 hidden units at 100 iterations had an average squared error of **0.107638**

Neural Network 4:

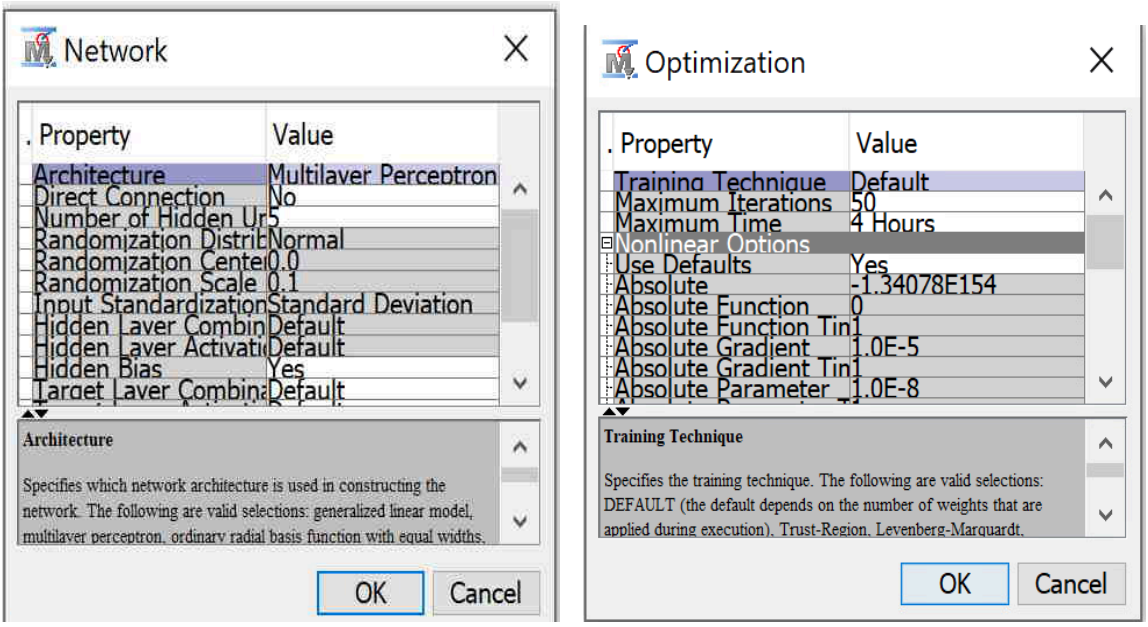


Here is the setting of the third neural network with **4 hidden units** and **50 iterations.**

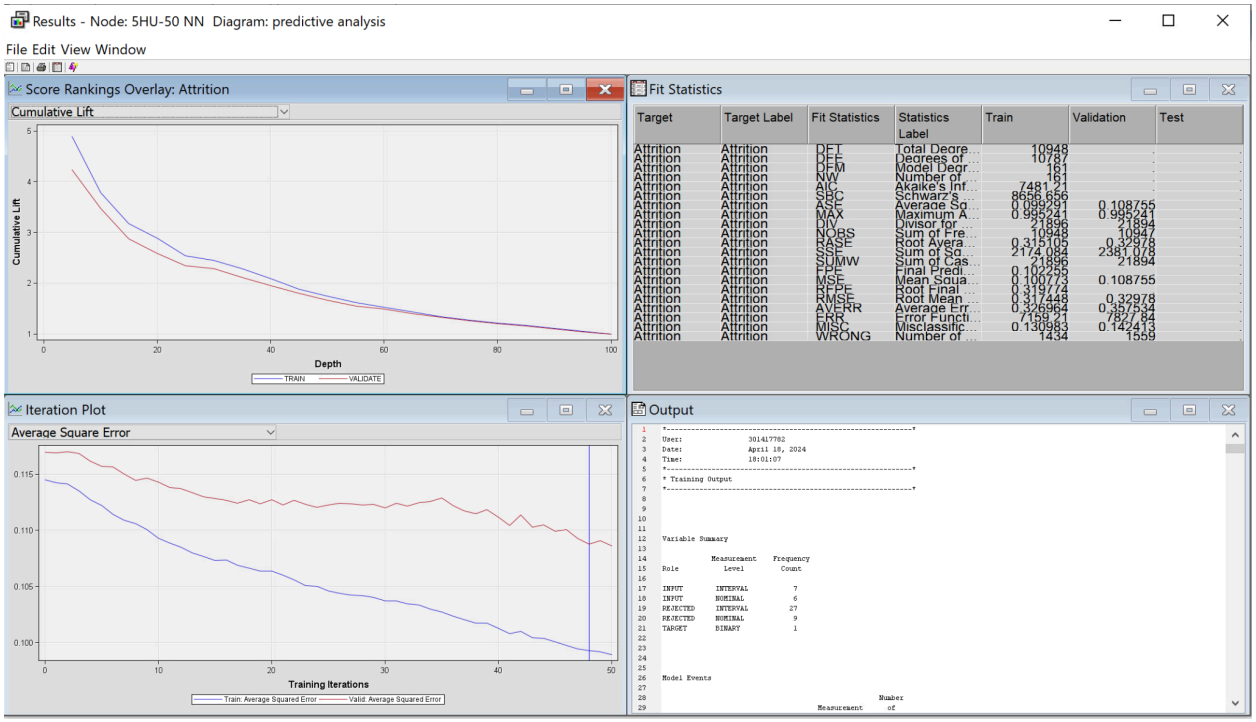


The third neural network with a configuration of 4 hidden units at 50 iterations had an average squared error of **0.109639**

Neural Network 5:

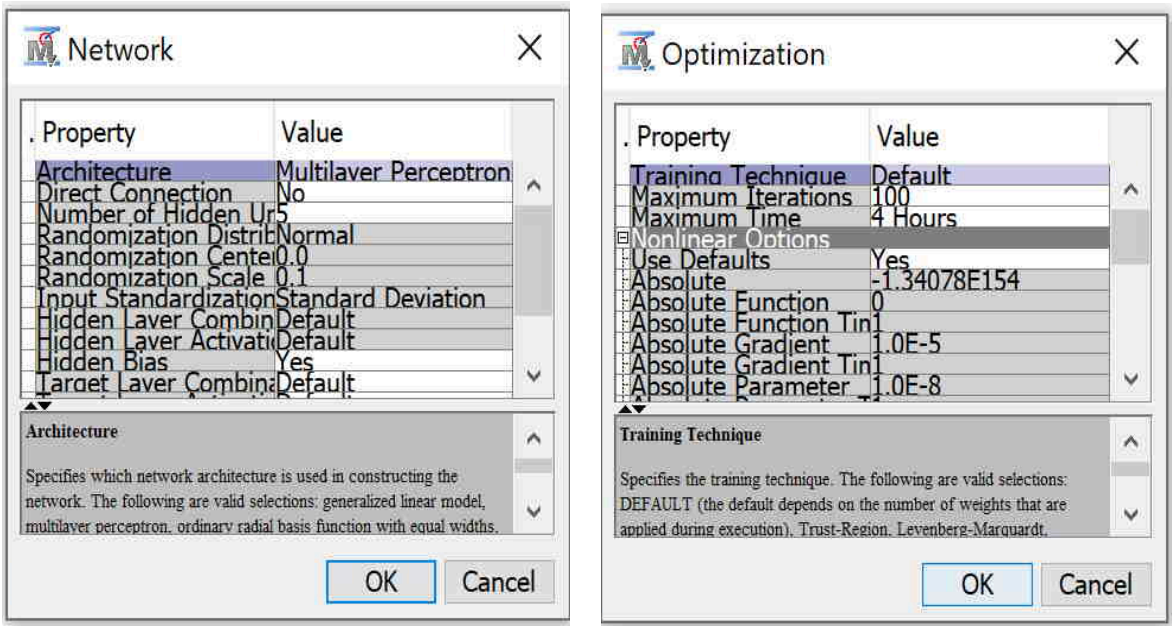


Here is the setting of the third neural network with **5 hidden units** and **50 iterations.**

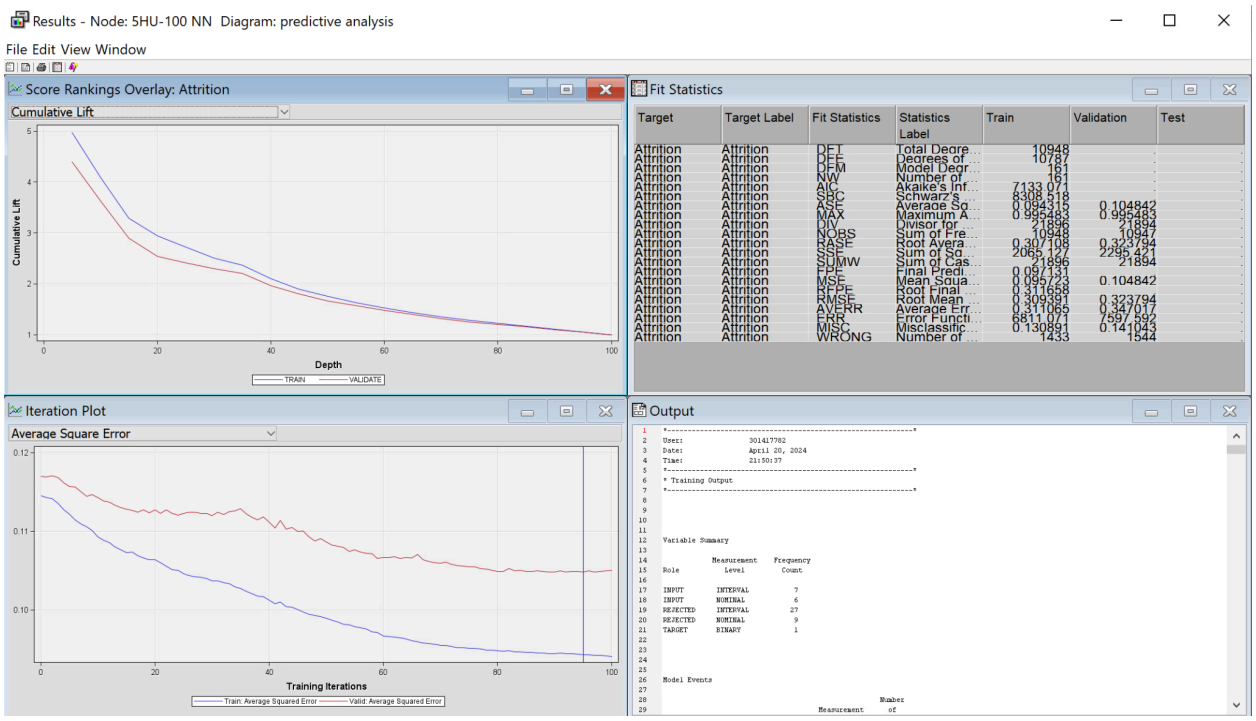


The third neural network with a configuration of 5 hidden units at 50 iterations had an average squared error of **0.108755**

Neural Network 6:



Here is the setting of the third neural network with **5 hidden units** and **100 iterations.**



The third neural network with a configuration of 5 hidden units at 100 iterations had an average squared error of **0.104842**

Here is the summary of the neural networks connected to the stepwise regression.

Neural Network	Hidden Units	Iterations	ASE
3HU-50NN	3	50	0.113814
3HU-100NN	3	100	0.112552.
4HU-50NN	4	50	0.109639.
4HU-100NN	4	100	0.107638.
5HU-50NN	5	50	0.108755
5HU-100NN	**5**	**100**	**0.104842**

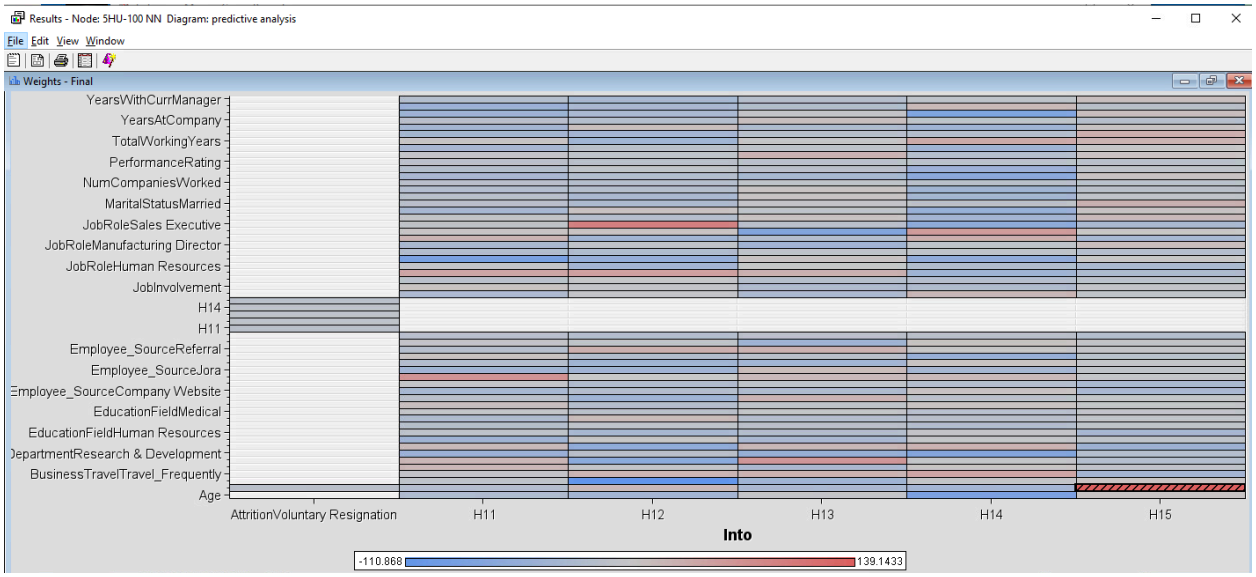
After performing 6 neural models, we arrived at our best model which is with 5 hidden units and 100 iterations. Another model was performed to get a more accurate result but the ASE was higher than before. Hence, we stopped our neural model as we got the best model with an ASE of 0.104842.

Neural Network Summary:

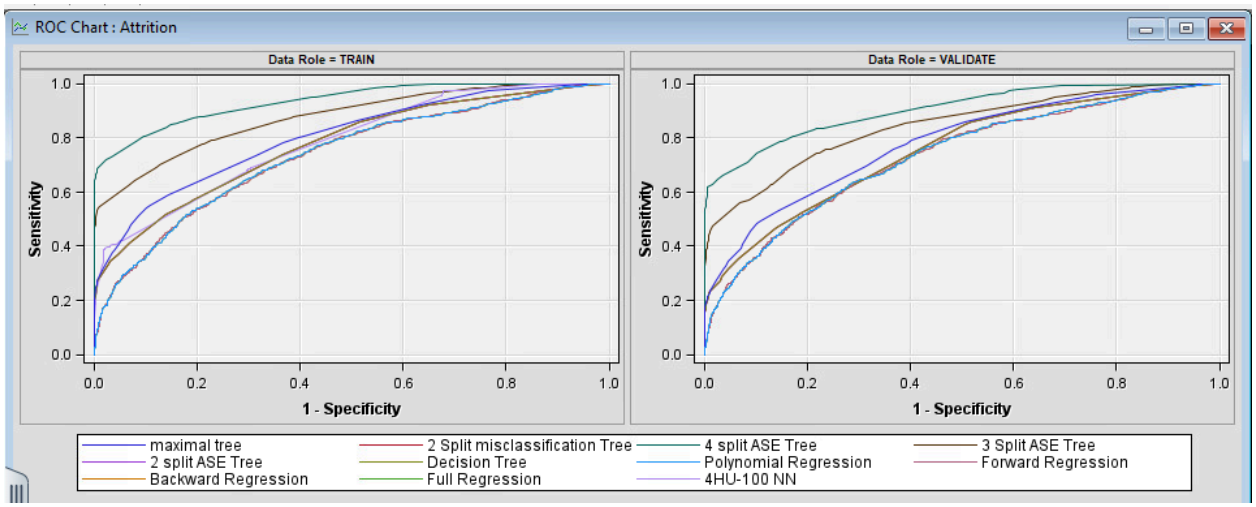
A total of 20 neural networks with different configurations were implemented for IBM employee data set.

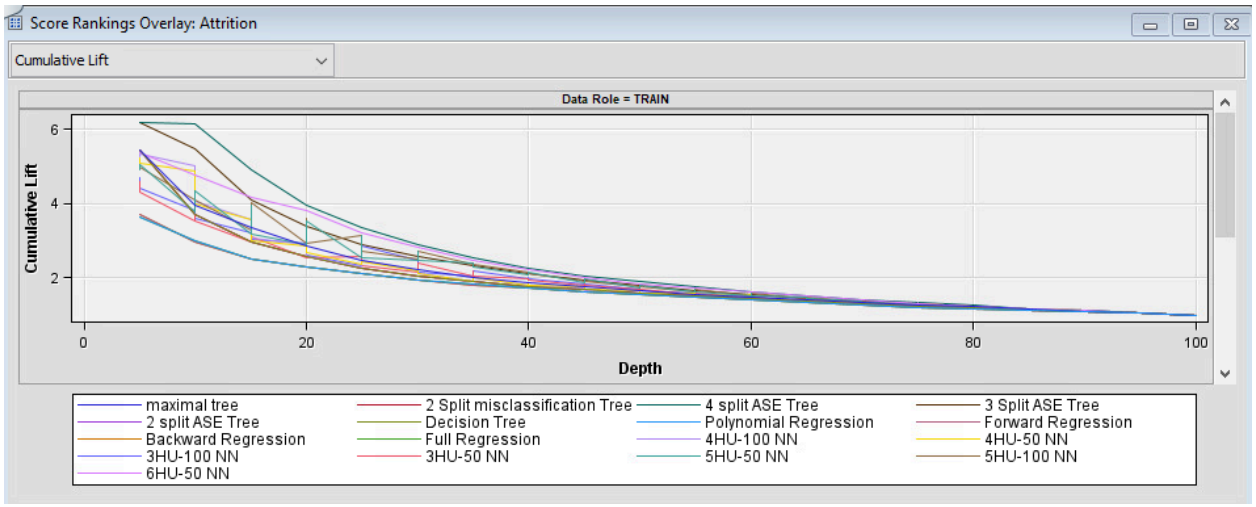
Neural Network	Hidden Units	Iterations	ASE
Imputed, 3HU-50NN	3	50	0.110849
Imputed, 3HU-100NN	3	100	0.109456
Imputed, 4HU-50NN	4	50	0.103028
Imputed, 4HU-100NN	4	100	0.099556
Imputed, 5HU-50NN	5	50	0.096481
Imputed, **5HU-100NN**	**5**	**100**	**0.94422**
Imputed, 6HU-50NN	6	50	0.96151
Forward, 3HU-50NN	3	50	0.113814
Forward, 3HU-100NN	3	100	0.112552.
Forward, 4HU-50NN	4	50	0.109639.
Forward, 4HU-100NN	4	100	0.107638.
Forward, 5HU-50NN	5	50	0.108755
Forward, **5HU-100NN**	**5**	**100**	**0.104842**
Forward, 6HU-50NN	6	50	0.10643
Stepwise, 3HU-50NN	3	50	0.113814
Stepwise, 3HU-100NN	3	100	0.112552.
Stepwise, 4HU-50NN	4	50	0.109639.
Stepwise, 4HU-100NN	4	100	0.107638.
Stepwise, 5HU-50NN	5	50	0.108755

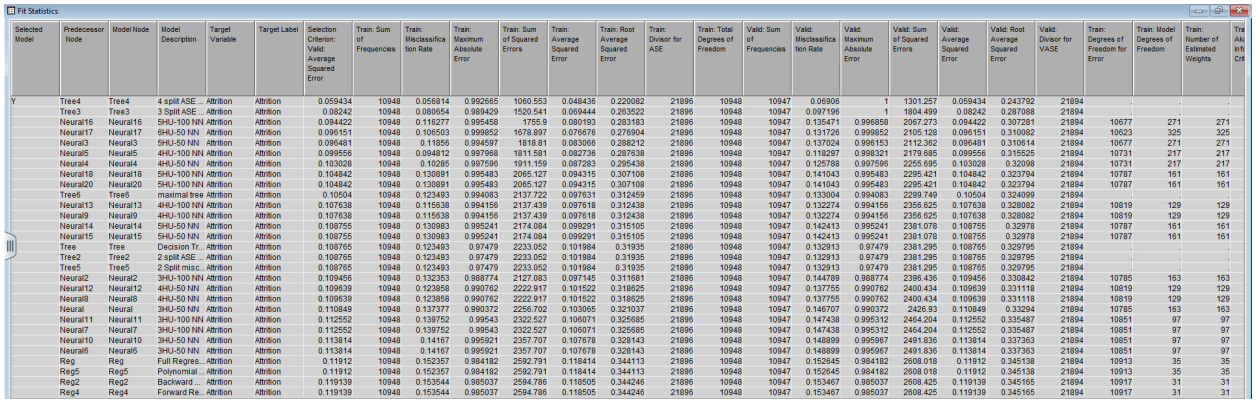
Stepwise, **5HU-100NN**	**5**	**100**	**0.104842**



MODEL COMPARISON







Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	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	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights
Y	Tree4	Tree4	4 split ASE ...	Attrition	Attrition	0.059434	10948	0.056814	0.992665	1060.553	0.048436	0.220082	21896	10948	10947	0.06906	1	1301.257	0.059434	0.243792	21894			
	Tree3	Tree3	3 Split ASE ...	Attrition	Attrition	0.08242	10948	0.080654	0.989429	1520.541	0.069444	0.263522	21896	10948	10947	0.097196	1	1804.499	0.08242	0.287088	21894			
	Neural16	Neural16	5HU-100 NN	Attrition	Attrition	0.094422	10948	0.116277	0.995458	1755.9	0.080193	0.283183	21896	10948	10947	0.135471	0.996858	2067.273	0.094422	0.307281	21894	10677	271	271
	Neural17	Neural17	6HU-50 NN	Attrition	Attrition	0.096151	10948	0.106503	0.999852	1679.897	0.076676	0.276904	21896	10948	10947	0.131726	0.999852	2195.129	0.096151	0.310082	21894	10623	325	325
	Neural3	Neural3	5HU-50 NN	Attrition	Attrition	0.096481	10948	0.11856	0.994597	1818.81	0.083066	0.288212	21896	10948	10947	0.137024	0.996153	2112.362	0.096481	0.310614	21894	10677	271	271
	Neural5	Neural5	4HU-100 NN	Attrition	Attrition	0.099556	10948	0.094812	0.997968	1811.581	0.082736	0.287638	21896	10948	10947	0.118297	0.998321	2179.685	0.099556	0.315525	21894	10731	217	217
	Neural4	Neural4	4HU-50 NN	Attrition	Attrition	0.103028	10948	0.10285	0.997596	1911.159	0.087283	0.295438	21896	10948	10947	0.125788	0.997596	2255.695	0.103028	0.32098	21894	10731	217	217
	Neural18	Neural18	5HU-100 NN	Attrition	Attrition	0.104842	10948	0.130891	0.995483	2065.127	0.094315	0.307108	21896	10948	10947	0.141043	0.995483	2295.421	0.104842	0.323794	21894	10787	161	161
	Neural20	Neural20	5HU-100 NN	Attrition	Attrition	0.104842	10948	0.130891	0.995483	2065.127	0.094315	0.307108	21896	10948	10947	0.141043	0.995483	2295.421	0.104842	0.323794	21894	10787	161	161
	Tree6	Tree6	maximal tree	Attrition	Attrition	0.10504	10948	0.123493	0.994083	2137.722	0.097631	0.312459	21896	10948	10947	0.133004	0.994083	2299.749	0.10504	0.324099	21894			
	Neural13	Neural13	4HU-100 NN	Attrition	Attrition	0.107638	10948	0.115638	0.994156	2137.439	0.097618	0.312438	21896	10948	10947	0.132274	0.994156	2356.625	0.107638	0.328082	21894	10819	129	129
	Neural9	Neural9	4HU-100 NN	Attrition	Attrition	0.107638	10948	0.115638	0.994156	2137.439	0.097618	0.312438	21896	10948	10947	0.132274	0.994156	2356.625	0.107638	0.328082	21894	10819	129	129
	Neural14	Neural14	5HU-50 NN	Attrition	Attrition	0.108755	10948	0.130983	0.995241	2174.084	0.099291	0.315105	21896	10948	10947	0.142413	0.995241	2381.078	0.108755	0.32978	21894	10787	161	161
	Neural15	Neural15	5HU-50 NN	Attrition	Attrition	0.108755	10948	0.130983	0.995241	2174.084	0.099291	0.315105	21896	10948	10947	0.142413	0.995241	2381.078	0.108755	0.32978	21894	10787	161	161
	Tree	Tree	Decision Tr...	Attrition	Attrition	0.108765	10948	0.123493	0.97479	2233.052	0.101984	0.31935	21896	10948	10947	0.132913	0.97479	2381.295	0.108765	0.329795	21894			
	Tree2	Tree2	2 split ASE ...	Attrition	Attrition	0.108765	10948	0.123493	0.97479	2233.052	0.101984	0.31935	21896	10948	10947	0.132913	0.97479	2381.295	0.108765	0.329795	21894			
	Tree5	Tree5	2 Split misc...	Attrition	Attrition	0.108765	10948	0.123493	0.97479	2233.052	0.101984	0.31935	21896	10948	10947	0.132913	0.97479	2381.295	0.108765	0.329795	21894			
	Neural2	Neural2	3HU-100 NN	Attrition	Attrition	0.109456	10948	0.132353	0.988774	2127.083	0.097145	0.311681	21896	10948	10947	0.144789	0.988774	2396.436	0.109456	0.330842	21894	10785	163	163
	Neural12	Neural12	4HU-50 NN	Attrition	Attrition	0.109639	10948	0.123858	0.990762	2222.917	0.101522	0.318625	21896	10948	10947	0.137755	0.990762	2400.434	0.109639	0.331118	21894	10819	129	129
	Neural8	Neural8	4HU-50 NN	Attrition	Attrition	0.109639	10948	0.123858	0.990762	2222.917	0.101522	0.318625	21896	10948	10947	0.137755	0.990762	2400.434	0.109639	0.331118	21894	10819	129	129
	Neural	Neural	3HU-50 NN	Attrition	Attrition	0.110849	10948	0.137377	0.990372	2256.702	0.103065	0.321037	21896	10948	10947	0.146707	0.990372	2426.93	0.110849	0.33294	21894	10785	163	163
	Neural11	Neural11	3HU-100 NN	Attrition	Attrition	0.112552	10948	0.139752	0.99543	2322.527	0.106071	0.325685	21896	10948	10947	0.147438	0.995312	2464.204	0.112552	0.335487	21894	10851	97	97
	Neural7	Neural7	3HU-100 NN	Attrition	Attrition	0.112552	10948	0.139752	0.99543	2322.527	0.106071	0.325685	21896	10948	10947	0.147438	0.995312	2464.204	0.112552	0.335487	21894	10851	97	97
	Neural10	Neural10	3HU-50 NN	Attrition	Attrition	0.113814	10948	0.14167	0.995921	2357.707	0.107678	0.328143	21896	10948	10947	0.148899	0.995967	2491.836	0.113814	0.337363	21894	10851	97	97
	Neural6	Neural6	3HU-50 NN	Attrition	Attrition	0.113814	10948	0.14167	0.995921	2357.707	0.107678	0.328143	21896	10948	10947	0.148899	0.995967	2491.836	0.113814	0.337363	21894	10851	97	97
	Reg	Reg	Full Regres...	Attrition	Attrition	0.11912	10948	0.152357	0.984182	2592.791	0.118414	0.344113	21896	10948	10947	0.152645	0.984182	2608.018	0.11912	0.345138	21894	10913	35	35
	Reg5	Reg5	Polynomial ...	Attrition	Attrition	0.11912	10948	0.152357	0.984182	2592.791	0.118414	0.344113	21896	10948	10947	0.152645	0.984182	2608.018	0.11912	0.345138	21894	10913	35	35
	Reg2	Reg2	Backward ...	Attrition	Attrition	0.119139	10948	0.153544	0.985037	2594.786	0.118505	0.344246	21896	10948	10947	0.153467	0.985037	2608.425	0.119139	0.345165	21894	10917	31	31
	Reg4	Reg4	Forward Re...	Attrition	Attrition	0.119139	10948	0.153544	0.985037	2594.786	0.118505	0.344246	21896	10948	10947	0.153467	0.985037	2608.425	0.119139	0.345165	21894	10917	31	31

From the model comparison found out that Tree 4 is the best model
With training ASE =0.048436.

SUMMARY

Tree 4 is the best model among the models we ran, in neural network neural 16 is best network.

From these models we can validate the attrition rate and predict employes having certain characteristics are tend to leave the company.