**SAS ENTERPRISE MINOR FINAL PROJECT**

**DATA SET: BANK–CHURN**

Submitted to

Prof. David Parent

Submitted by

Chandra Poudyal

December, 2021

**Table of Content**

List of figures 2

1. Introduction : Business Objective 3
2. Data Analysis 3
3. Project Bank Churn – Analysis 4
4. Data Modelling 6
   1. Decision Tree 6
   2. Regression 8
   3. Neural Network 11
5. Model Comparison 14
6. Conclusion 16
7. Bibliography 18

**List of figures**

Fig 1: Variables used in this analysis

Fig. 2: StatExplore details

Fig. 3: StatExplore and data partition model picture

Fig. 4: Maximal Interactive Tree

Fig. 5: Misclassification Tree

Fig. 6: Average Square Tree

Fig. 7: StatExplore to know skewness

Fig. 8: Transformation of the skewed variable

Fig. 9: Full Regression

Fig. 10: Backward Regression

Fig. 11: Stepwise Regression

Fig. 12: Forward Regression

Fig. 13: Neural Network having 3 hidden nodes

Fig. 14: Neural Network having 4 hidden nodes

Fig. 15: Neural network having 2 hidden nodes

Fig. 16: Neural network having 6 hidden nodes

Fig. 17: Neural network having 9 hidden nodes

Fig. 18: Final result of comparison

Fig. 19: ROC of all models

Fig. 20: Node Model Comparison result

Fig. 21: All process to get final model of the comparison

Fig. 22: Final output of Average Squared Error showing best model

1. **Introduction:**

The SAS Enterprises minor is used to perform Churn modeling for customer attributes to analyze the customer attrition, customer turnover, or customer defection, in the loss of clients or customers in the bank.

The most significant concert is the retention of existing client and acquisition of new clients to determine the business growth. The new companies are trying to acquire new customers and at the same time mature companies are focusing on the retention of the existing ones to providing opportunities of cross selling.

This analysis is focused on the bank customers’ behaviors that are more likely to leave the bank (i.e. close their bank account). To find out the most striking behaviors of customers through data analysis, and use some of the predictive analytics techniques to determine the customers who are most likely to churn. The SAS Enterprise Minor is used to build different analytical models for the bank customer’s behavior to monitor their customers risk in a real-time setting.

1. **Data Analysis**

Data is downloaded from <https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling>. The data record found in the form of CSV file and it have to change into excel file format and then can loaded in the SAS Enterprise Minor. While analysing this data, it is found that the data have the following characteristics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Full label** | **Model Role** | **Measurement Level** | **Description** |
| RowNumber | Row Number | Rejected |  | Row number |
| CustomerId | Customer ID | ID | Nominal | Identification number of customer |
| Surname | Surname | Rejected |  | Surname of the customer |
| CreditScore | Credit Score | Input | Interval | The range of credit score is from 350 to 850 |
| Geography | Geography | Input | Nominal | Customers are from three countries: France, Germany and Spain |
| Gender | Gender | Input | Binary | Customer Gender |
| Age | Age of Customer | Input | Interval | The customer’s age ranges from 18 to 92 |
| Tenure | Tenure | Input | Interval | Customers stayed in the bank in years |
| Balance | Balance | Input | Interval | The available amount for withdrawal |
| NumberOfProduct | Number of Product issued | Input | Interval | How many products are used by customers |
| HasCrCard | Has Cr. Card | Input | Binary | If customer have credit card indicated by 1 |
| IsActiveMember | Is Active Member | Input | Binary | Customer are active indicated by 1 |
| EstimatedSalary | Estimated Salary | Input | Interval | Customer’s self reported annual salary |
| Exited | Exited | Target | Binary | whether the customer has churned (closed the bank account) 1, and 0 indicates still in use |

1. **Project Bank Churn- Analysis**

For the analysis a new project called Bank Churn in the SAS Enterprise Minor Workstation created for the bank business analysis. As the file is in csv format it will saved into format, then loaded into the system.

The Churn Modelling contains 14 variables and 10000 line items. For the business purposes it is decided to reject 2 variables (Surname and Row number) and use one target variable (Exited), listed in the data dictionary above. The variables taken in account for the input are sufficient to get necessary information for the present analysis.

The variables are changes into the nominal, interval and binary as per required and the exited variable remains binary and will play a decisive role in this analysis.

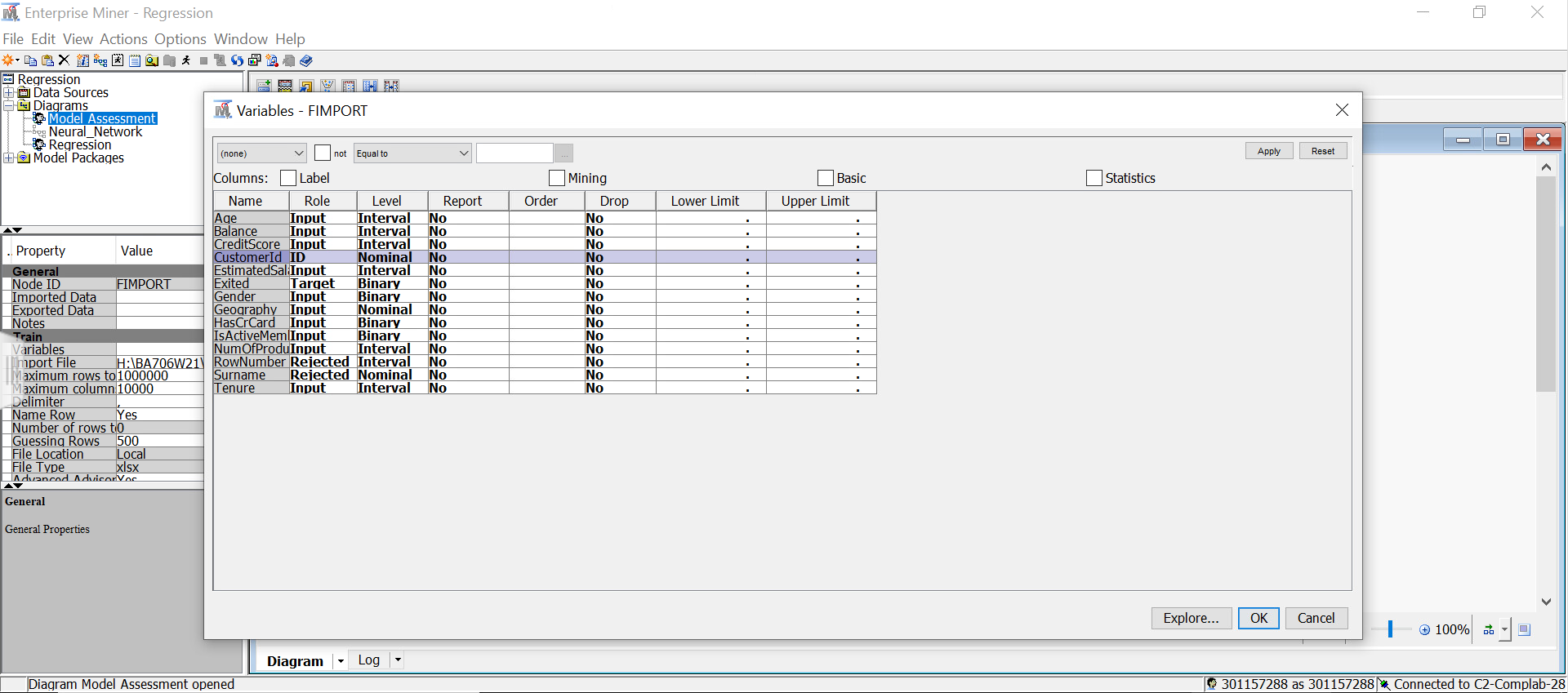


Fig. 1: Variables used in this analysis

While running the StatExplore node on the data set and found that there were no any missing values in all variables, so there will be no loss of data. As per the validation purpose the whole data are partitioned into two equal half and run the system.

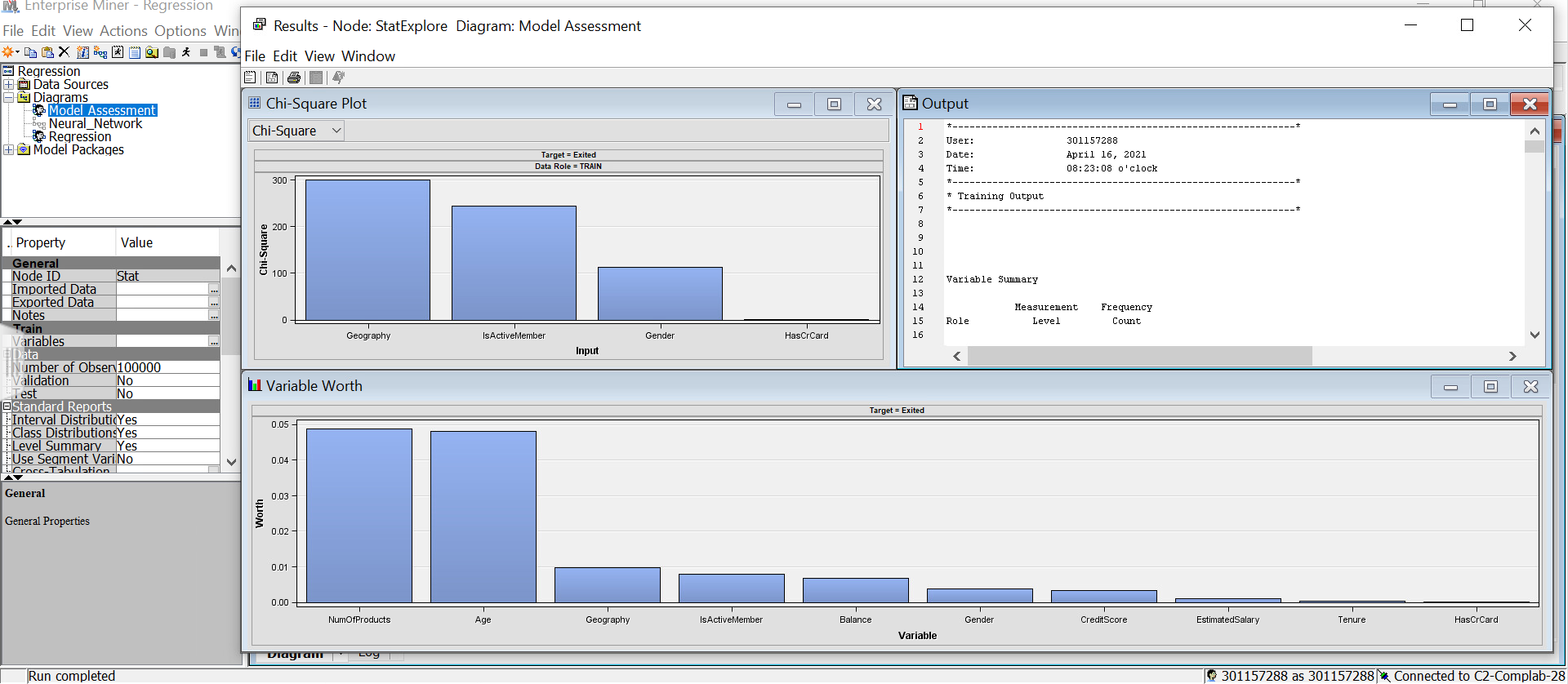


Fig. 2: StatExplore details

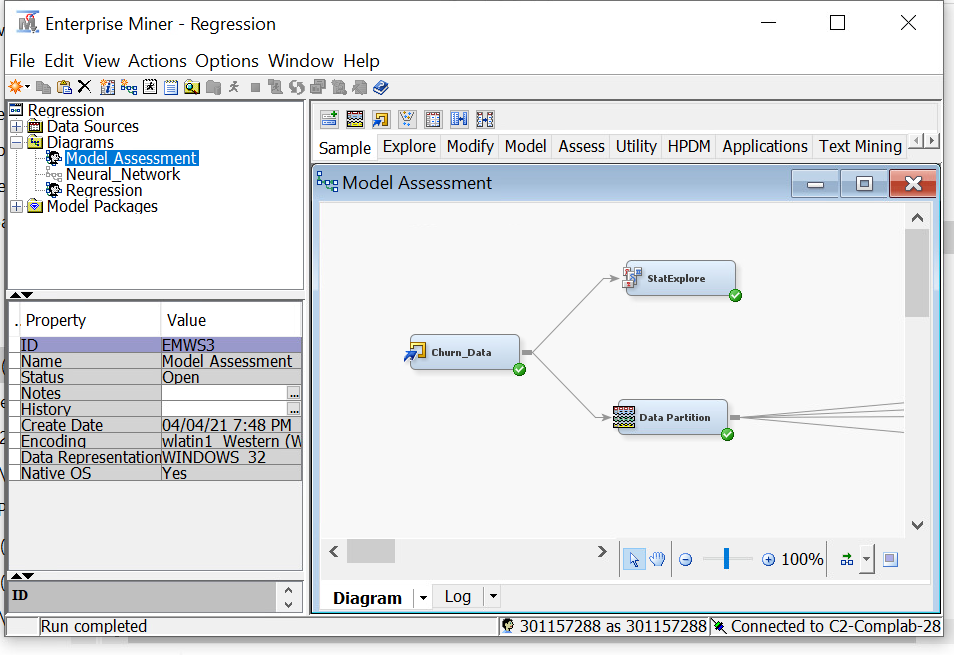


Fig. 3: StatExplore and data partition model picture

1. **Data Modelling**

The standardization of the data is necessary if the range or scale of one variable has larger difference with other variables. This difference can be compensated by the inter-correlation among the variables. Cluster analysis can be used to summarize data rather than finding natural or real cluster. Clustering is one of the unsupervised classification and grouping of data attempted based on the similarities in the input variables.

The most common mean of clustering is through the k-means algorithm and is primary clustering tool in SAS Enterprise Miner. The data variables are clustered in order to get at least 3 meaningful clusters to make easy cluster interpretation. The more stable clusters will produce due to low correlation in input variables. The lower value of skewness and kurtosis indicates the chances of creating small outlier cluster.

There are 10,000 bank customers data used for this analysis. The response or target variable is EXITED indicates whether a customer is still in bank or left. The variables with little or no relation with the target were excluded from the analysis.

* 1. **Decision Tree**

Decision tree analysis is one of the best methods of analysis showing a graphical representation. This method answers various questions in a series taking in account of previous questions answer after each affirmative or negative answer until making the final choice. The graphical representation is looks like a tree structure where problems in decision making represents a flow chart having each alternative choices in each branch. The decision tree provides clarity about the possible alternatives and also visible with the consequences of any decisions that will be made. During the decision tree analysis the branch splitting stops when log worth values is less than 0.7.

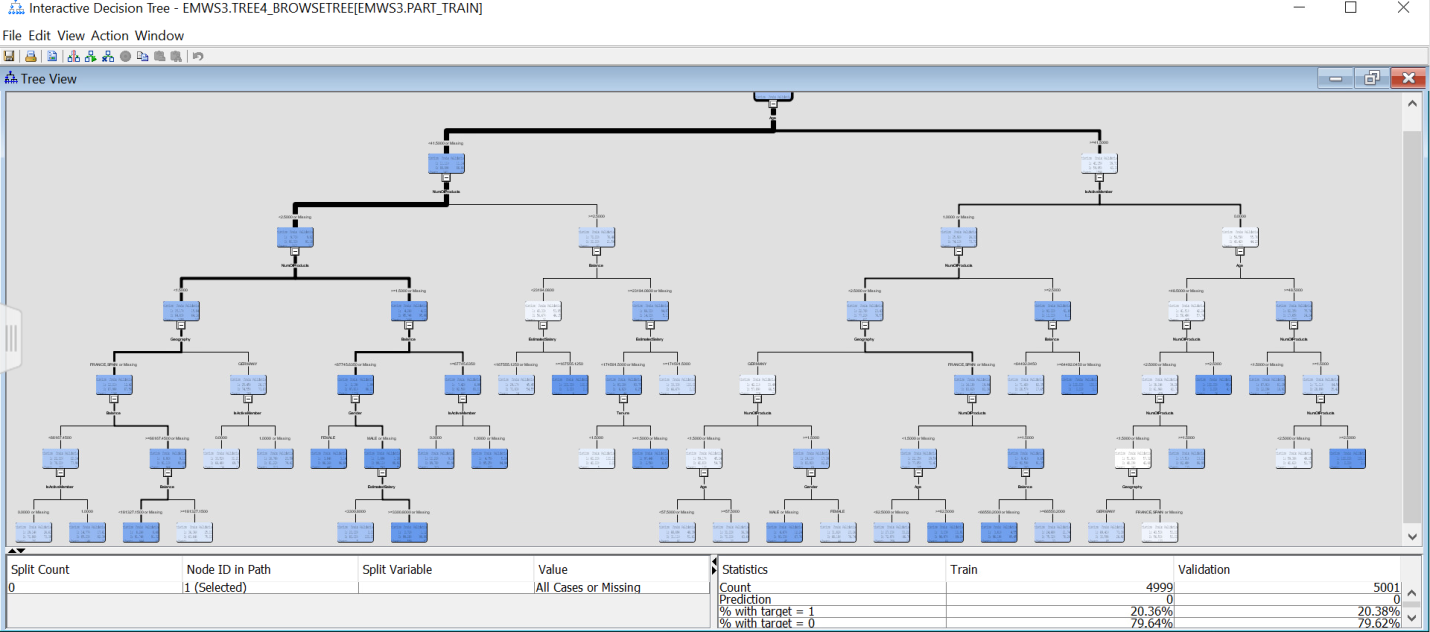


Fig. 4: Maximal Interactive Tree

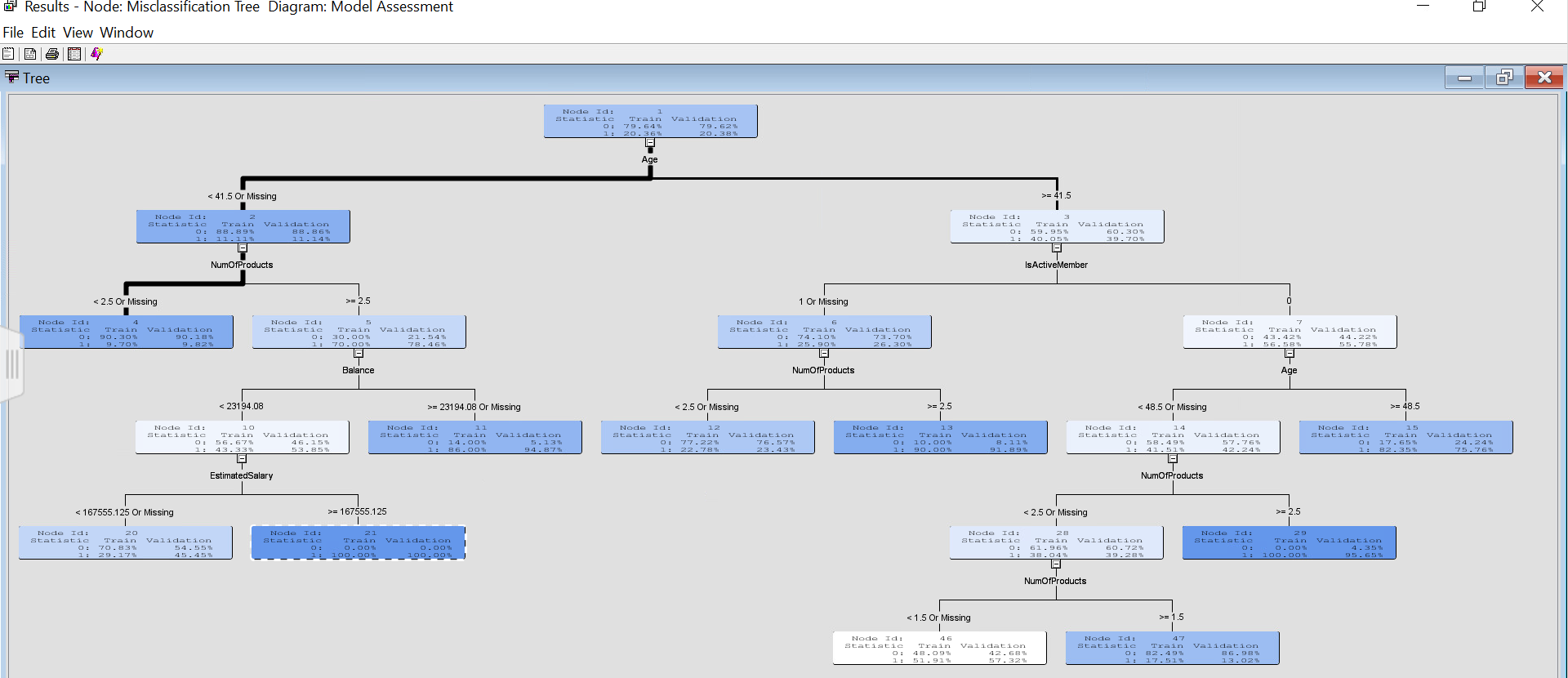


Fig. 5: Misclassification Tree

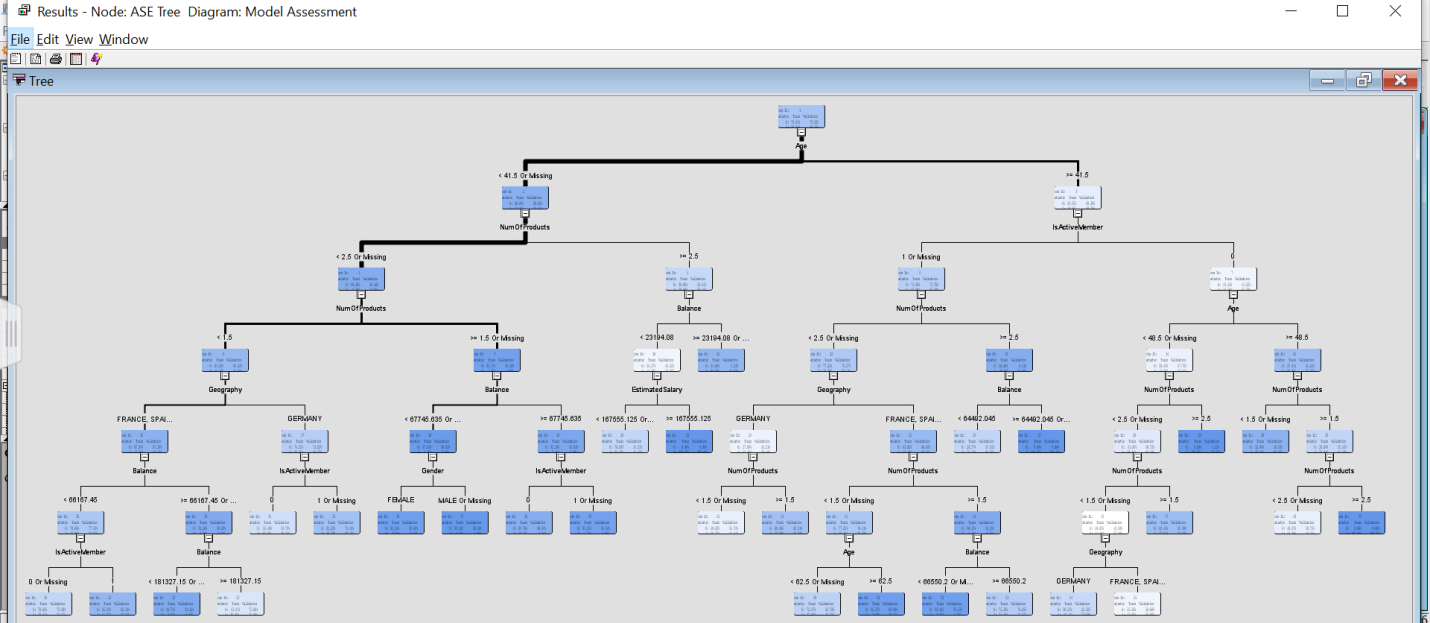


Fig. 6: Average Square Tree

* 1. **Regression**

The regression analysis is one of the statistical method to estimate the relationships among dependent variables and one or more independent variables. The strength of the relationship between variables can be accessed through the regression analysis. There are several regression models in use and common is linear regression. Nonlinear regression is commonly used to deal with the complicated data sets and showing nonlinear relationships between the dependent and independent variables. In the SAS Enterprise Minor there are full regression, stepwise regression, forward regression and backward regression processes and it gives best result for the large size data variables.

The Cap and Floor node introduced before running the regression to replace extreme values as it performs standard deviation and the mean which will change the value higher than mean into the mean value. After running Cap and Floor, it is good to check whether there is still skewed values remaining in the system or not. Afterward another node of StatExplore is connected with the Cap and Floor node. The result gives the high skewed value and then that value will be minimize using the log of the variables then our data distribution will be normal using the node Transform Variables. As these data does not have too many variables so there will be no recode dummies node added in this analysis.

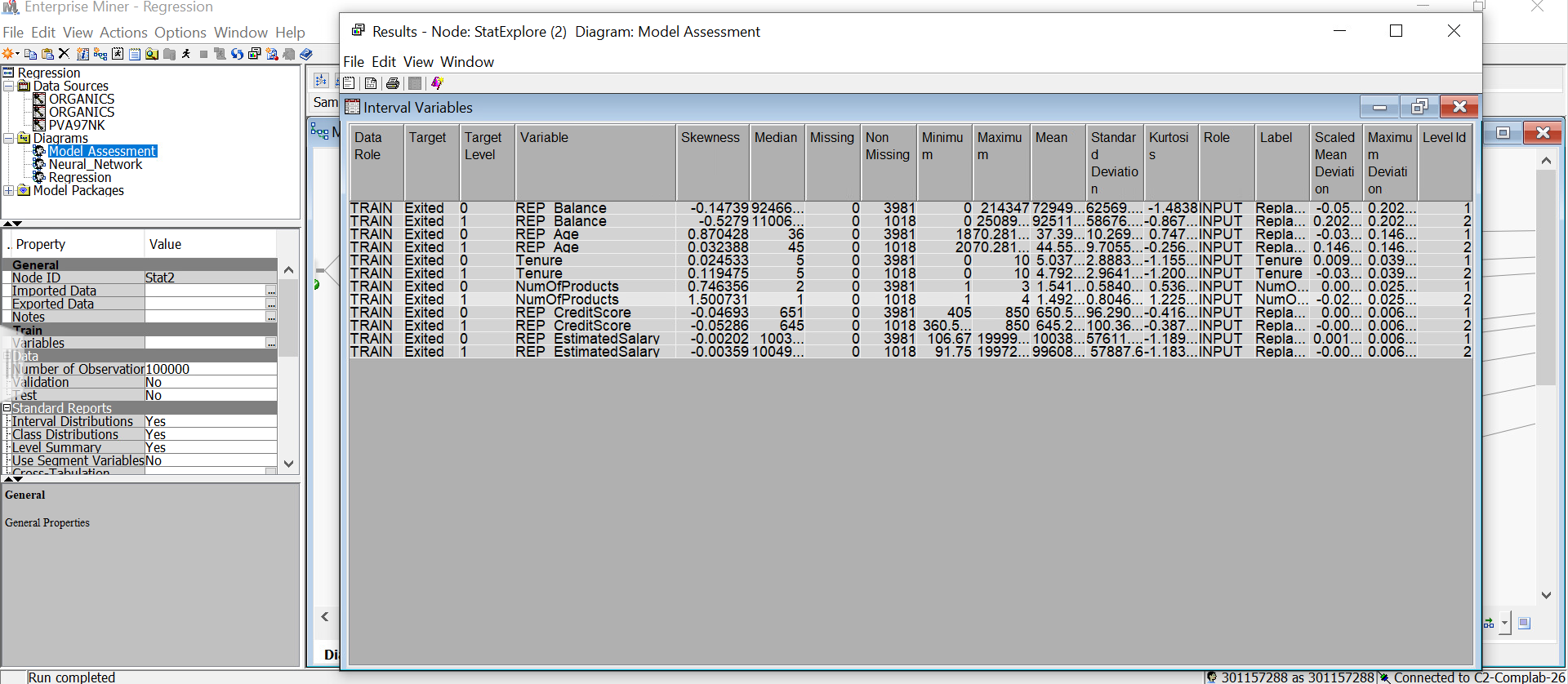


Fig. 7: StatExplore to know skewness

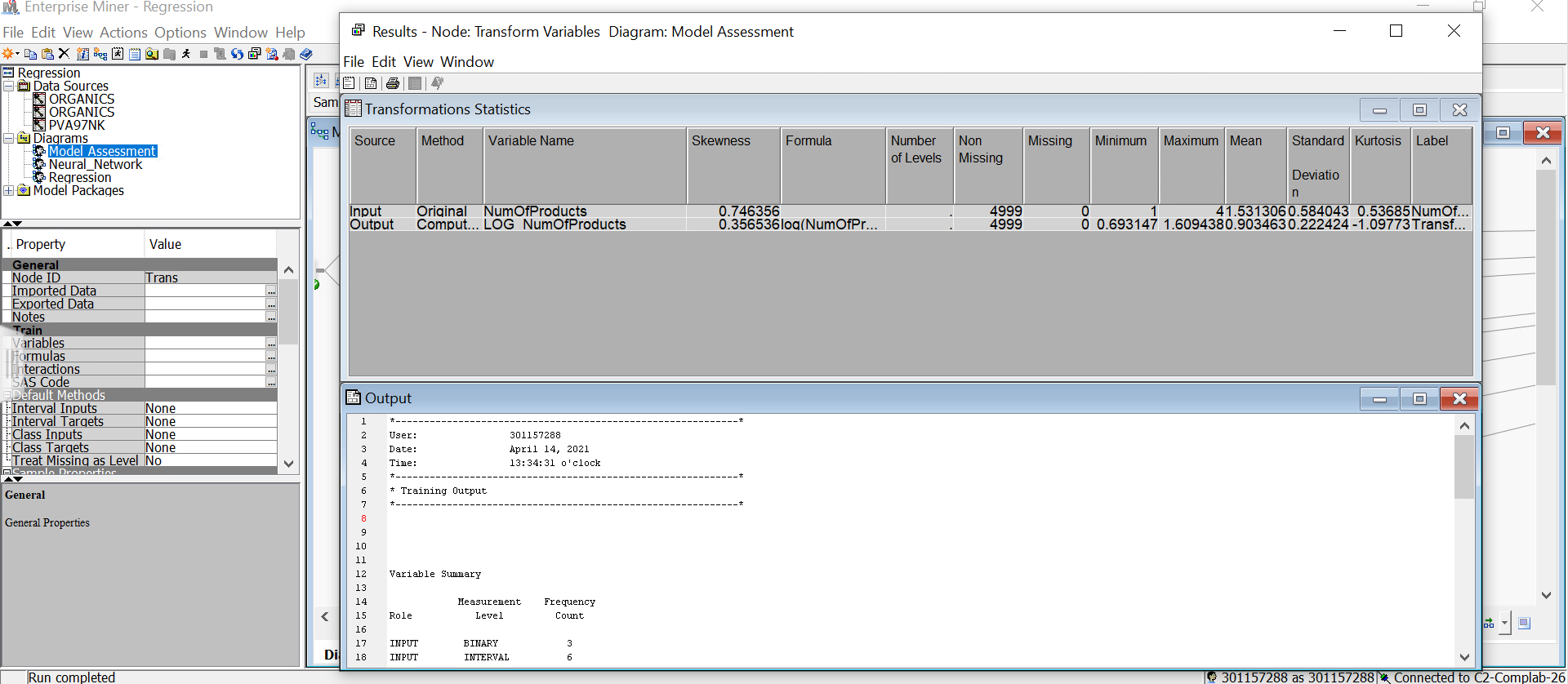


Fig. 8: Transformation of the skewed variable

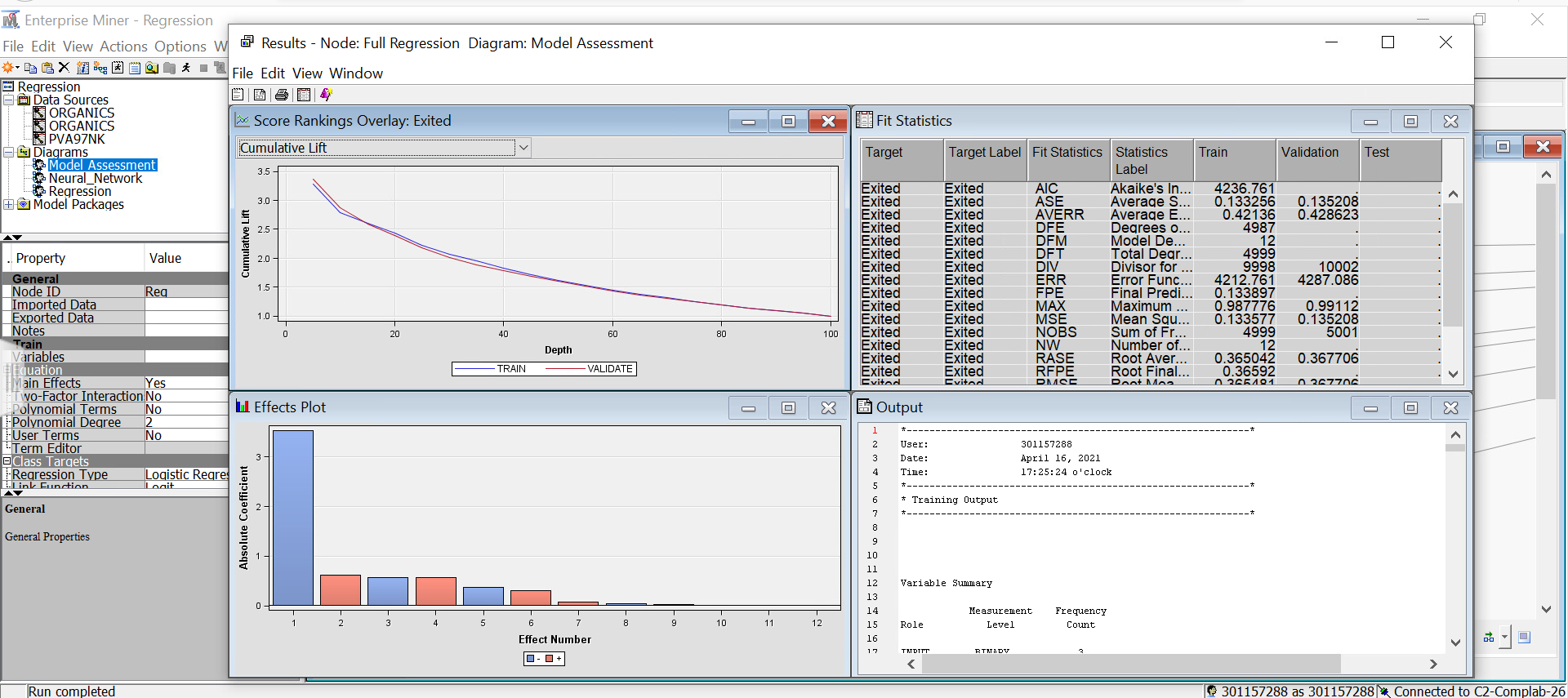


Fig. 9: Full Regression

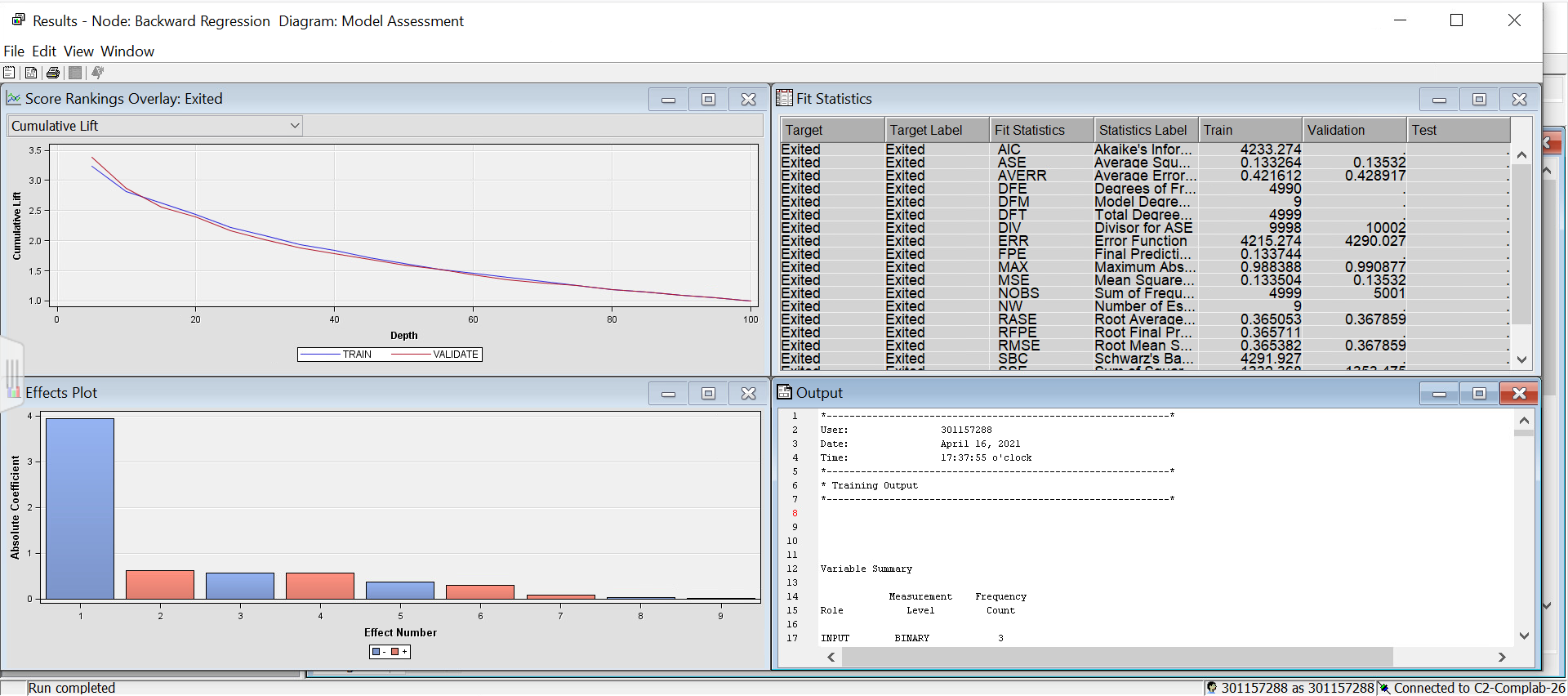


Fig. 10: Backward Regression

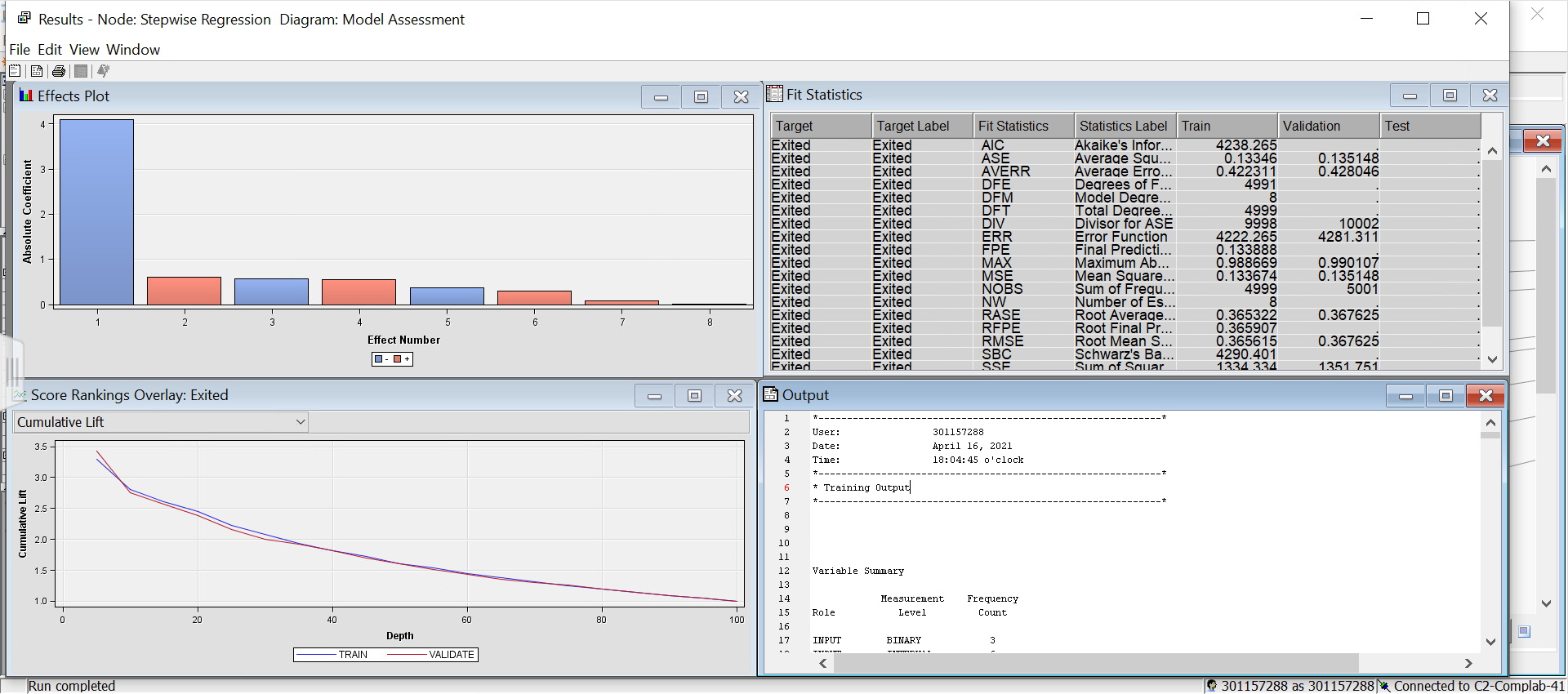


Fig. 11: Stepwise Regression

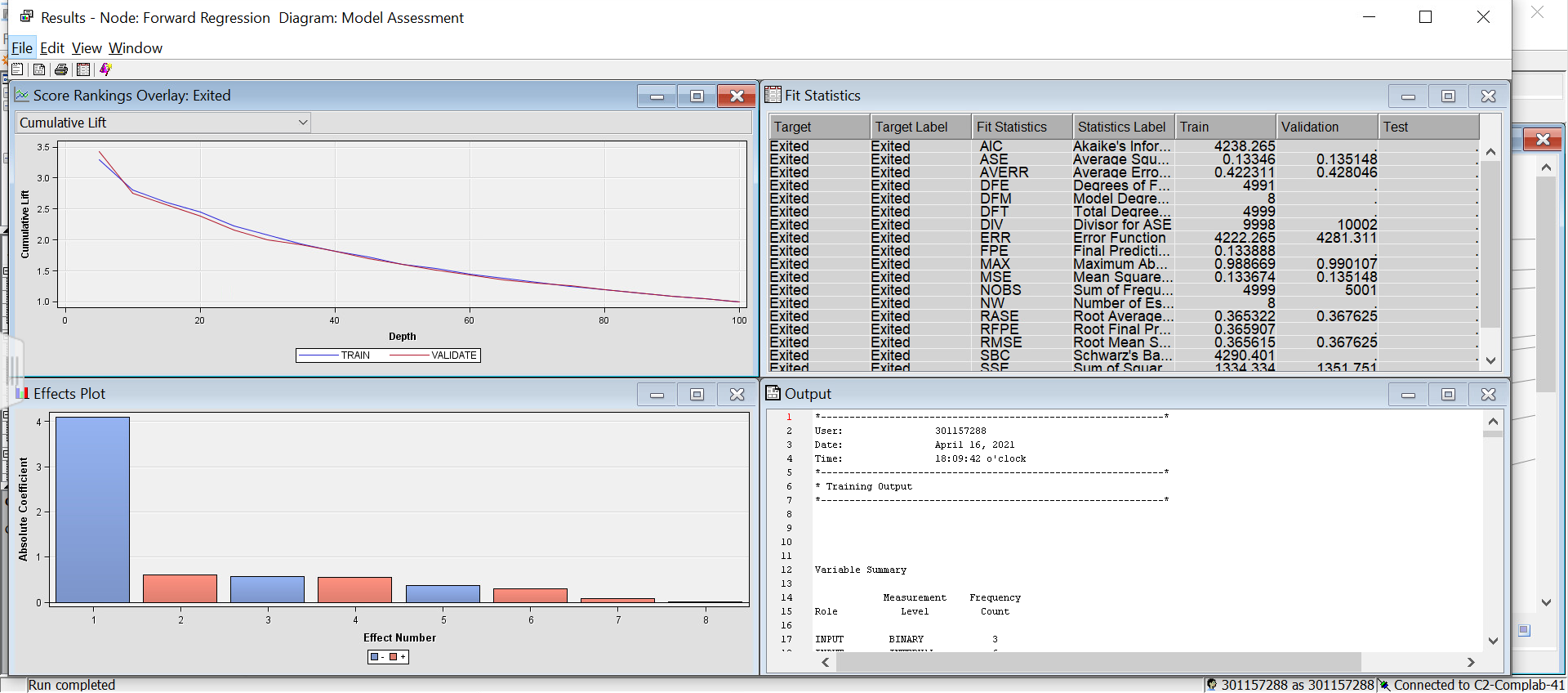


Fig. 12: Forward Regression

* 1. **Neural Network**

A neural network is just like a network of neurons or algorithms that recognize the relationships in between data through a process similar to the human brain operates. The process of referring similar to the neurons of human brain artificially is often called as artificial neural network and it works using mathematical function by collecting and classifying information towards specific architecture.



Fig. 13: Neural Network having 3 hidden nodes

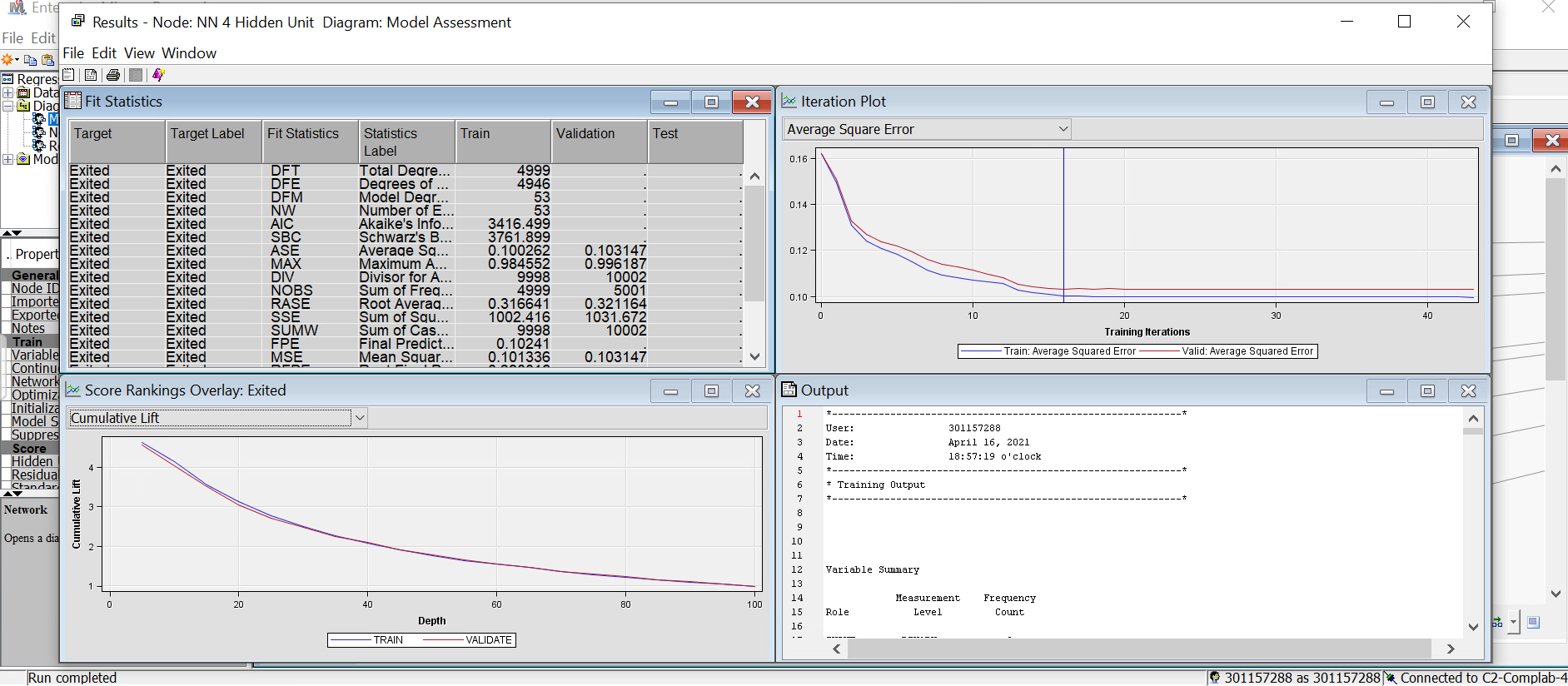


Fig. 14: Neural Network having 4 hidden nodes

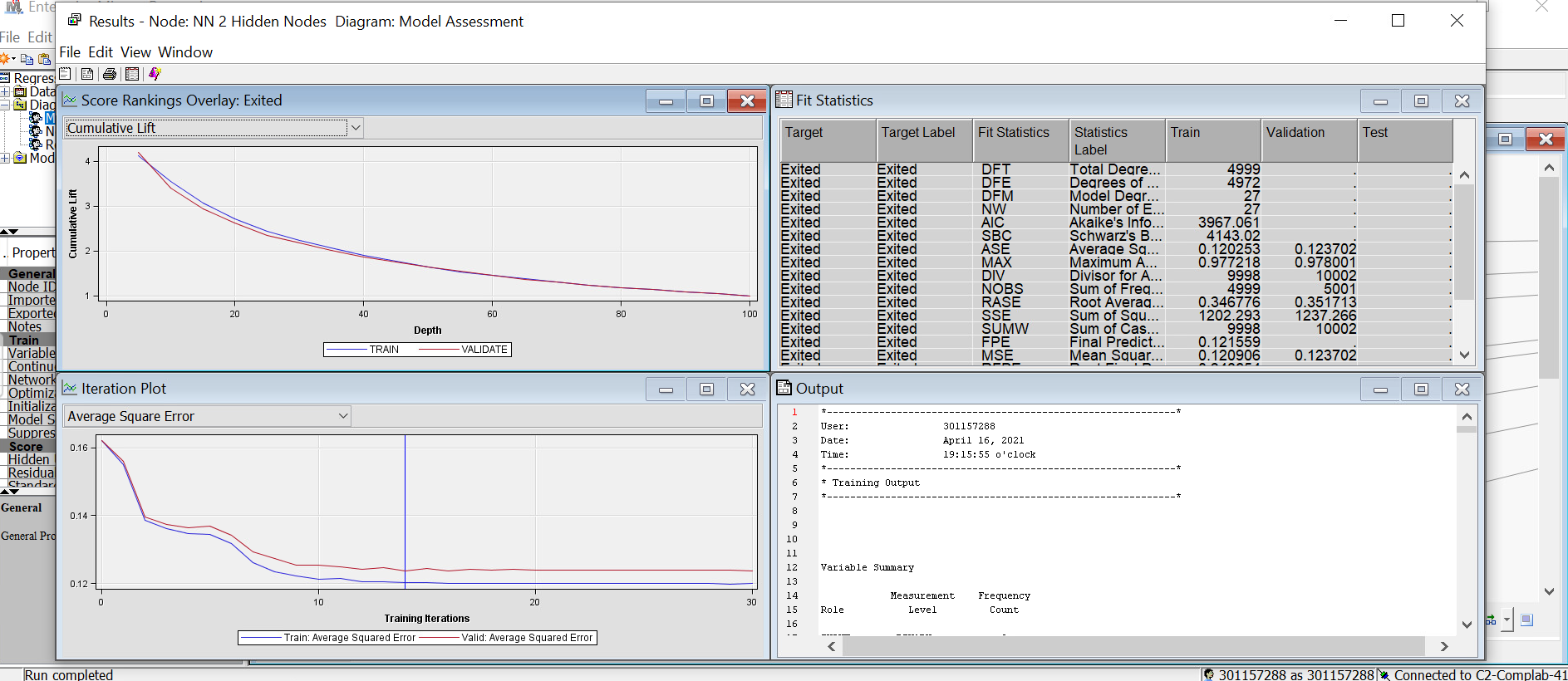


Fig. 15: Neural network having 2 hidden nodes

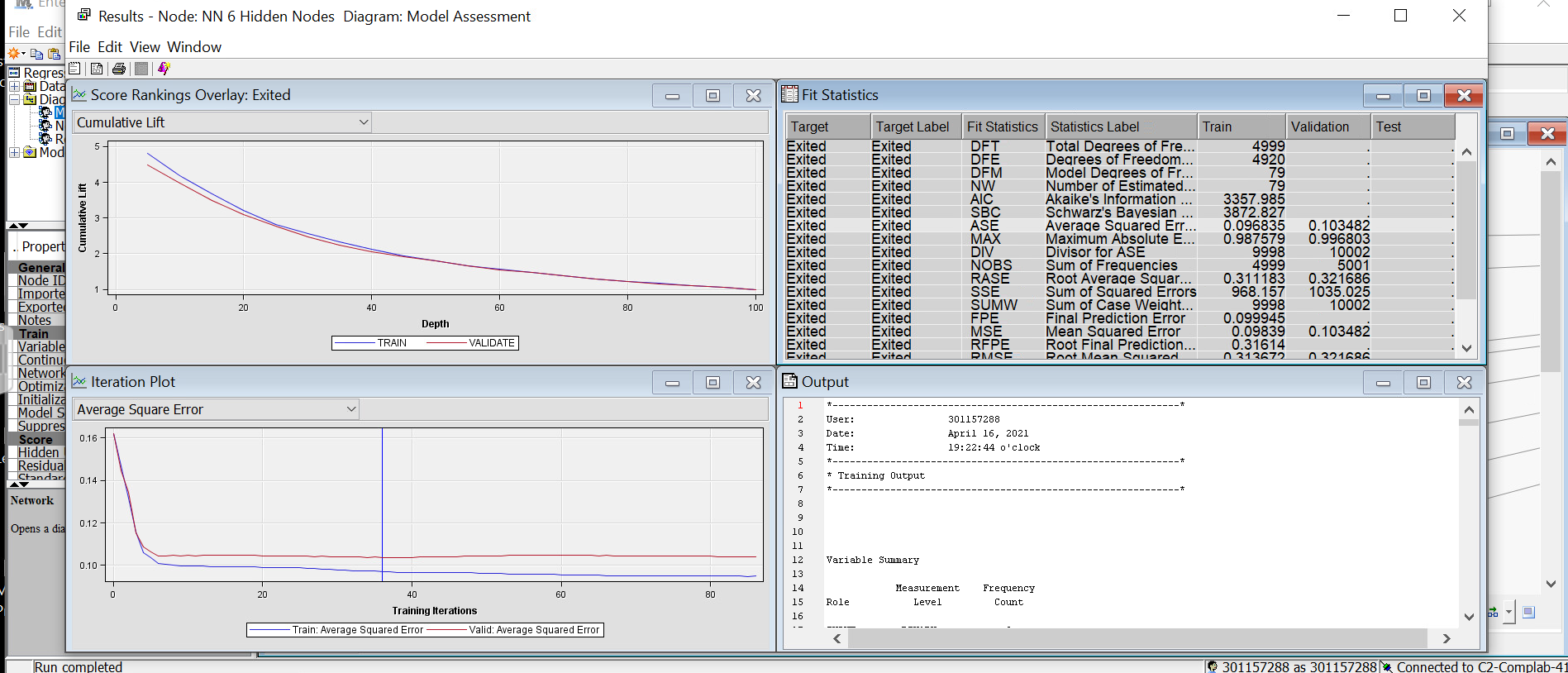


Fig. 16: Neural network having 6 hidden nodes

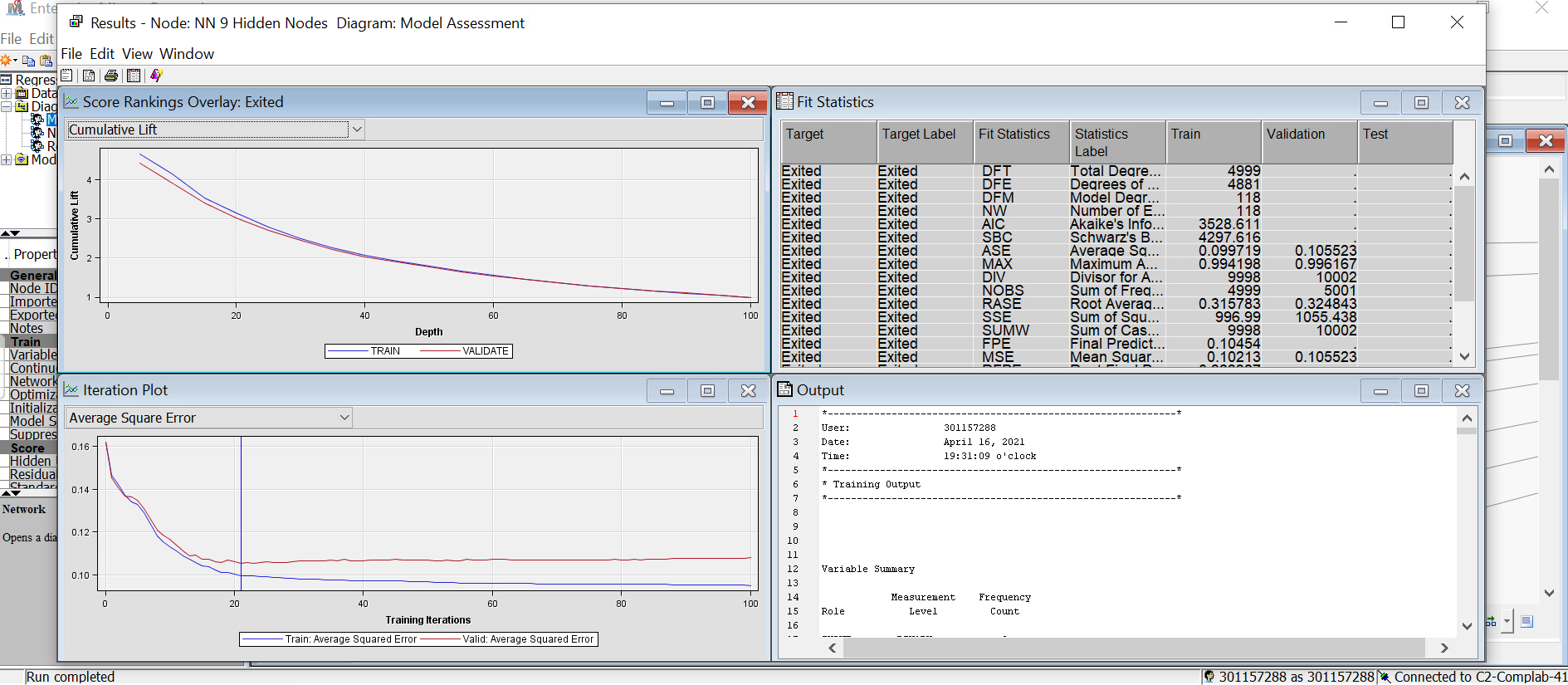


Fig. 17: Neural network having 9 hidden nodes

1. **Model Comparison**

In this section the different models (Decision tree, regression and the neural network model) explained and their results comparison to get best one. These different models are compared to give a best fit model for the purposes of answering the questions of the main objective.

The statistical model selects note by-default and run through all algorithms to determine the best fit algorithm. There are various algorithms are found and can implement what we need to answer or according to our objectives. The most used algorithms are average squared error, mean square error, misclassification rate, receiver operative curve (ROC) and many more.

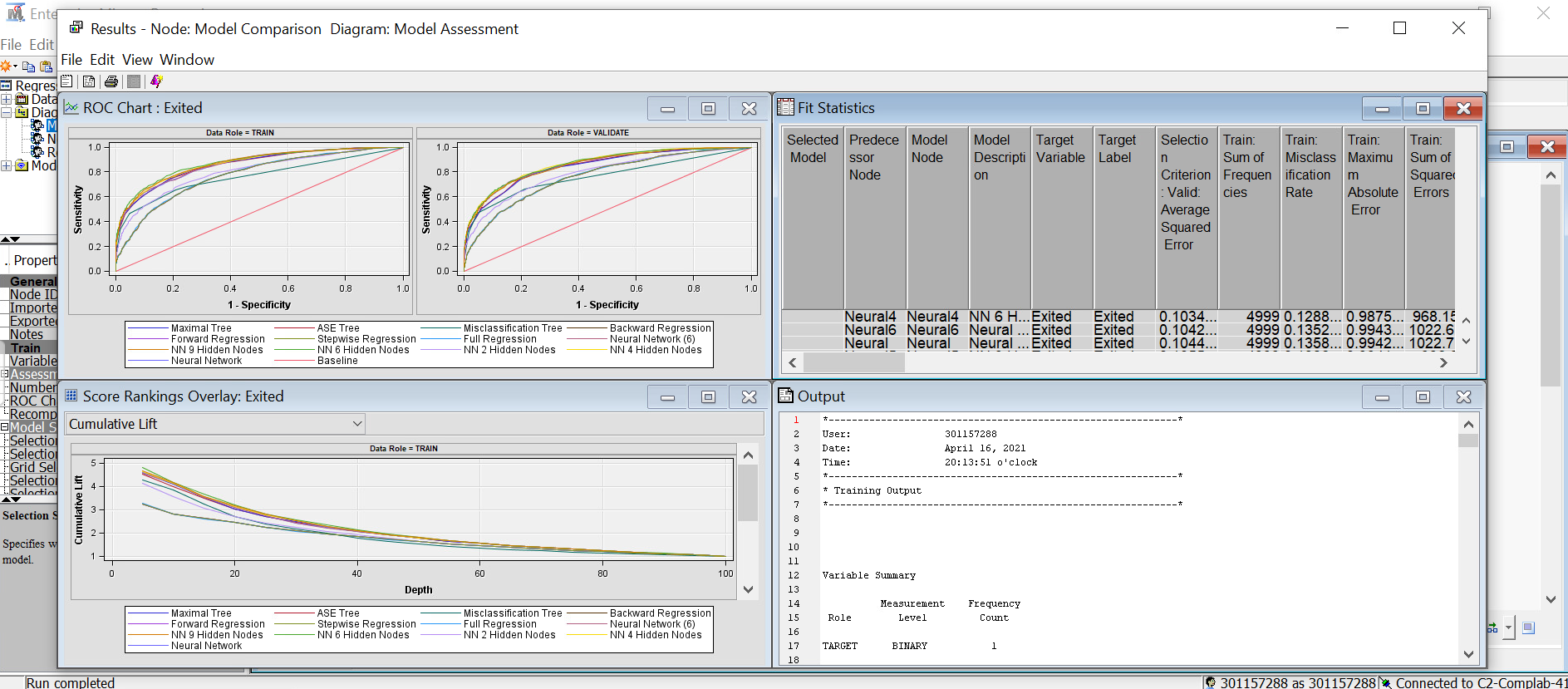


Fig. 18: Final result of comparison

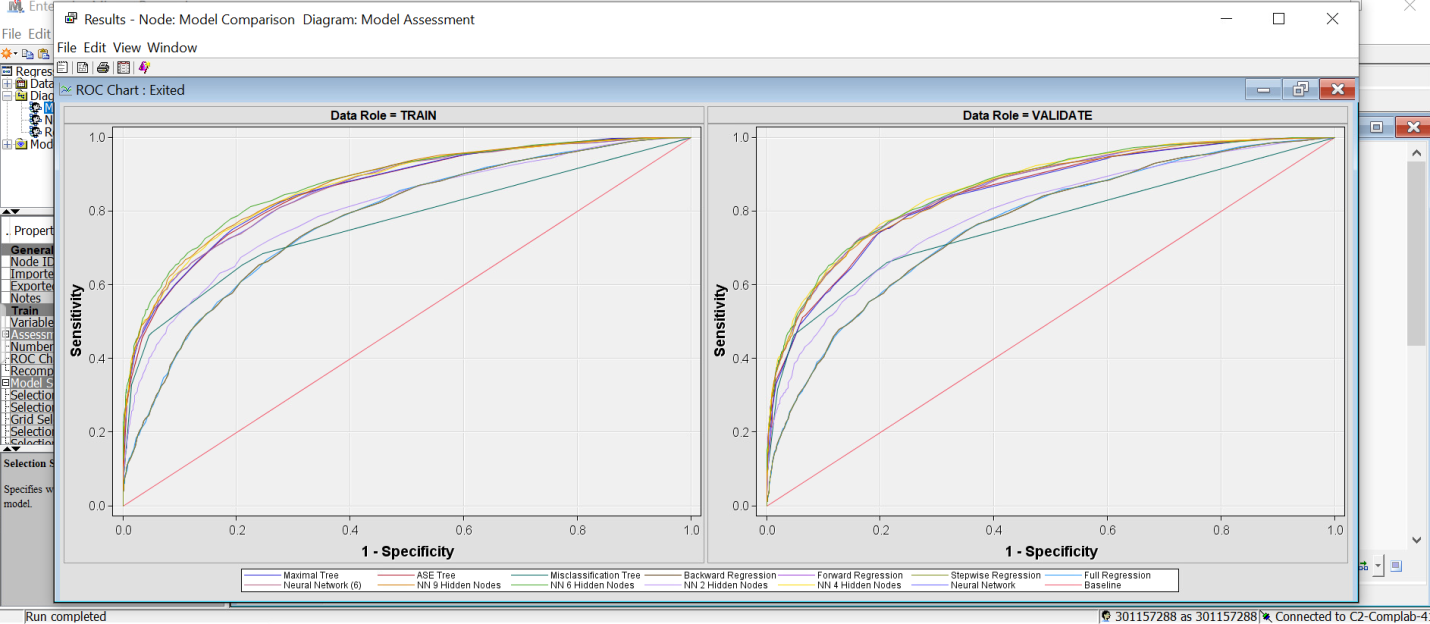


Fig. 19: ROC of all models

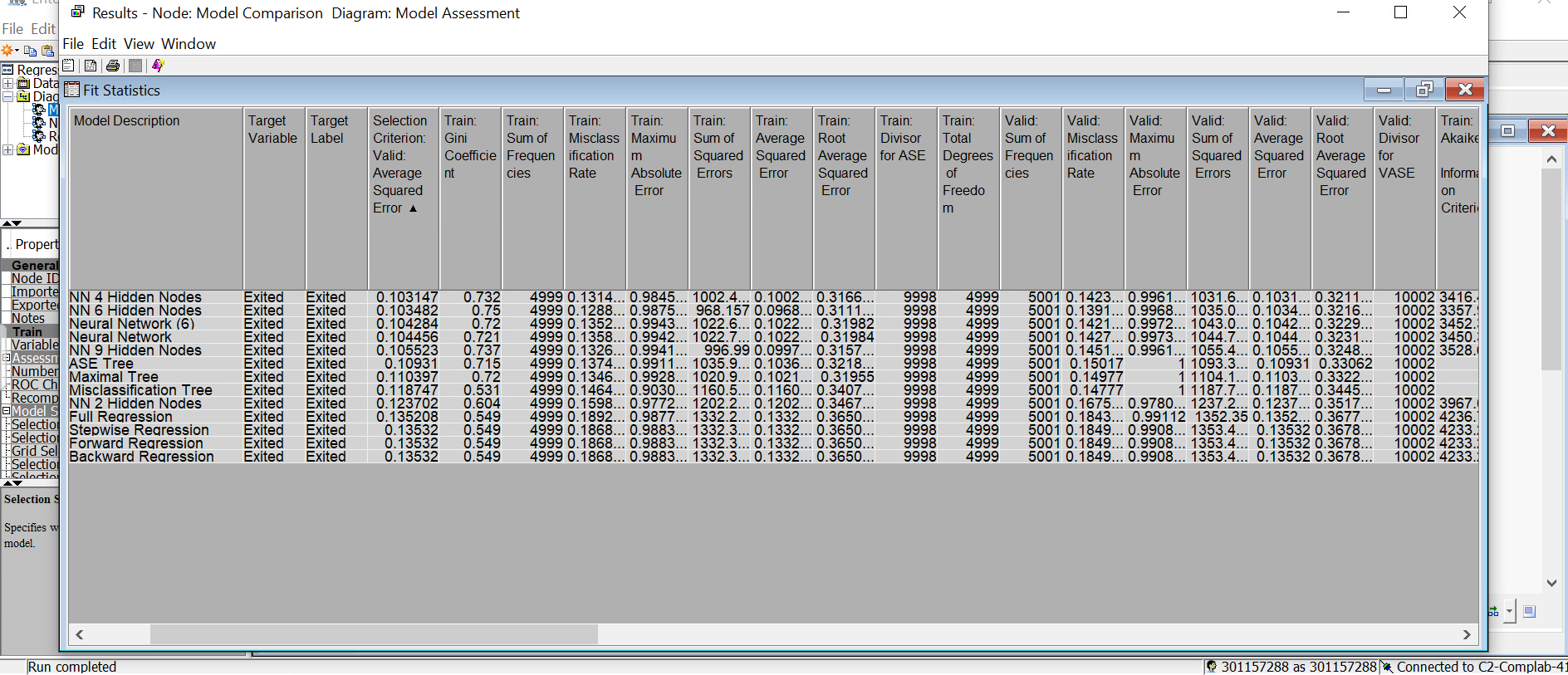


Fig. 20: Node Model Comparison result

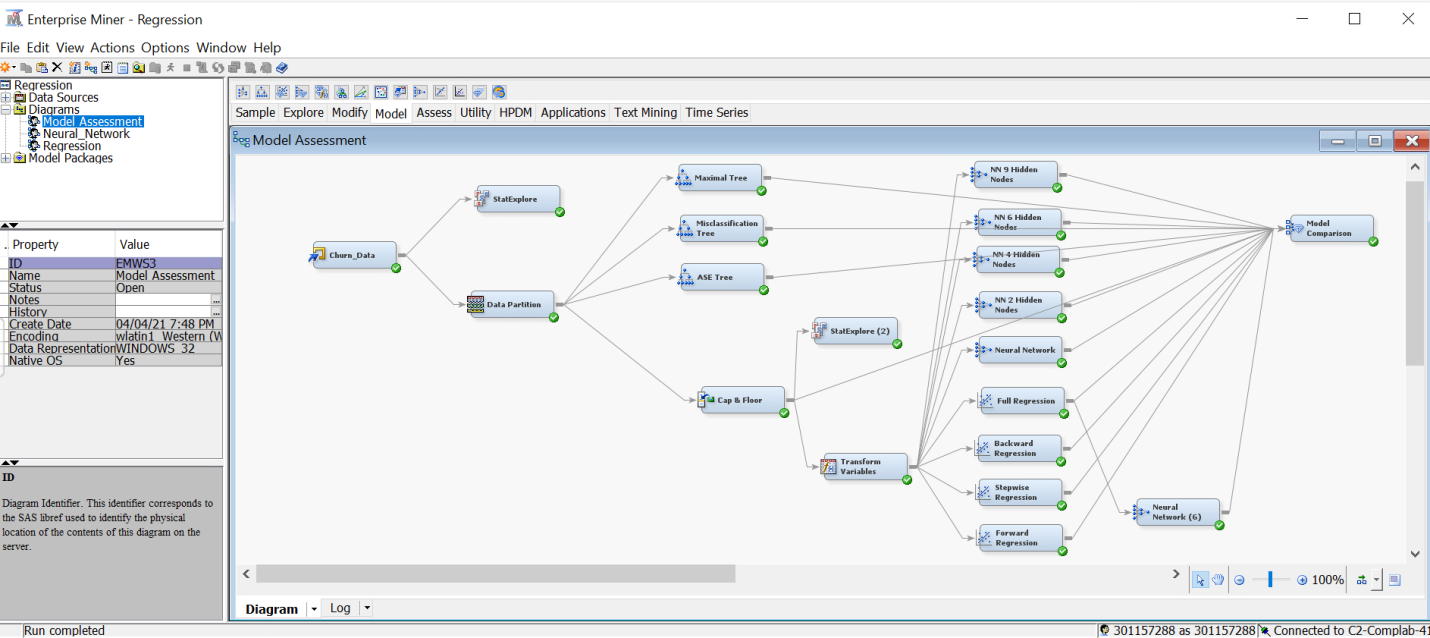


Fig. 21: All process to get final model of the comparison

1. **Conclusion**

The SAS Enterprise Minor can be used to analyze customer’s behavior and this analysis is also analyzing customer risk in a real-time setting using Bank-Churn data. To determine the best model of customer behavior of bank-Churn data, there are several models performed. The decision tree, regression and neural networks are three models used in this analysis, but inside these three different models there are also sub-divisions on each. To get the best result, the analyses of several variables in different combinations were used in different models. These models give us best results in terms of validation data and training data that splits equally half and optimally reduce the skewness of the data for this analysis.

Using the receiver operating curve (ROC) the best model comparison shows the Neural Network having four hidden nodes best describes the churn exited for both training data and validation data. Not only from ROC but also using average squared error, the neural network having 4 hidden nodes the best promised model. The full regression is best among all other regression models. In consideration with the average squared error the Decision Tree model is best fit in the bank-churn data to determine Exited.

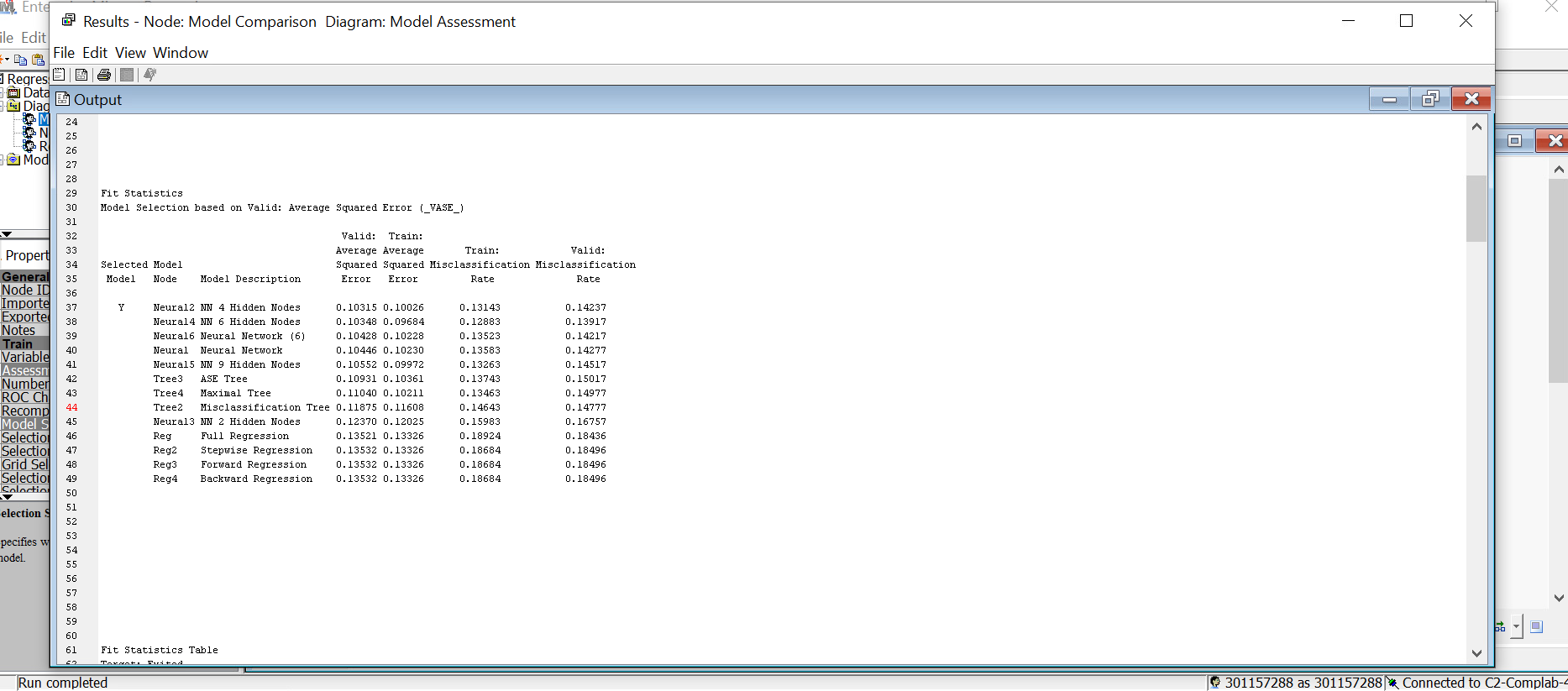


Fig. 22: Final output of Average Squared Error showing best model

1. **Bibliography**

* SAS Enterprise Miner Manual and Help guides
* Data Mining Using SAS Enterprise Miner – A Case Study
* Getting Started with SAS Enterprise Miner 13.1
* <http://support.sas.com/resources/papers/proceedings13/086-2013.pdf>
* <https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/>
* <https://www.investopedia.com/terms/n/neuralnetwork.asp>
* <https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling>