**BA 706**

**Final Group Project**

**Predicting & Understanding Loan Defaults.**

**Team -**

**Dev Desai - 301473593**

**Sherwin D’souza – 301483113**

**Venmukil Maharajan - 301431272**

**Dataset:**

[**Loan Default**](https://www.kaggle.com/datasets/nikhil1e9/loan-default)

**Contents:**

1. Introduction
2. Diagram
3. Stat Explore
4. Decision Trees
5. Regression Models
6. Neural Networks
7. Conclusion

**Introduction**

Understanding the factors contributing to loan defaults is crucial for banks and financial institutions. Loan defaults not only affect a bank's profitability but also pose risks to its long-term stability. Identifying the key drivers behind loan defaults can enable banks to make informed lending decisions, reduce risk exposure, and enhance the credit assessment process.

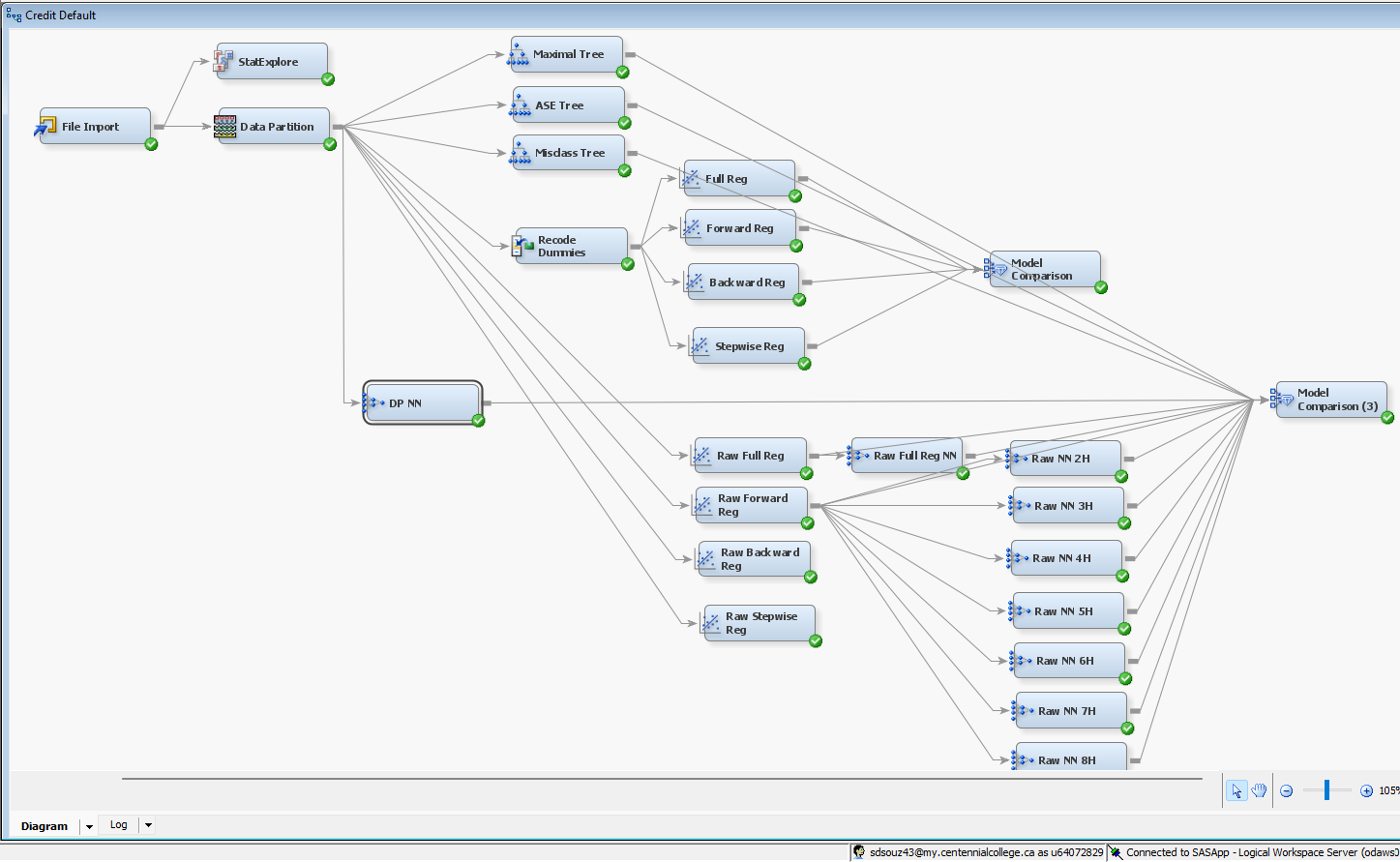
In this project, we are going to employ a range of analytical techniques like **Decision Trees**, **Regression Models**, and **Neural Networks** to uncover patterns and relationships within the data. These methods allow us to:

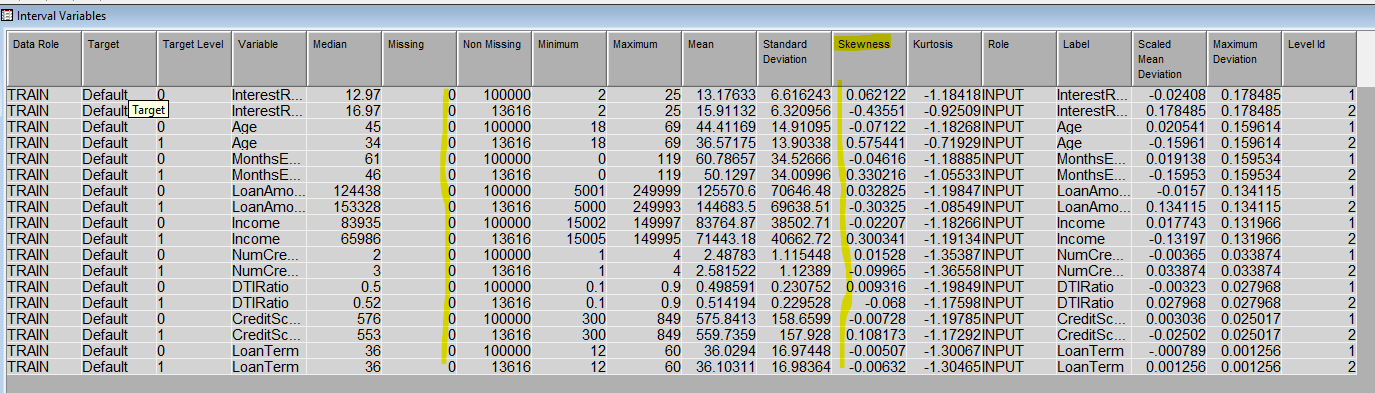
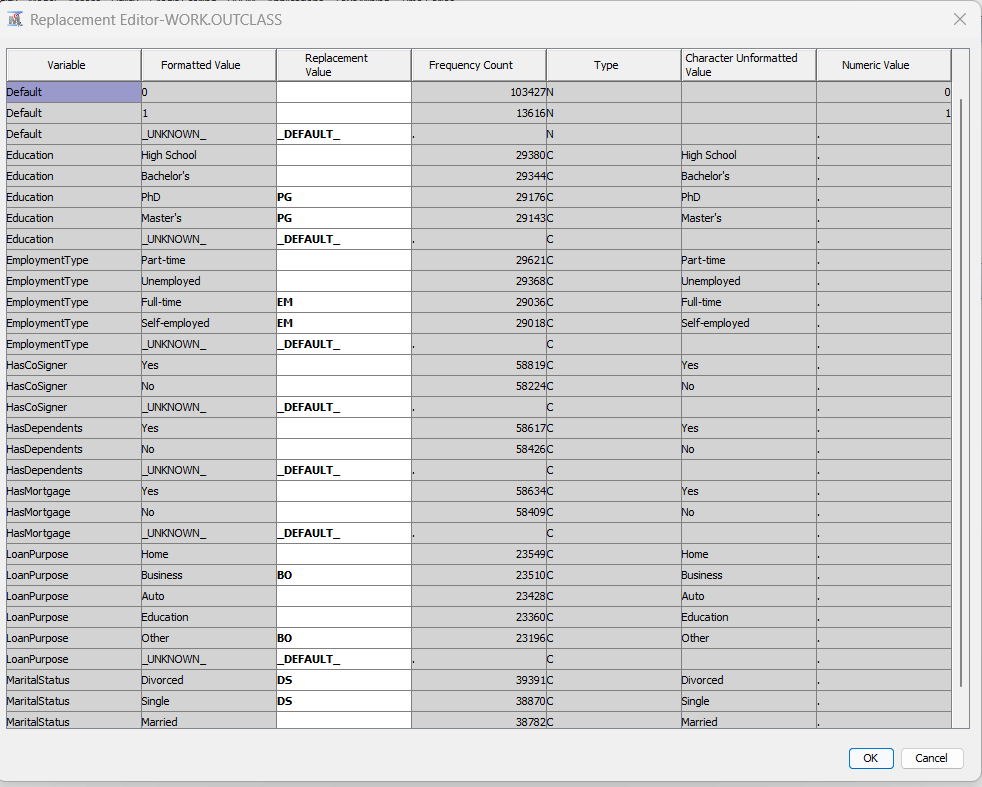
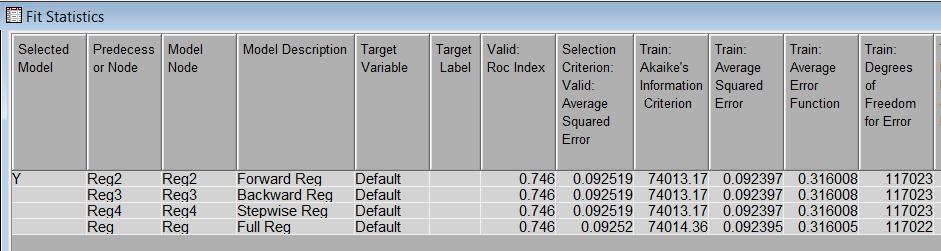
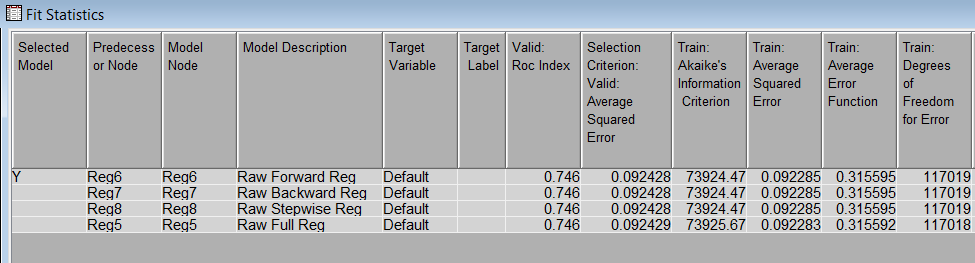
1. Identify critical predictors of loan defaults (e.g., income, debt-to-income ratio, credit score).
2. Build robust predictive models to classify loan applicants based on their likelihood of default.
3. Provide actionable insights to help banks improve loan approval processes and minimize default rates.

The decision tree approach provides an interpretable way to identify decision rules for defaults, regression models can help in finding relationships between predictors and loan outcomes, and neural networks capture complex, non-linear patterns within the data. By combining these techniques, we aim to deliver a thorough analysis that can be interpreted for business relevance.

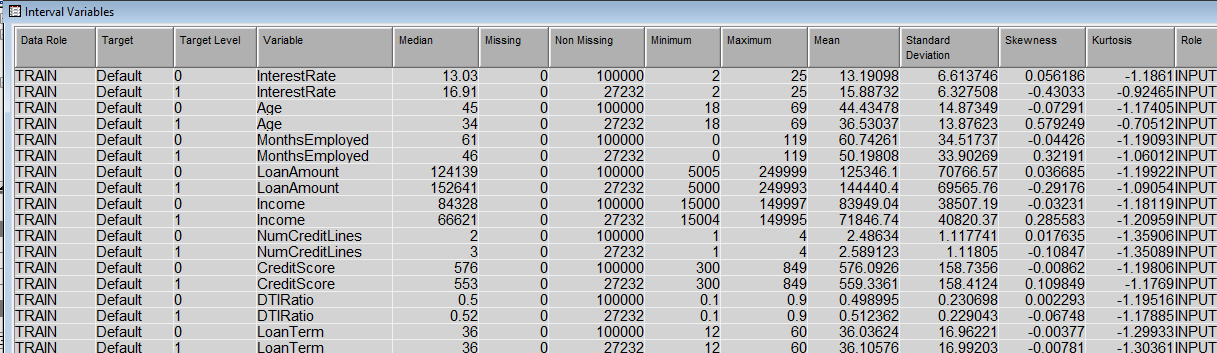
Ultimately, this project will help the bank refine its risk assessment framework, leading to informed lending practices and a sustainable future.

**Diagram**



* The data set was imported.
* The data was split into training and validation sets in a 50-50 ratio so that a reliable model performance could be ascertained.
* 3 decision tree models were created.
* An impute node was not added since there are no missing values.
* A replacement node was not added, since the data is not skewed.  
  
* Re-coding dummies/variables: Groups under a few variables were combined to reduce the curse of dimensionality & computational complexity:  
    
  1. Under education, ‘PHD’ & ‘Masters’ were combined under a new group called PG, since these groups are quite similar & candidates from this group will end up earning money post graduation that will help in repaying the loan.  
    
  2. Under employment type, ‘Full time’ & ‘Self-employed’ were combined under the new group EM, since candidates from both groups are earning a considerably stable income, in comparison with the other two groups.  
    
  3. Under loan purpose, ‘Business’ & ‘Other’ were combined under the new group BO, since we are assuming that other types of loans taken are most likely to be business related.  
    
  4. Under Marital Status, ‘Single’ & ‘Divorced’ were combined under the new group DS, since the candidates of these two groups are all individuals.   
  
* 4 different regression models were created (Full, Forward, Backward & Stepwise).
* Before proceeding, a comparison was made to check if re-coding made the model better or worse. Hence, the 4 regression models were compared, one set of which was connected to the recode dummies node & the other set was connected directly to the data partition.  
    
  1. Below is the model comparison results for the regressions **after re-coding dummies.**  
  Recoded Regression - ROC gives values up to 3 decimals only and the ROC of all the regressions are the same, hence ASE was used to compare models.  
    
    
    
  2. Below is the model comparison results for the regression **without re-coding dummies**.  
   This ASE is better without re-coded variables. Therefore, the re-coded step has not been considered any further.  
  
* Hence, the 7 Neural Networks models with 2 to 8 hidden units were connected to our best regression model
* We have also connected a Neural Network directly to the data partition & one more to the full regression.
* Finally, all the models have been connected to the model comparison node for assessment.

**Stat Explore**



The table summarizes interval variables for a **loan default analysis**, separating the data by **Target = 0 (Non-default)** and **Target = 1 (Default).**

**General understanding of the dataset:**

**1.** **Interest Rate:  
 Default:** Mean = 15.88, Standard Deviation = 6.32  
 **Non-Default:** Mean = 13.19, Standard Deviation = 6.61  
 **Insight:** Defaults are associated with higher average interest rates. Borrowers facing higher interest payments may be more likely to default.

**2.** **Age:  
 Default**: Mean = 44.5, Standard Deviation = 14.87  
 **Non-Default**: Mean = 46.3, Standard Deviation = 13.87  
 **Insight**: Borrowers in default tend to be very slightly younger on average compared to non-defaulting borrowers. This suggests that age may be a factor in loan defaults, but this difference is too little to conclude.

**3.** **Months Employed:  
 Default**: Mean = 53.30, Standard Deviation = 33.90  
 **Non-Default**: Mean = 60.74, Standard Deviation = 34.51  
 **Insight**: Non-defaulting borrowers have, on average, more months of employment. This indicates that job stability (longer employment history) is associated with lower default rates.

**4.** **Loan Amount:  
 Default**: Mean = 144,440, Standard Deviation = 71,462  
 **Non-Default**: Mean = 125,340, Standard Deviation = 70,766  
 **Insight**: Defaults are associated with higher loan amounts. Borrowers with larger loans may face greater financial burdens, increasing their likelihood of defaulting.

**5.** **Income:  
 Default**: Mean = 71,846, Standard Deviation = 40,402  
 **Non-Default**: Mean = 83,994, Standard Deviation = 38,507  
 **Insight**: Borrowers who default tend to have lower average incomes compared to non-defaulters. Income is a significant predictor of the ability to repay loans.

**6.** **Number of Credit Lines:  
 Default**: Mean = 2.5, Standard Deviation = 1.11  
 **Non-Default**: Mean = 2.48, Standard Deviation = 1.11  
 **Insight**: There is minimal difference between the groups. This suggests that the number of credit lines might not be a strong predictor of loan default.

**7.** **Credit Score:  
 Default**: Mean = 533.70, Standard Deviation = 158.47  
 **Non-Default**: Mean = 576.09, Standard Deviation = 158.73  
 **Insight**: Defaulting borrowers have slightly lower average credit scores. This suggests that credit score could be a factor in loan default prediction, but this is too small a margin to conclude.

**8.** **DTI Ratio (Debt-to-Income Ratio):  
 Default**: Mean = 0.55, Standard Deviation = 0.23  
 **Non-Default**: Mean = 0.49, Standard Deviation = 0.23  
 **Insight**: Higher debt-to-income ratios are observed in the default group, indicating that borrowers who allocate more of their income toward debt are at greater risk of defaulting.

**9.** **Loan Term:  
 Both Groups**: Mean = 36, Standard Deviation ≈ 17  
 **Insight**: Loan term is identical for both groups, suggesting it may not play a major role in predicting loan defaults.

### **Insights based on the dataset:**

One of the **major talking points** with this dataset is that the observations & any inferences from this study can be **only** applicable to loans that are less than 60 months (5 years), since the maximum loan terms for this entire data set is only 60 months.

1. **Interest Rate**: Defaults are linked to higher interest rates.
2. **Income**: Borrowers with lower incomes are more likely to default.
3. **Credit Score**: Lower credit scores are a significant predictor of default.
4. **Debt-to-Income Ratio**: Higher DTIRatio increases the risk of default.
5. **Loan Amount**: Larger loans are associated with a higher likelihood of default.
6. **Employment Duration**: Longer employment duration is associated with fewer defaults.
7. **Age**: Younger borrowers are slightly more likely to default.

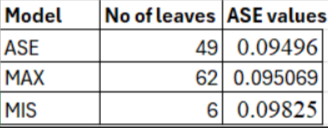
**Decision Tree**

**Decision Trees for Predicting Loan Defaults: Validation ASE Comparison**

**Overview**

The various decision tree models were validated strictly on the basis of their performance as assessed by validation ASE values because their performance on unseen data is paramount. By restricting ourselves to validation ASE alone, we avoid biases inherent in training results and focus our attention on those models that generalize better on new observations.

The validation ASE was used to assess model accuracy and generalizability.



### **Observations:**

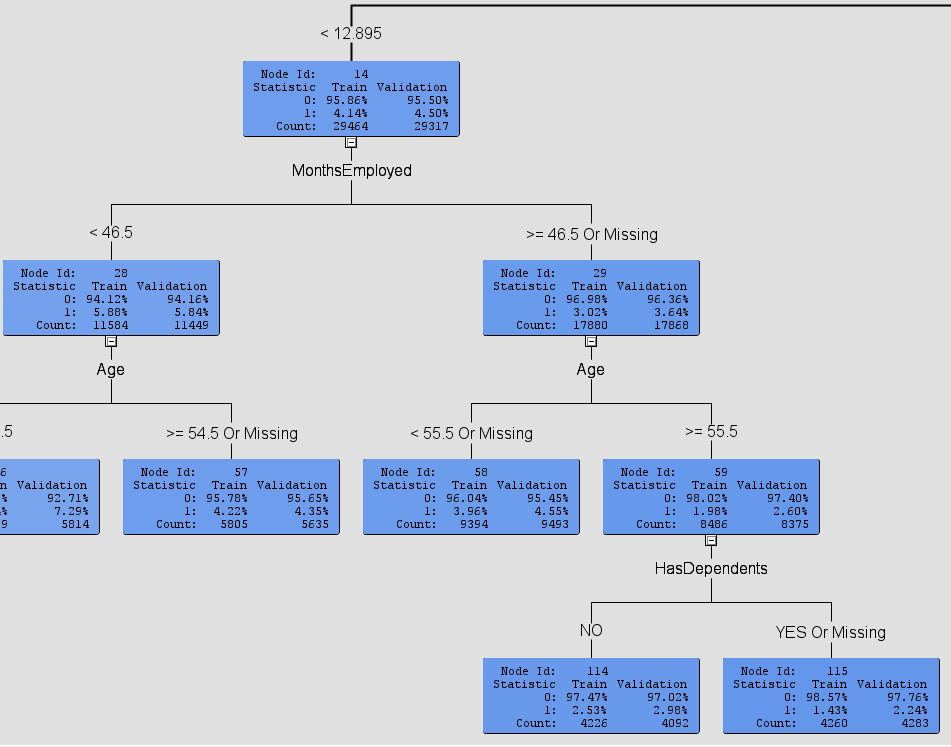
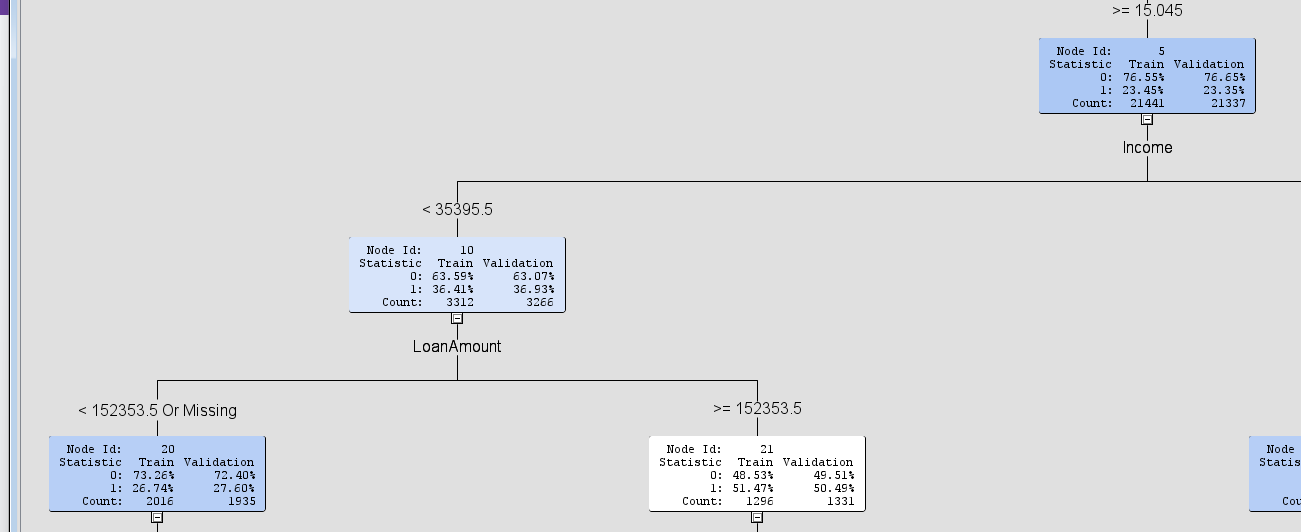
1. **ASE Model:  
   No of Leaves**: 49  
   **ASE Value**: 0.09496  
   This model has a moderate level of complexity (49 leaves) and achieves the lowest ASE value, indicating the best performance among the three models.  
   **Interpretation:** This model strikes a balance between complexity and accuracy.
2. **MAX Model:  
   No of Leaves**: 62  
   **ASE Value**: 0.095069  
   This model has the highest number of leaves (62), indicating the most complex tree. However, its ASE value (0.095069) is slightly higher than the ASE model.  
   **Interpretation**: Despite its complexity, this model does not improve performance and might be overfitting.
3. **MIS Model:  
   No of Leaves**: 6  
   **ASE Value**: 0.09825  
   This model has the fewest leaves (6), representing the simplest model. However, it has the highest ASE value, indicating poorer performance.  
   **Interpretation:** This model under fits the data due to its simplicity, failing to capture key patterns.

### **Recommendations**

1. **Preferred Model**: The **ASE model** is the most optimal choice because it achieves the lowest ASE value with a reasonable level of complexity (49 leaves).
2. **Avoid Overfitting**: The **MAX model** adds unnecessary complexity without significantly improving accuracy.
3. **Avoid Underfitting**: The **MIS model** is too simple and does not perform well, as indicated by the higher ASE value.

**Analysis of leaves in ASE tree:**

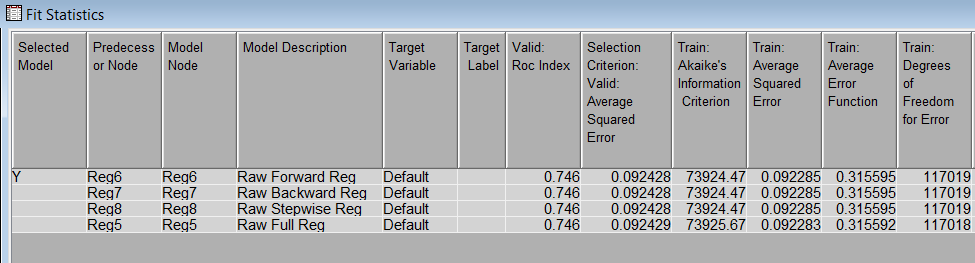
**Providing loans** -

1. **The least chances of default** - customers with age is equal to 39 or greater than 39.5 or missing whose interest rate is less than 12.895 and have been employed for 46.5 months or more, having a dependent would make it even less likely for them to default these applicants have a 97.76% chance of not defaulting.  
   
2. **The highest chance of default:** Applicants whose age is less than 39.5 whose interest rate is greater than 15.04% with a income of less than 35,395 and who are appling for the loan that are greater than the amount of 152,353 these applicants have a 50.49% chance of defaulting. These candidates are highly likely to default. 

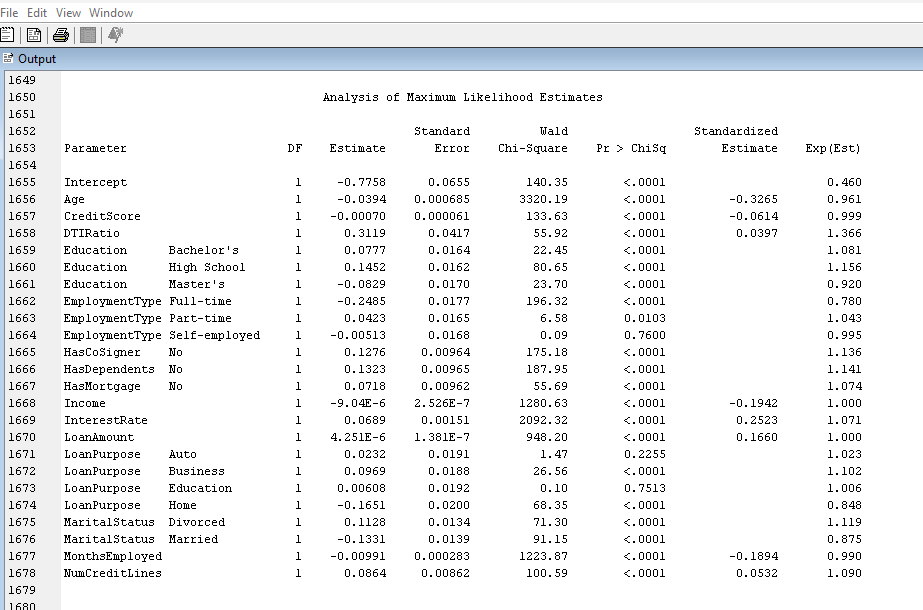
**On the whole, the decision tree models identified critical variables and decision rules to predict loan defaults, thus enabling banks to make informed decisions and manage risk effectively.**

**Regression Models**

The model comparison results of the 4 regression models. The Raw - Forward, Backward & Stepwise all have the same ASE values and hence are the best models, the Raw Forward regression will be used for the rest of the assessment.

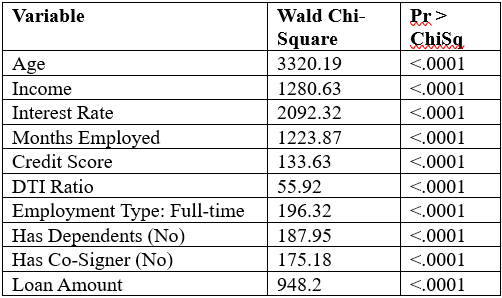


**Results of the Raw Forward Regression**



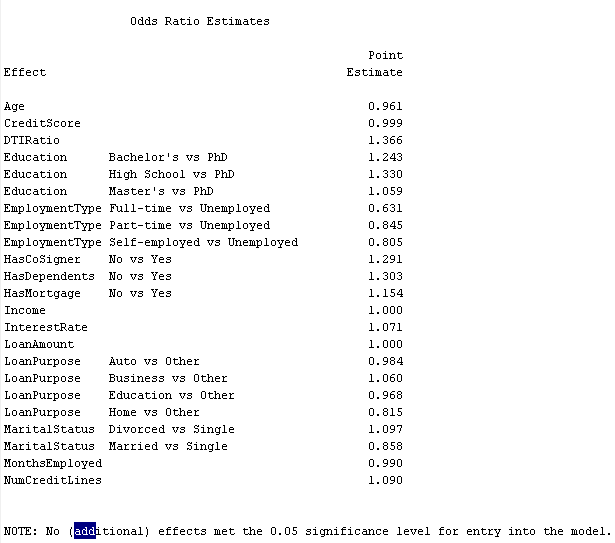
**10 most important variables & predictors of default:**

The Wald Chi-Square and p-values (Pr > ChiSq) are used to analyze the above table. Variables with large Wald Chi-Square values and statistically significant p-values (typically < 0.05) are the most important predictors. Only 10 most important variables are being considered in this case to give us a better understanding of the data.

****

The most influential variables include Age, Income, Interest Rate, Months Employed, and Credit Score, along with categorical variables like Employment Type and whether the applicant has a cosigner or dependents. These predictors are statistically significant and have practical relevance for identifying loan default risk.

**Odds Ratio Explanation:**



#### Odds Ratio (Exp(Estimate))

* Odds ratios represent the factor by which the odds of the event (loan default) change for a one-unit increase in a predictor variable or when comparing categories.
* Odds ratio > 1: Increases the odds of default.
* Odds ratio < 1: Decreases the odds of default.
* Odds ratio = 1: No effect on the odds.

**Variable-Wise Interpretation:**

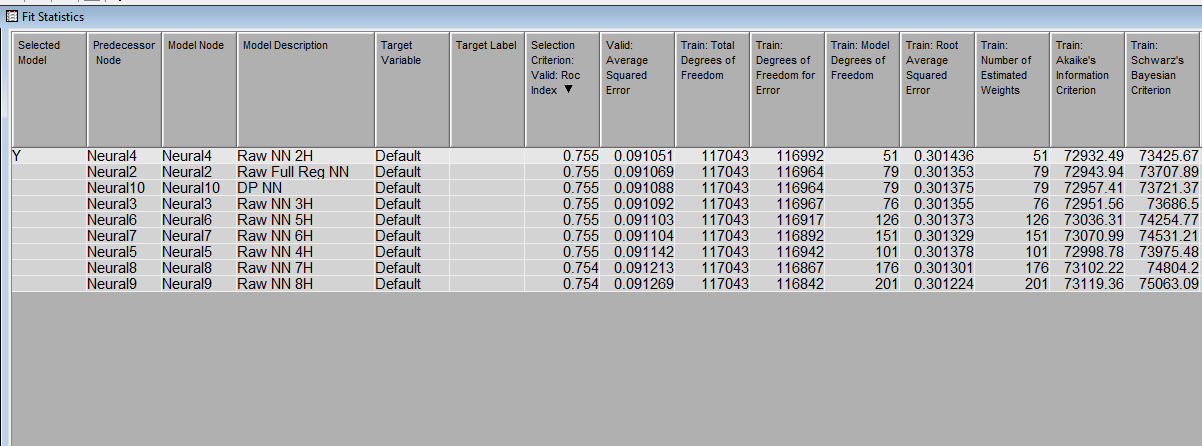
1. **Age:** Odds Ratio: 0.961  
    Interpretation: For each additional year of age, the odds of loan default decrease by 3.9% (1 - 0.961 = 0.039).
2. **Credit Score:** Odds Ratio: 0.999  
    Interpretation: A higher credit score slightly reduces the odds of default (negligible change, as the odds ratio is close to 1).
3. **DTI Ratio (Debt-to-Income Ratio):** Odds Ratio: 1.366  
    Interpretation: For every unit increase in DTI, the odds of loan default increase by 36.6%.
4. **Education:** Bachelor's vs PhD: Odds Ratio = 1.243  
    High School vs PhD: Odds Ratio = 1.330  
    Master's vs PhD: Odds Ratio = 1.059  
    Interpretation: Lower levels of education (e.g., high school) are associated with higher odds of default compared to a PhD.
5. **Employment Type:** Full-time vs Unemployed: Odds Ratio = 0.631  
    Part-time vs Unemployed: Odds Ratio = 0.845  
    Self-employed vs Unemployed: Odds Ratio = 0.805  
    Interpretation: Being employed (full-time, part-time, or self-employed) reduces the odds of default compared to being unemployed, with full-time employment offering the strongest reduction (36.9% decrease in odds).
6. **Has Co-Signer:** Odds Ratio: 1.291  
   Interpretation: Not having a cosigner increases the odds of loan default by 29.1%.
7. **Has Dependents:** Odds Ratio: 1.303  
   Interpretation: Not having dependents increases the odds of loan default by 30.3%.
8. **Has Mortgage:** Odds Ratio: 1.154  
    Interpretation: Not having a mortgage increases the odds of loan default by 15.4%.
9. **Income:** Odds Ratio: 1.000  
    Interpretation: The model shows no effect of income on the odds of default.
10. **Interest Rate:** Odds Ratio: 1.071  
    Interpretation: A higher interest rate increases the odds of default by 7.1%.
11. **Loan Amount:** Odds Ratio: 1.000  
    Interpretation: The effect of loan amount is negligible on the odds of default.
12. **Loan Purpose:** Auto vs Other: Odds Ratio = 0.984  
    Business vs Other: Odds Ratio = 1.060  
    Education vs Other: Odds Ratio = 0.968  
    Home vs Other: Odds Ratio = 0.815  
    Interpretation: Loan purpose has minimal effects on default, except for "Home," which decreases the odds by 18.5%.
13. **Marital Status:** Divorced vs Single: Odds Ratio = 1.097  
    Married vs Single: Odds Ratio = 0.858  
    Interpretation: Being divorced slightly increases the odds of default (by 9.7%), while being married reduces the odds (by 14.2%).
14. **Months Employed:** Odds Ratio: 0.990  
     Interpretation: Longer employment tenure slightly reduces the odds of default (by 1% per additional month).
15. **Num Credit Lines:** Odds Ratio: 1.090  
     Interpretation: More credit lines increase the odds of default by 9% per additional credit line.

**Recommendations for the Bank:**

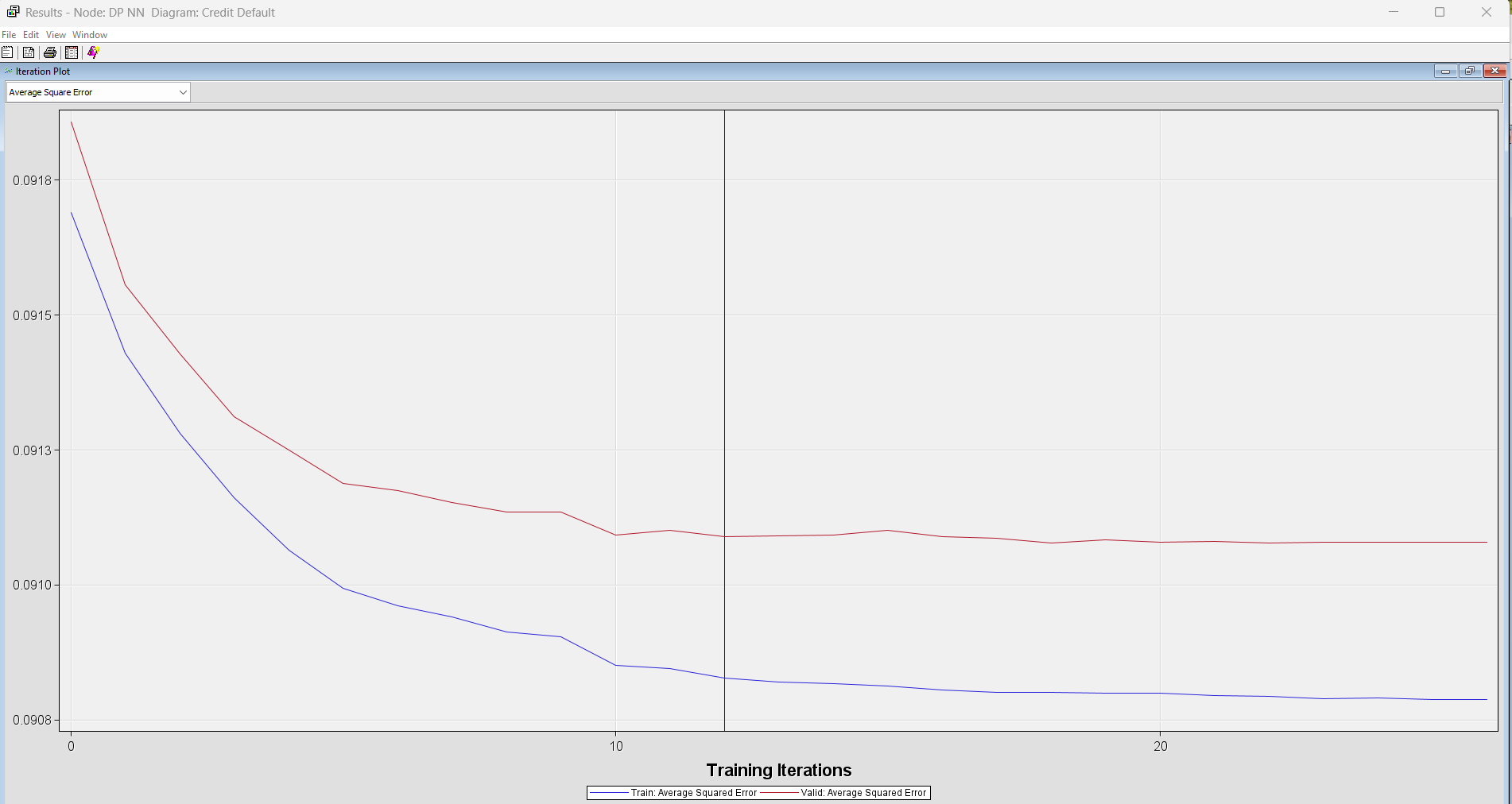
1. **Prioritize Risk Factors**: Focus on **Age, DTI ratio, Credit Score, Employment Type**, and **Interest Rate** when assessing loan applications.
2. **Adjust Lending Policies**:  
    - Implement stricter thresholds for high-risk factors like DTI ratio and interest rates.  
    - Require cosigners for applicants with high-risk profiles.
3. **Encourage Financial Stability**: Offer financial counselling to help borrowers manage debt and improve credit scores.
4. **Monitor Loan Amounts**: Limit loan amounts for high-risk applicants and gradually increase them after demonstrating repayment capability.
5. **Leverage Employment Stability**: Use employment history as a key risk assessment tool.

**Neural Networks**

Model comparison of all neural networks



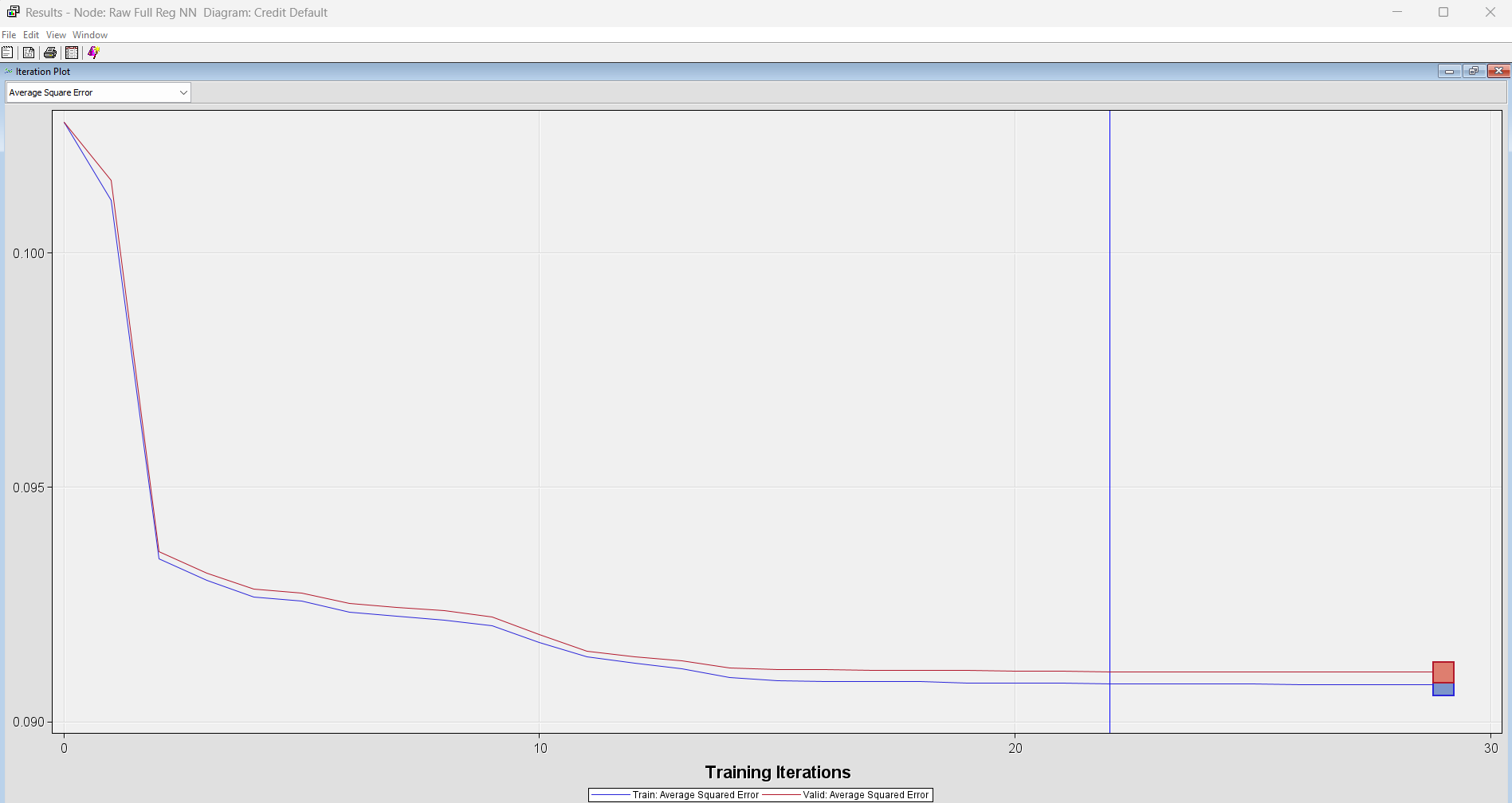
**DPNN**



**DP NN:**

* + Convergence: Yes (12 iterations)
  + Strengths: Fastest convergence (12 iterations), potentially indicating a simpler and more efficient model. This is important for rapid model development and deployment. In banking, where timely decisions are crucial, faster models can be advantageous.
  + Potential Weaknesses: We need to ensure it isn't oversimplified. Oversimplification can lead to a model that misses crucial risk factors, resulting in inaccurate predictions and potential financial losses.
  + Good or Bad for Predicting Default: Could be good for quick initial screening of loan applications, but might miss subtle risk factors that a more complex model would capture.
  + Banking Relevance: If this model proves accurate, it could be used for real-time credit scoring, enabling faster loan approvals and potentially reducing the risk of lending to high-risk borrowers.

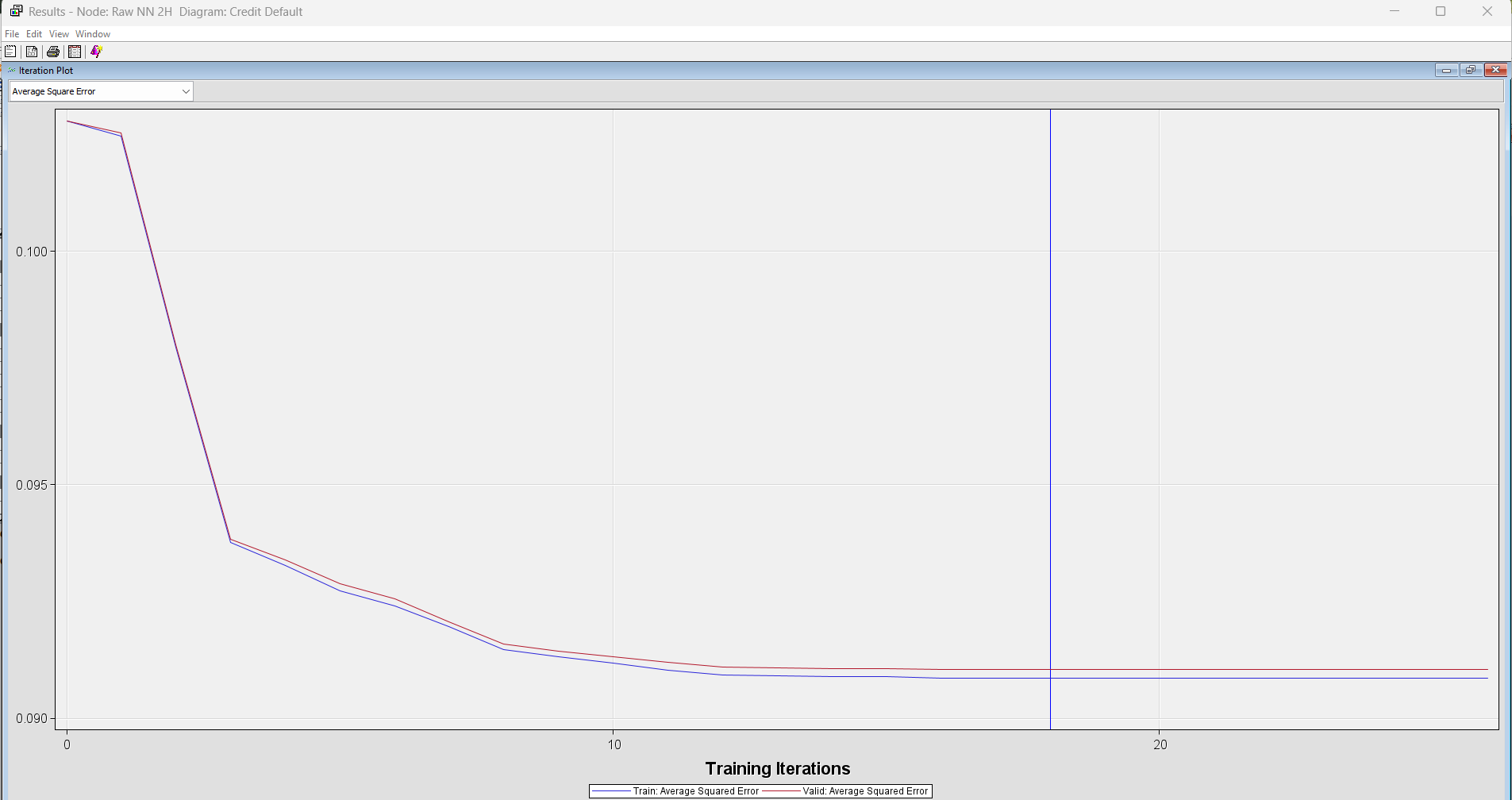
**RAW FULL REG NN**



Raw Full Reg NN:

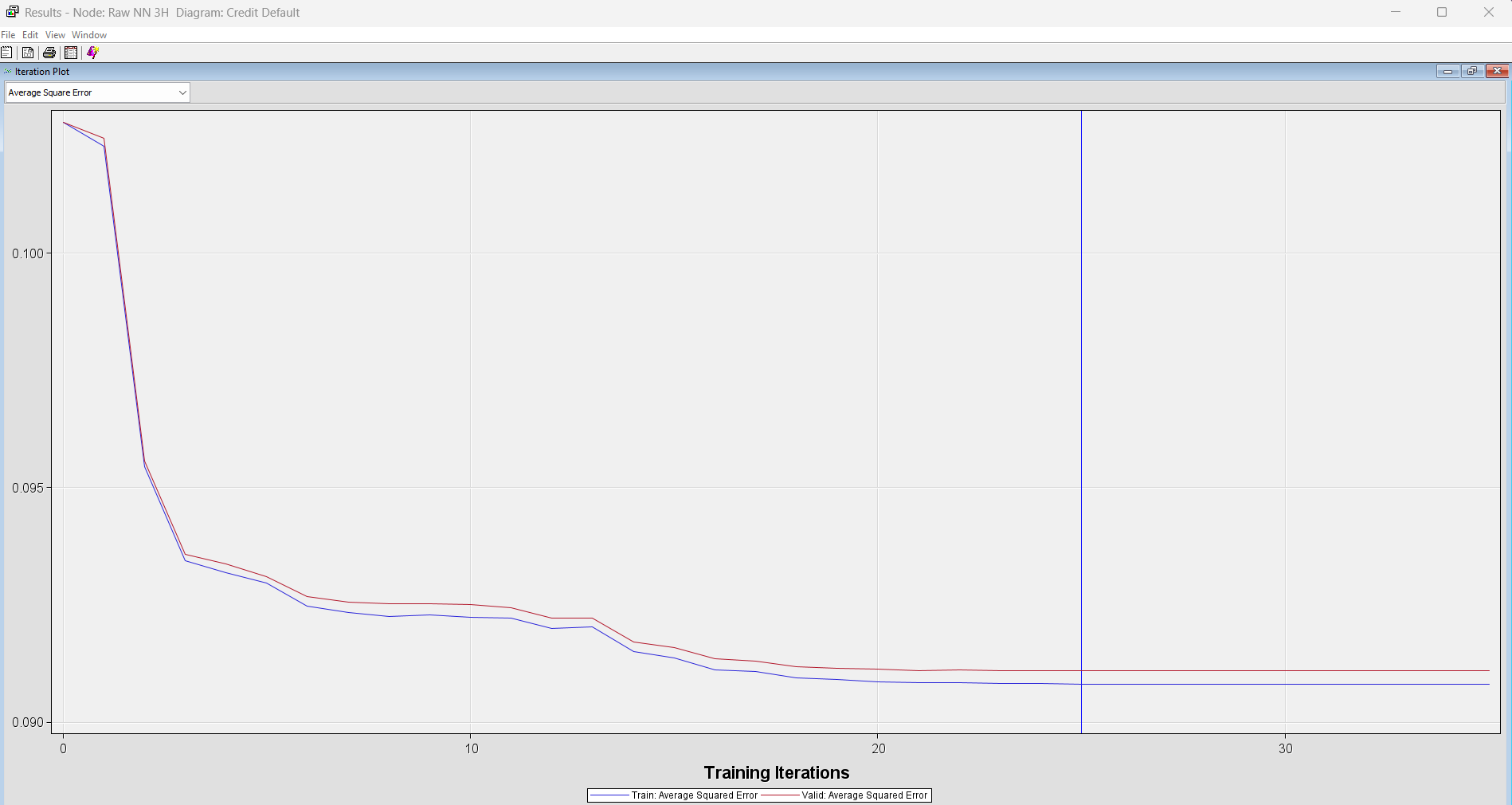
* + Convergence: Yes (22 iterations)
  + Strengths: Relatively fast convergence (22 iterations).
  + Potential Weaknesses: Similar to DP NN, we need more performance data to draw conclusions.
  + Good or Bad for Predicting Default: Similar to DP NN, it might be good for initial screening but could miss some nuances in risk assessment.

**RAW NN 2H**



1. Raw NN 2H:  
   * Convergence: Yes (23 iterations)
   * Potential Strengths: Converged reasonably quickly (23 iterations).
   * Potential Weaknesses: Higher validation ASE compared to DP NN and Raw Full Reg NN might indicate lower predictive accuracy.
   * Good or Bad for Predicting Default: Might be better at capturing non-linear relationships in the data, potentially leading to more accurate default predictions than simpler models, but this needs to be confirmed with further analysis.

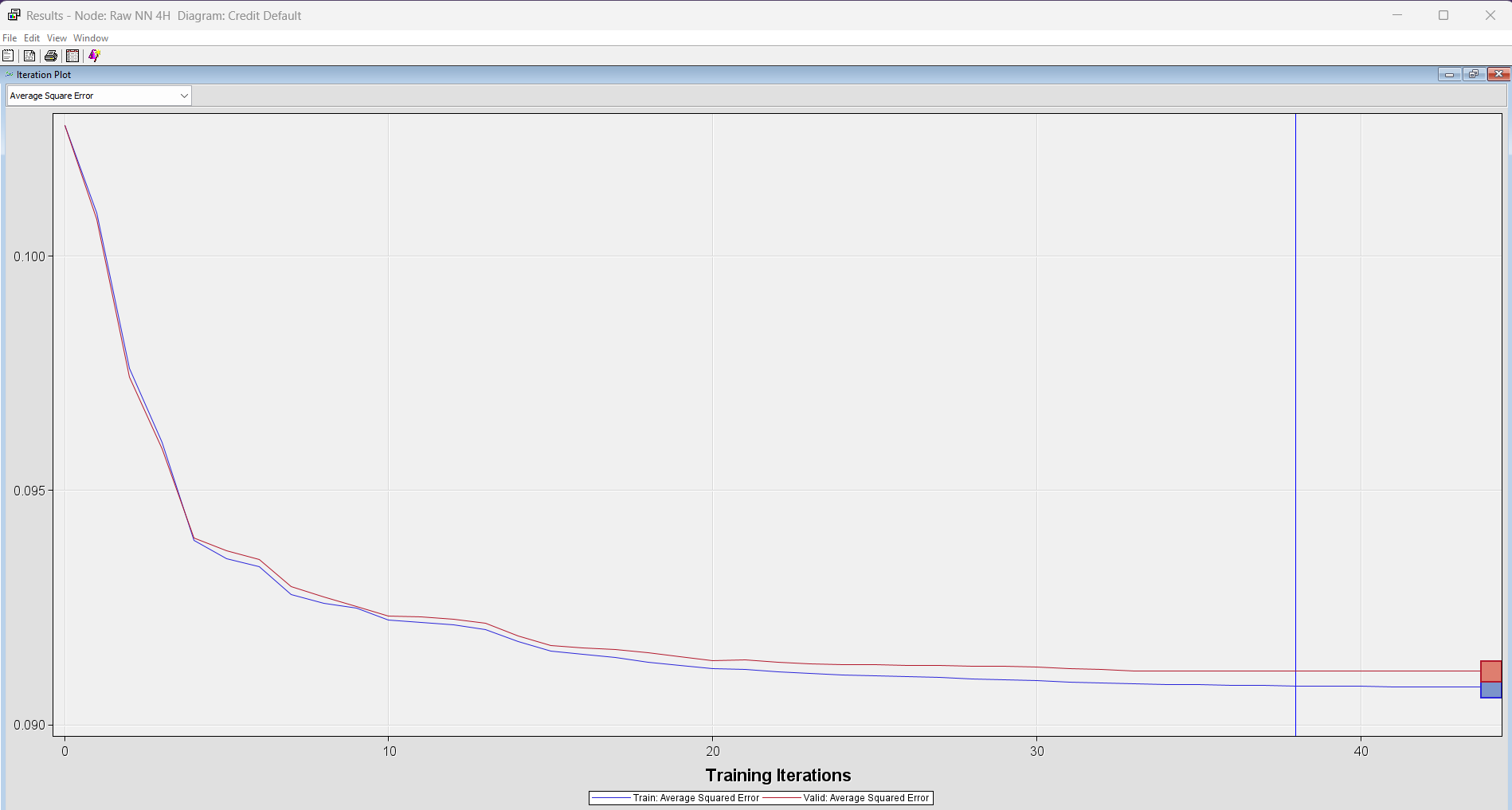
RAW NN 3H



**Raw NN 3H:**

* + Convergence: Yes (31 iterations)
  + Potential Strengths: Converged (31 iterations).
  + Potential Weaknesses: Moderate convergence speed (31 iterations) and validation ASE.
  + Good or Bad for Predicting Default: Similar to Raw NN 2H, it could potentially capture more complex relationships but requires careful evaluation of its accuracy.

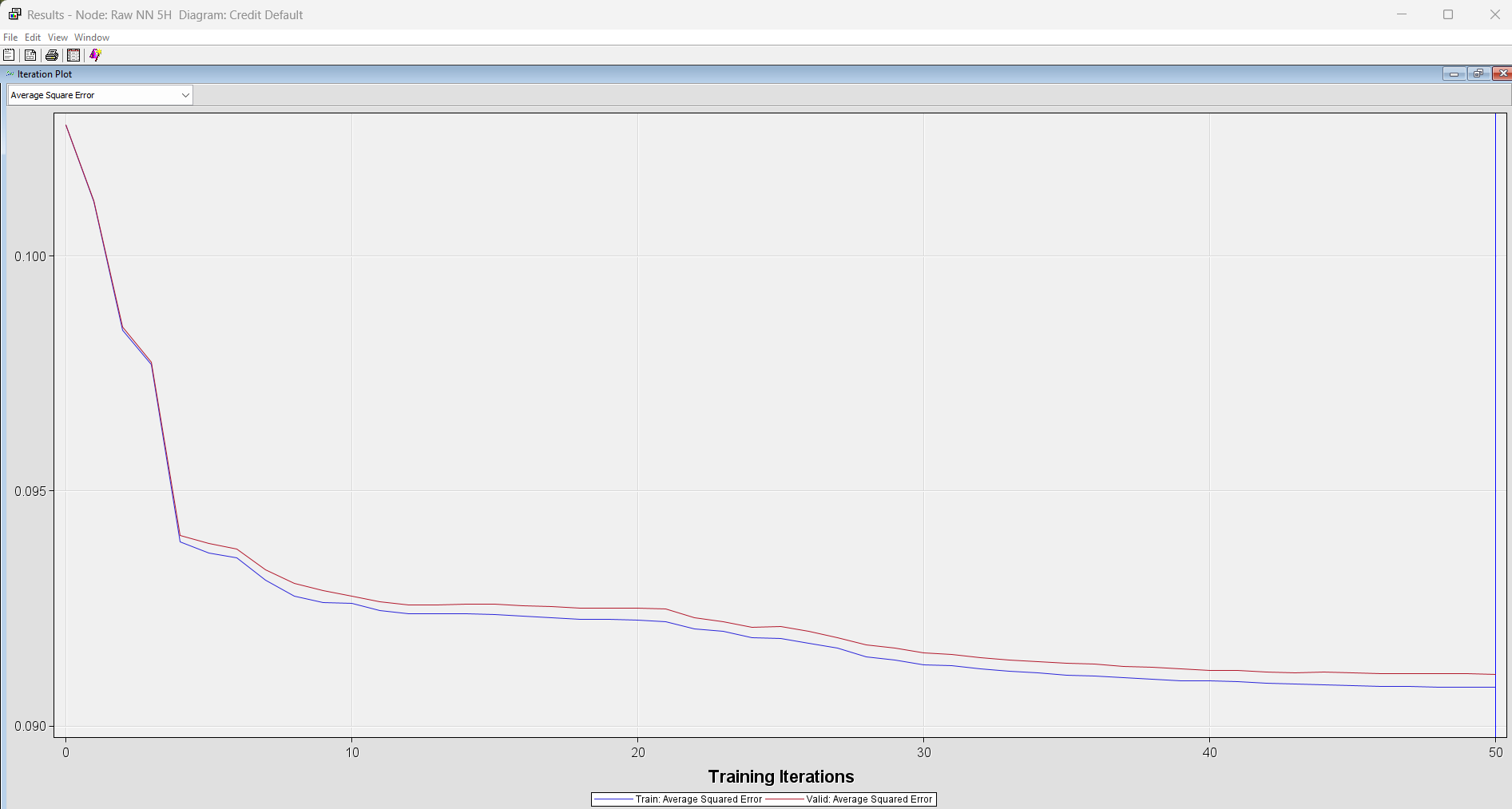
**RAW NN 4H**



**Raw NN 4H**

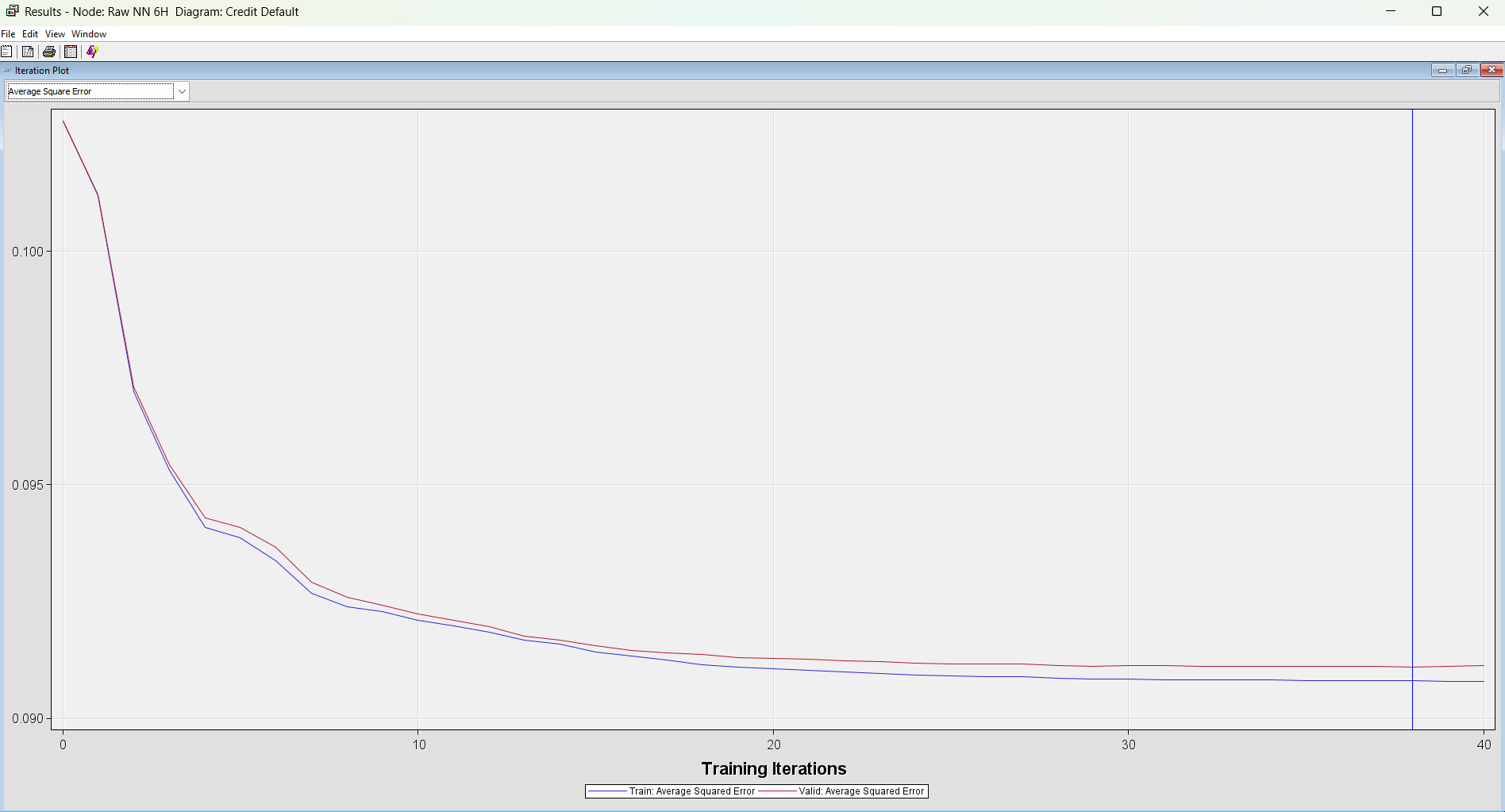
* Convergence: Yes (38 iterations)
* Potential Strengths: Converged successfully, it was able to find a stable solution during training.
* Possible Weaknesses: Moderate convergence speed (38 iterations) may suggest that it is more complex than, for example, the DP NN or Raw Full Reg NN. We have yet to see how its accuracy compares to justify extra training time.
* Good or Bad to Predict Default: Unlike any other simpler models, this model has the capability of capturing complicated relationships in the data and may lead to good predictions; though this needs confirmation with analysis from the performance metrics-Roc Index, ASE, among others.

**RAW NN 5H**



1. Raw NN 5H:  
   * Convergence: No (50 iterations)
   * Concerns: The failure to converge within 50 iterations raises a red flag. This could indicate instability in the model or the need for more data or adjustments to the learning process. In a banking context, model stability is critical for reliable risk assessment.
   * Good or Bad for Predicting Default: Potentially bad due to the lack of convergence, which suggests it might be unreliable or difficult to train effectively.

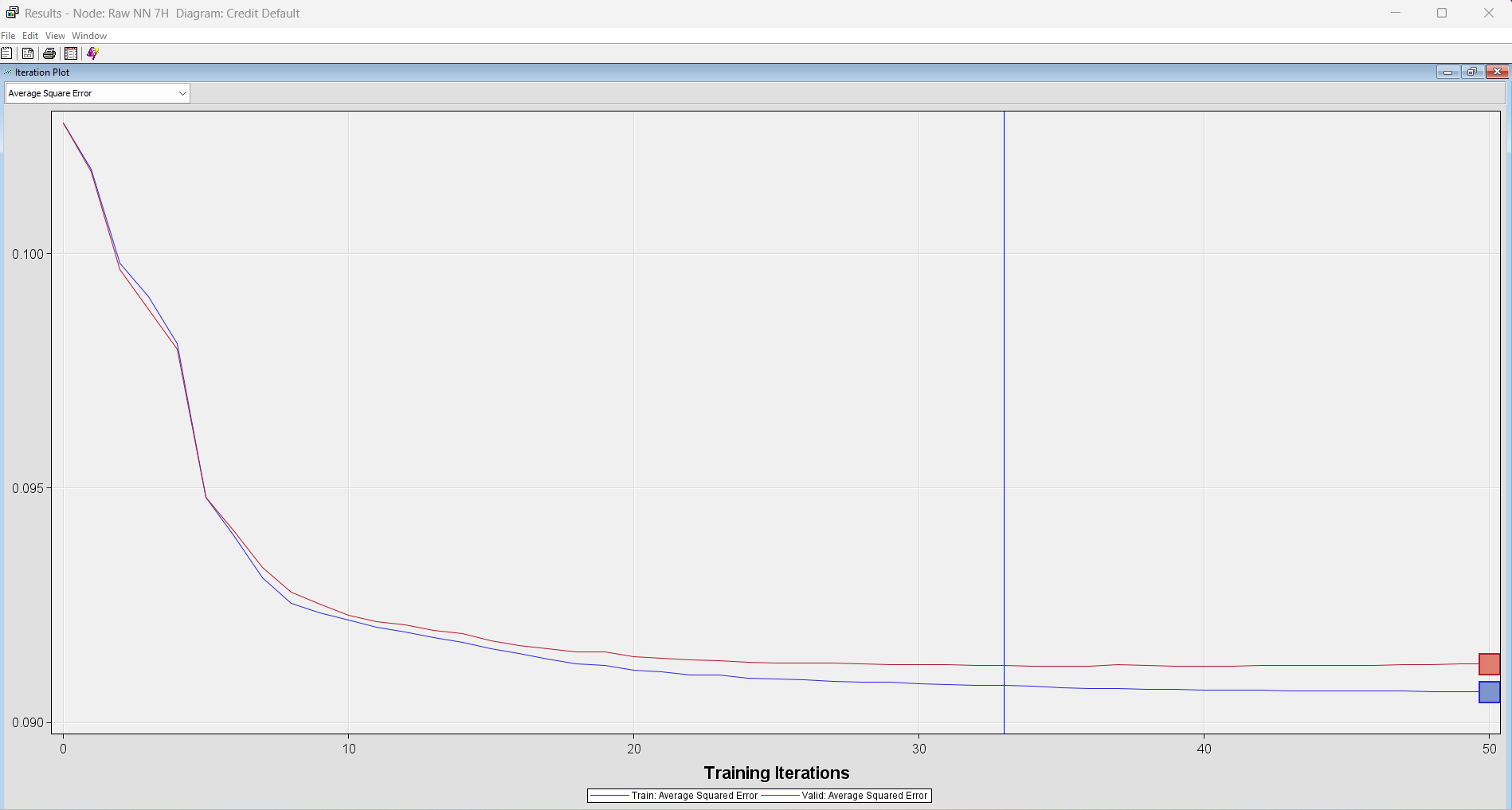
**RAW NN 6H**



Raw NN 6H

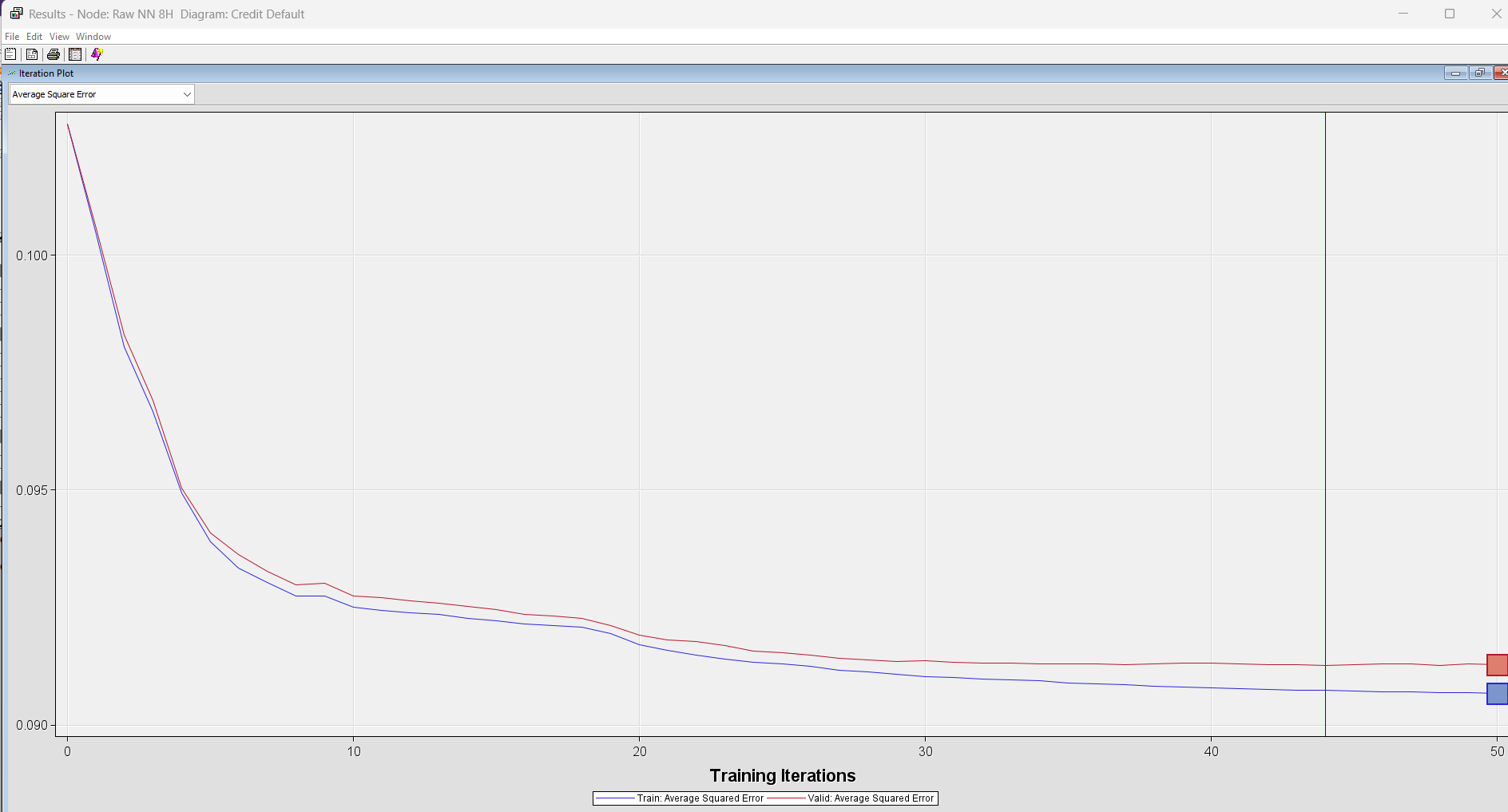
* Convergence: Yes (38 iterations)
* Potential Strengths: Successfully converged, suggesting it has learned well from training data.
* Possible Weaknesses: Convergence speed was moderate-38 iterations-and hence needs to be balanced against the likely gain in accuracy.
* Good or Bad for Predicting Default: Perhaps, good to pick out a complex relationship existing between different attributes that would lead to increased defaults, but this may further have to be verified for unseen cases and against less-complex models.

**Raw NN 7H:**



* Convergence: Yes (58 iterations)
* Potential Strengths: Converged eventually (58 iterations).
* Potential Weaknesses: Slowest convergence (58 iterations), indicating a complex model that requires significant training time.
* Good or Bad for Predicting Default: Could be good for capturing very complex relationships and potentially identifying high-risk loans that simpler models might miss, but the increased complexity needs to be justified by significantly improved accuracy.

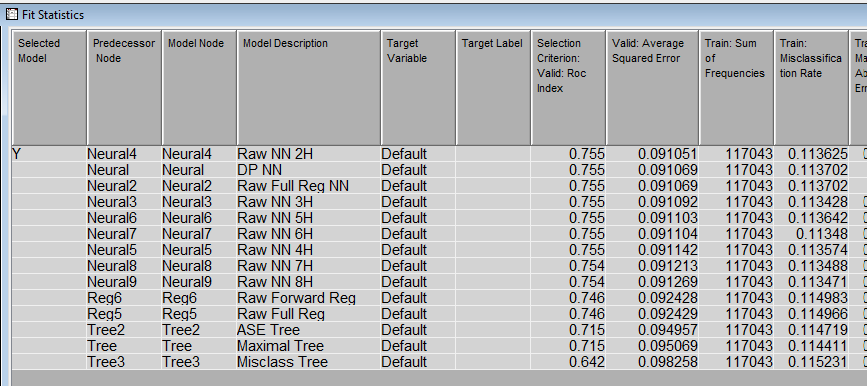
**RAW NN 8H**



Raw NN 8H:

* Convergence: Yes (58 iterations)
* Potential Strengths: Converged (58 iterations).
* Potential Weaknesses: Similar to Raw NN 7H, it has slow convergence (58 iterations).
* Good or Bad for Predicting Default: Similar to Raw NN 7H, it has the potential to capture complex relationships but needs to demonstrate a clear advantage over simpler models

**Model Comparison**



**Model Performance Insights**

Based on the SAS output, some key observations are as follows:

**ROC Index:**

Most of the neural network models, starting from Raw NN 2H to Raw NN 8H, have a ROC Index of 0.755, indicating good discriminatory power.

Raw Forward Reg and Raw Full Reg models have relatively lower ROC Indices, 0.746, which may imply a reduction in the ability to discriminate between defaulters and non-defaulters.

ASE Tree and Maximal Tree present reasonable results as well-0.715. Misclass Tree has the poorest ROC Index of 0.642.

**Average Squared Error (ASE):**

The neural network models give an ASE that varies between 0.091051 and 0.091269. The Raw NN 2H presents the smallest ASE among these different versions. Raw Forward Reg and Raw Full Reg models perform poorer with their ASE around 0.0924, suggesting potentially low prediction accuracy.

Tree-based models have higher values of ASE, compared to most neural networks, which is indicative of overall lower accuracy.

**Misclassifications:**

The misclassification rates are lower for most neural networks, at around 0.113, indicating better accuracy in classifying loans either as default or non-default.

Raw Forward Reg and Raw Full Reg models follow closely behind with misclassification rates of about 0.114.

Tree-based models, particularly the Misclass Tree, show higher misclassification rates above 0.115.

**Model Complexity:**

Neural networks, especially those with more hidden layers, such as Raw NN 7H and Raw NN 8H, are likely to be more complex and capture intricate relationships in the data.

Raw Forward Reg and Raw Full Reg models are likely to be less complex compared to the multi-layer neural networks.

Tree-based models are varied in complexity, with the Misclass Tree being the most simplistic and thus potentially less accurate of all.

Preliminary Recommendations

**Top Performers:**

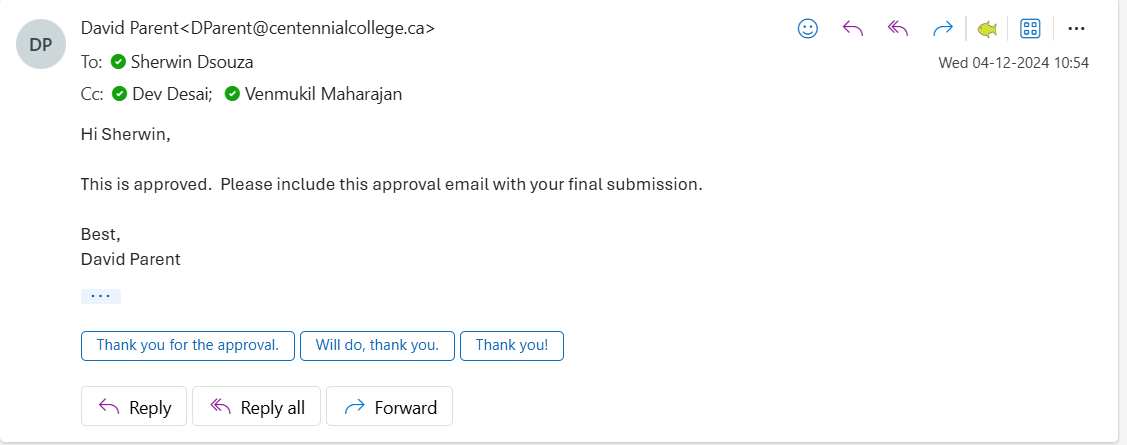
Based on the ROC Index and ASE, neural network models, in particular, **Raw NN 2H,** look quite promising.

Further Evaluation: We should check the complexity and interpretability of these models. There might be a preference for a simpler model if they exhibit similar performance with much ease of understanding and, perhaps, implementation.

Business Context: The chosen final model should be aligned with our risk appetite, regulatory requirements, and business objectives. Other key factors include explainability, implementation costs, and the potential for biases.

**Final Conclusion:**

While the Neural Network with 2 hidden units provides the best performance, the forward regression model will be considered when:  
- Interpretability is a priority (e.g - explaining results to stakeholders).  
- Resource constraints exist (e.g - limited computation power).

****