Bank Marketing Trend

**Team Number: 04**

Team Members: NetID:

Akanksha Mishra axm170031

Harman Singh hxs134930

Lokesh Kumar Chaturvedi Lolla lxl163230

Yugandhara Rawool yer170030

**Introduction:**

The verge of financial crisis back in 2008 hit the financial markets bad around the world. European banks are no exception for this and they needed customers to increase capital to cope up with the situation. A fitting model to predict the marketing result using phone calls to sell term deposits would provide the Portuguese banks with valuable information that facilitate selection of better customer base.

The aim of our project is to predict and choose a fitting model, which help the Portuguese bank to be aware of significant attributes that lead a customer into term deposits and make decisions accordingly to improve the marketing campaign strategies.

We have used primary data modelling approach for prediction. We have used decision tree, regression and neural network for the analysis.

**Project Background:**

The data set we used is a second-hand data set, derived from the ‘UCI Machine Learning repository’. The data is captured from direct marketing campaigns of a Portuguese banking institution. It contains the details of the individual customers and their responses to phone calls. A customer was reportedly contacted more than once as part of this campaign to know if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

**Description of Data:**

The data contains rows and columns related to the customer and their interaction with telemarketing campaign.

**Output Variable (Target):** **Term deposit**

It is the output or dependent variable which is binary (‘Yes’ or ‘No’). All the variables constituted in the data set are presented below with descriptions of each.

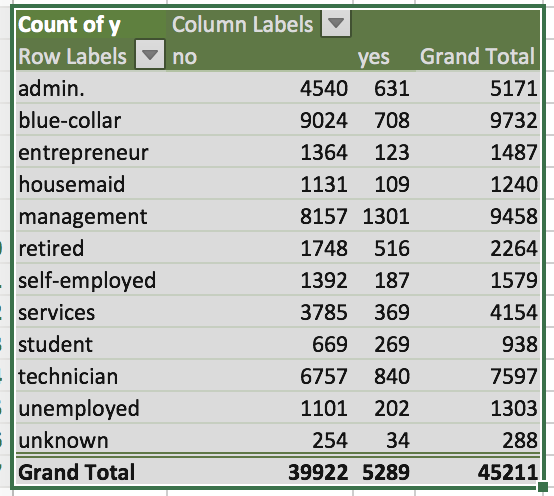
|  |  |  |
| --- | --- | --- |
| **Variables** | **Description** | **Type of data** |
| Age | Age of client | Numeric |
| Job | Type of job | Categorical |
| Marital | Marital status | Categorical |
| Education | Level of education | Categorical |
| Default | Has credit in default? | Binary |
| Housing | Has housing loan? | Binary |
| Loan | Has personal loan? | Binary |
| Contact | Contact communication type | Categorical |
| Month | Last contact month of year | Categorical |
| Duration | Last contact duration | Numeric |
| Campaign | Number of contacts performed during this campaign and for this client | Numeric |
| Pdays | Number of days that passed by after the client was last contacted from a previous campaign | Numeric |
| Previous | Number of contacts performed before this campaign and for this client | Numeric |
| Poutcome | Outcome of the previous marketing campaign | Categorical |
| Term deposit | Has the client subscribed a term deposit? | Binary: 'yes’, ‘no' |

**Data Preparation (Pre-processing):**

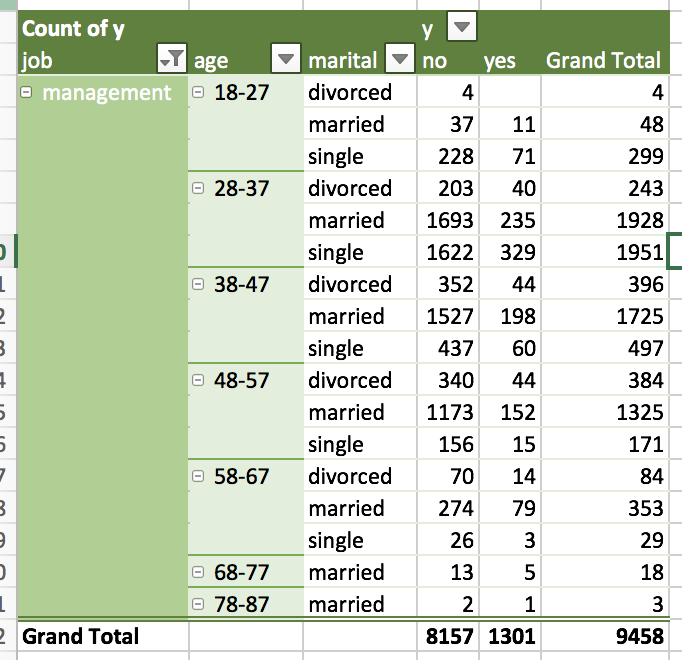
In every data mining problem, data preparation is the most crucial, difficult and longest part of the mining process. In this section, we will be performing the following tasks using SAS enterprise Miner:

1. Exploring statistical properties of the variables in the input dataset.
2. Partitioning the input data into Training and Validation data sets.
3. Handling of the missing data.

Initially, using the Microsoft Excel we arranged the dataset into respective columns and did a preliminary analysis on it. The results are shown below.



Using a pivot, we attempted to throw the raw data into various views that could help us understand what we could infer. Based on the table above we see that majority of successful sales came from employees in a management position. This revelation led us to believe that we should look at other variables in tandem such as age and marital status



Once again, the results help us find out that those management employees, in the age range of 28-37, and who are single are most likely to purchase the bank’s products. Ultimately, we knew that this was just a first pass at analysing the data, but it showed us what variables we knew we had to drill down on. Our goal with this manual data dive was to create a baseline that would allow us to measure the accuracy of our thought process and the models we will soon create.

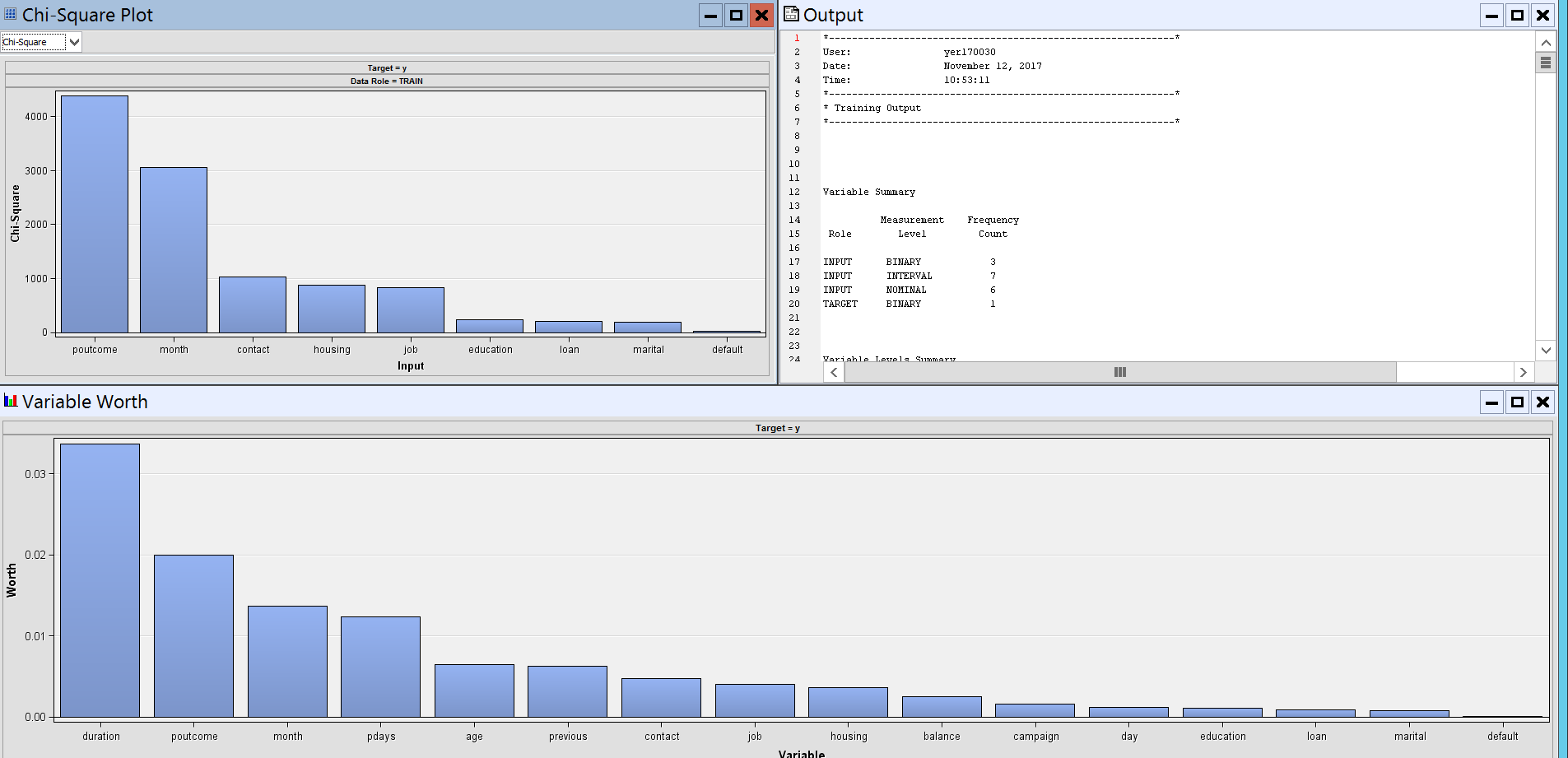
A new diagram is added and using **File Import** node the data is loaded data from Excel workbook. We are keeping the role as input for all the variables and have set the role of Term Deposit (Variable y) as Target. The level used in the dataset are Nominal, Interval and Binary.

1. **Summary Statistics:**

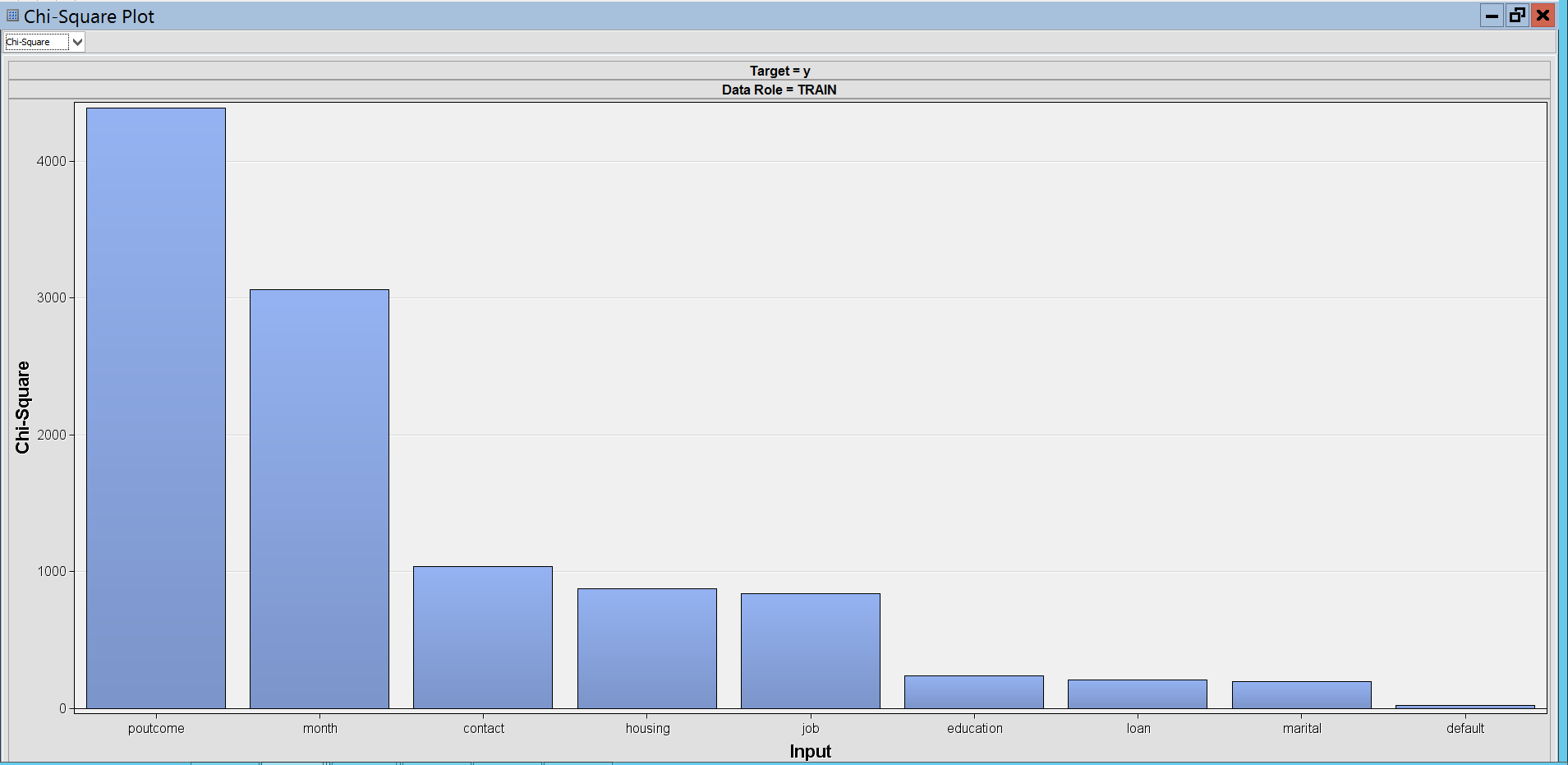
**Stat Explore**, a multipurpose exploration tool is used to examine variable distributions and related summary statistics. The description of the Term deposit is as below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Role** | **Level** | **Reasons** |
| Term Deposit | Target | Binary | Target Variable |

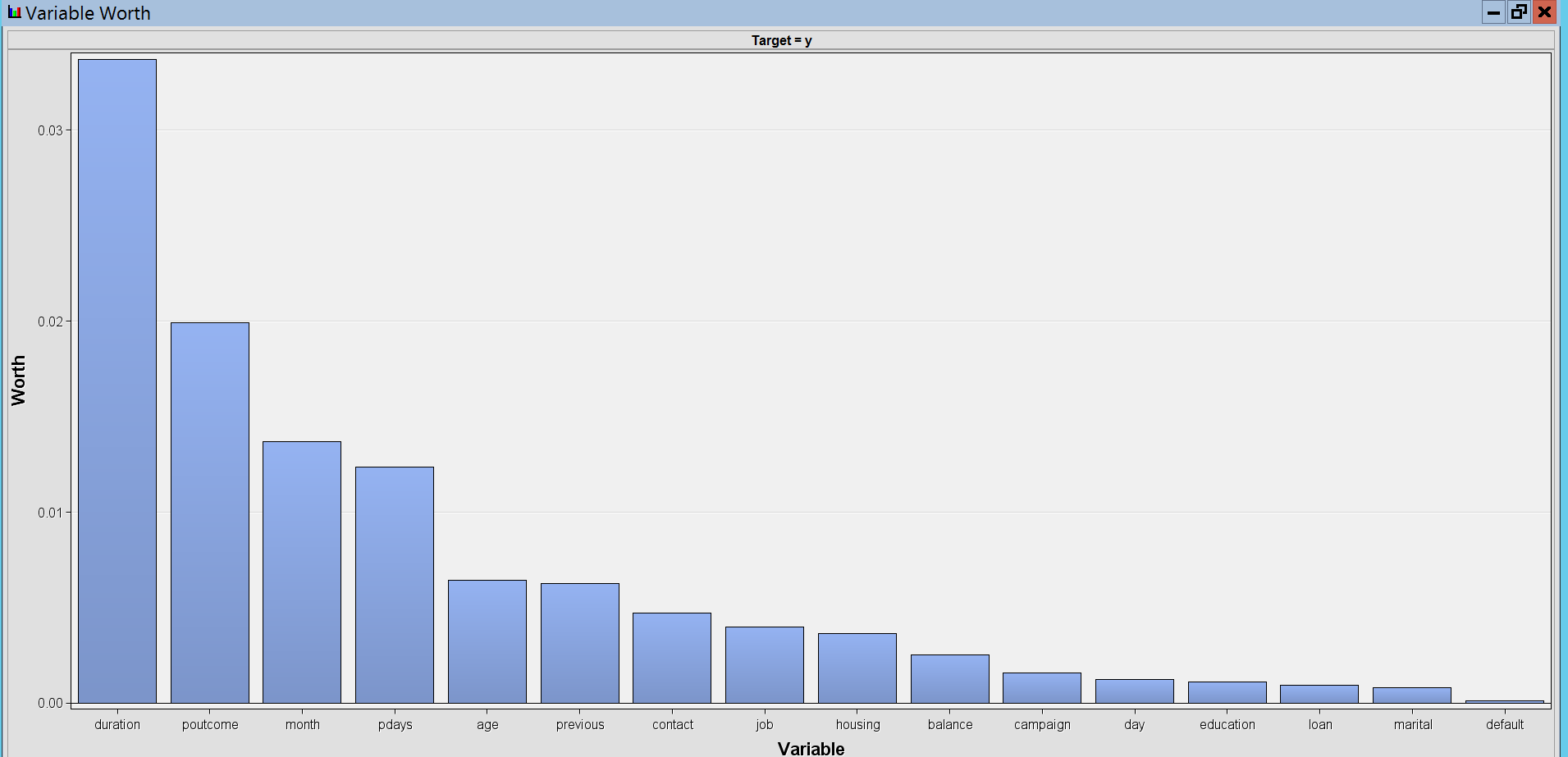
The Stat Explore statistics are shown below. The Result window provide us with the Chi-Square Plot, Variable worth and the Output.



The Chi-Square Plot for the input variable is as below:



The variable worth for the input variable is as below:



We can see that the Variable worth is highest for the variable duration, followed by poutcome, month, pdays, age, previous, contact, job, housing etc.

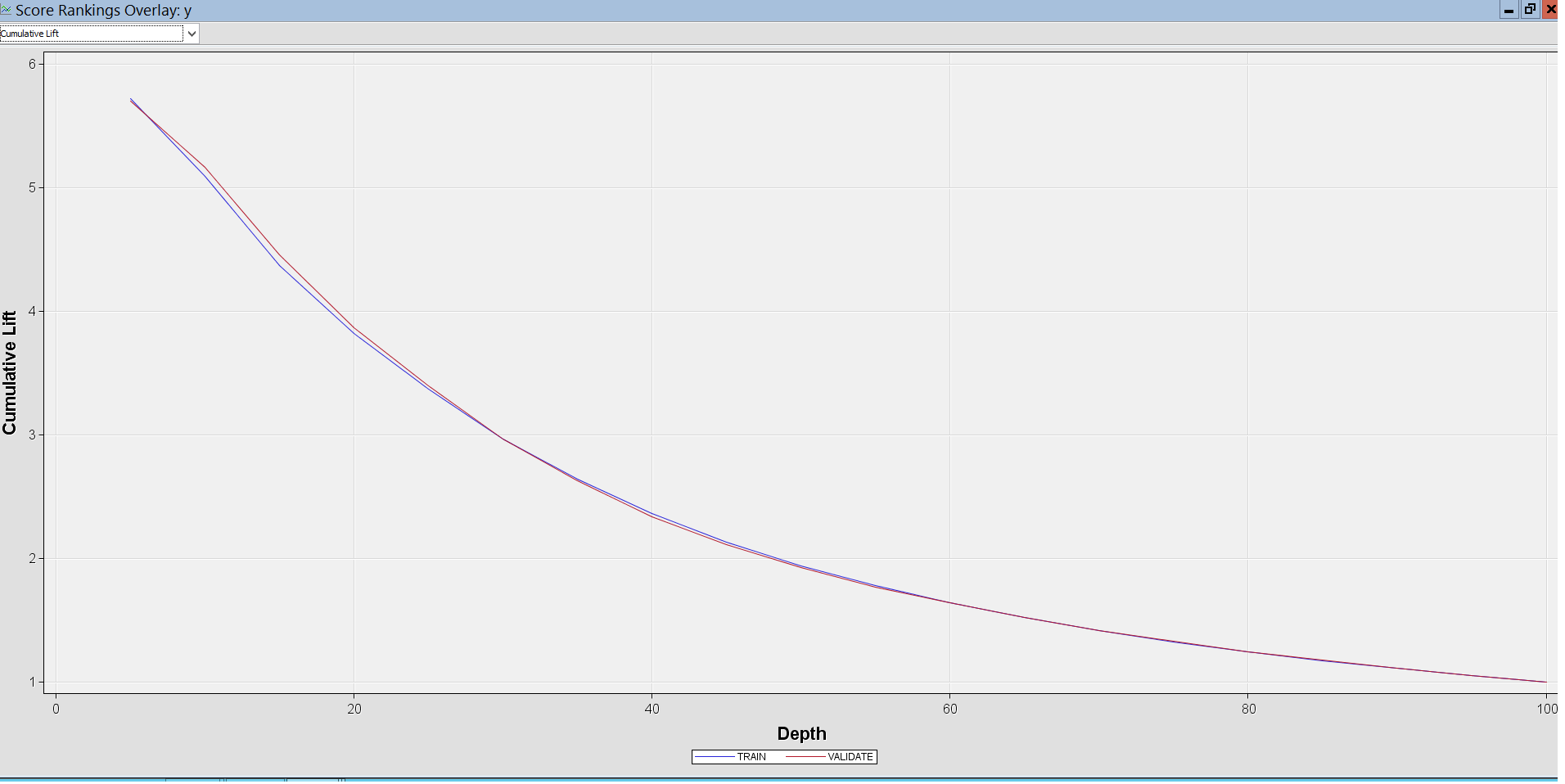
1. **Data Partitioning:**

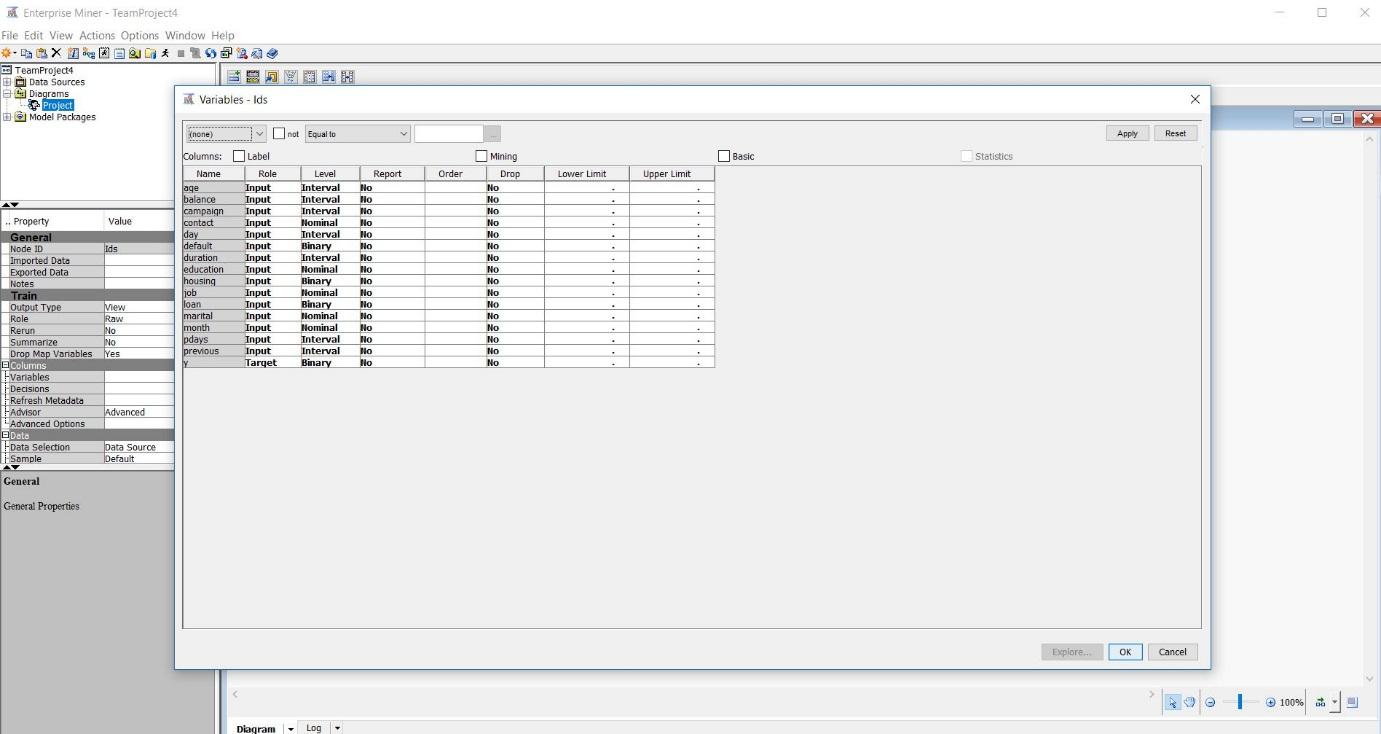
Data partitioning helps in managing the quality of the model during fitting. For preliminary model fitting, training data is used whereas the validation data serves in testing the model empirically without overfitting the data. We partitioned the data into:

1. Training data – 50% of data
2. Validation data – 50% of data

To avoid overfitting problem, we divided our dataset into 50-50% as we are getting least error in validation dataset when we divide our data set like this.

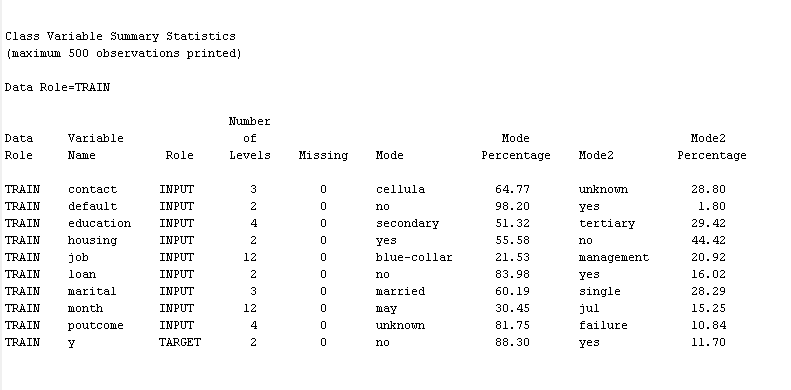
We selected the **Data Partition** node to partition the raw data into training and validation data sets and ran the node to complete the partitioning process. The cumulative lift for the Output ‘y’ is as below.





1. **Missing Values:**

We did not have any missing data in our dataset. The output of Stat Explore confirms the same as mentioned below.



**Data Modelling:**

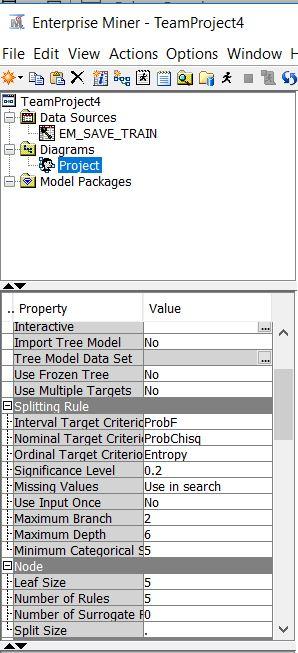
After the inspection of the data we have decided to go with the three primary approaches: Decision trees, Logistic Regression and Neural Networks to work upon our data and then compare the results to choose the best among them.

1. **Decision Trees:**

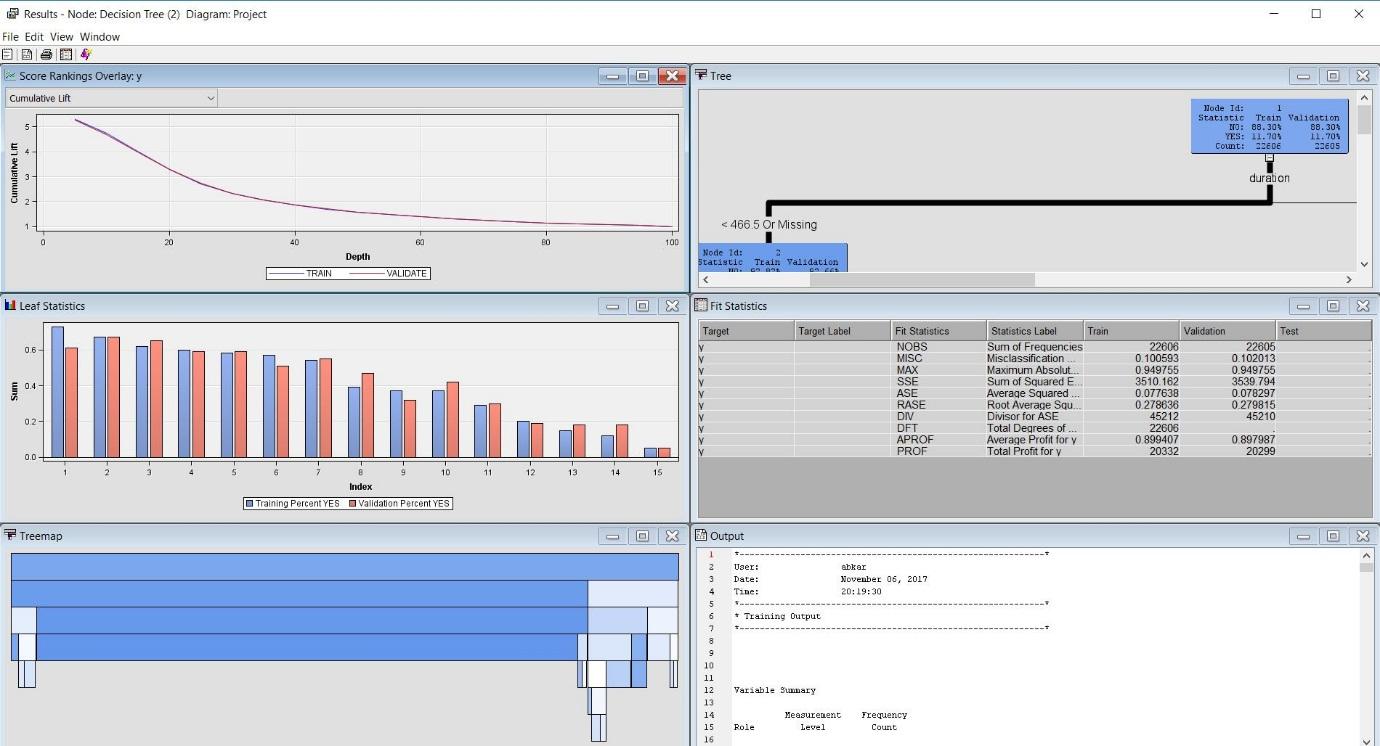
We have considered average squared error to be the assessment measure for decision trees.We model the input data using different decision trees:

1. To maximize the split decision logworth, we automatically train a full decision tree and prune it to size by selecting split rules to 2-splits
2. To maximize the split decision logworth, we automatically train a full decision tree and prune it to size by selecting split rules to 3-splits
3. **Default Decision Tree(2-Split):**

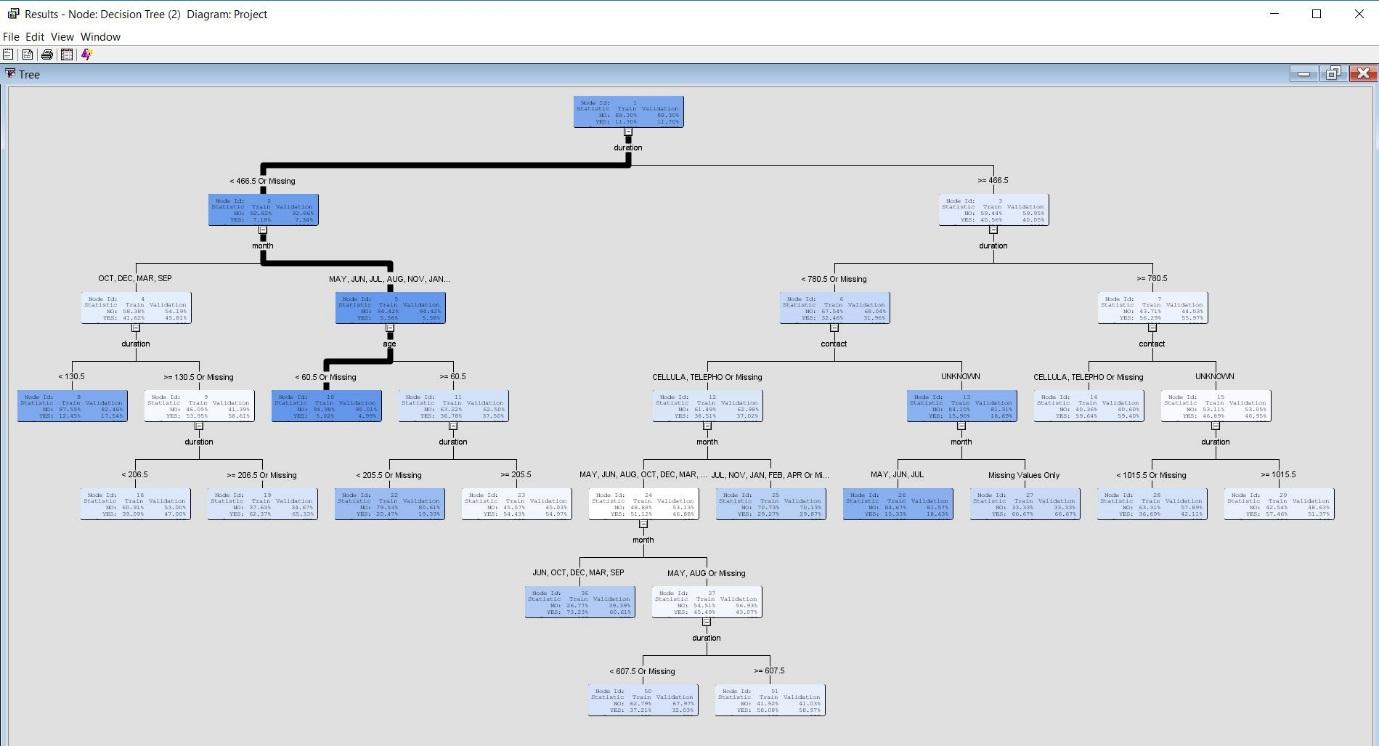
We ran the decision tree with default setting. The below are the parameters for this model:



The Result generated is as follow. It provides us with the Cumulative lift, Leaf statistics, Treemap, Tree diagram, Fit Statistics and output.



The Tree generated is as below.

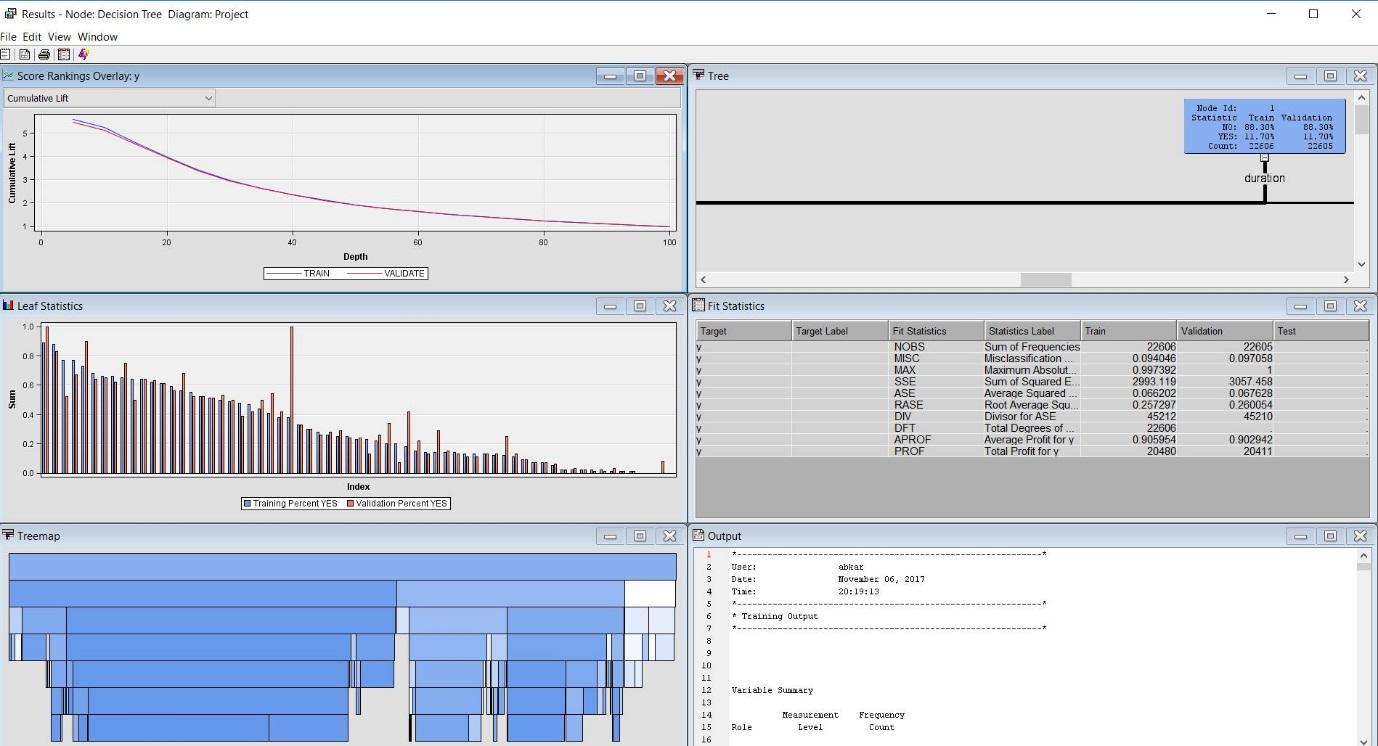


From the above we can observe the following

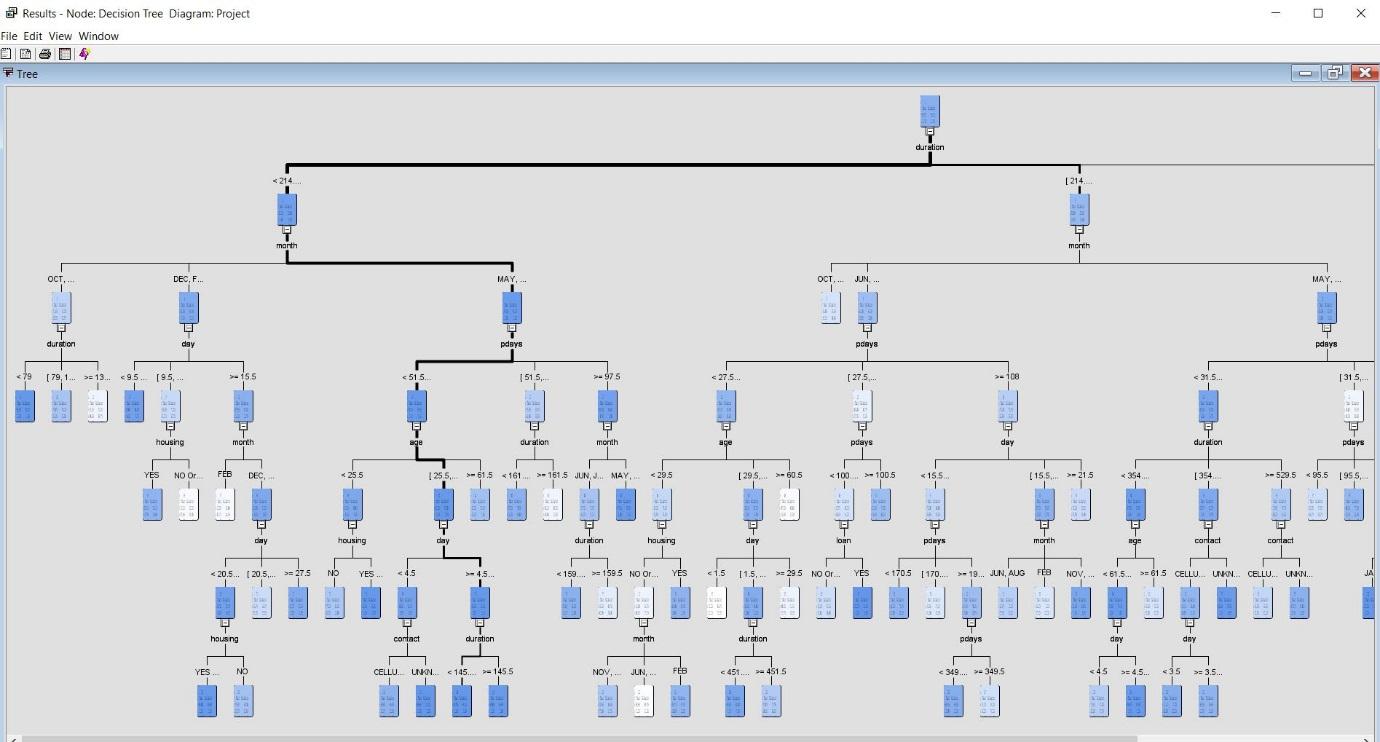
* The first split variable by the model was **duration.**
* Total number of leaves: 22
* Overall depth of the tree generated: 6

1. **Forced Split Decision Tree (Split -3):**

We ran the decision tree after setting the value of **Maximum Branch** in **Splitting Rule** to 3. The below are the parameters for this model:



The Tree Diagram is as below.



From the above we can observe the following

* The first split variable by the model was **duration.**
* Total number of leaves: 55
* Overall depth of the tree generated: 6

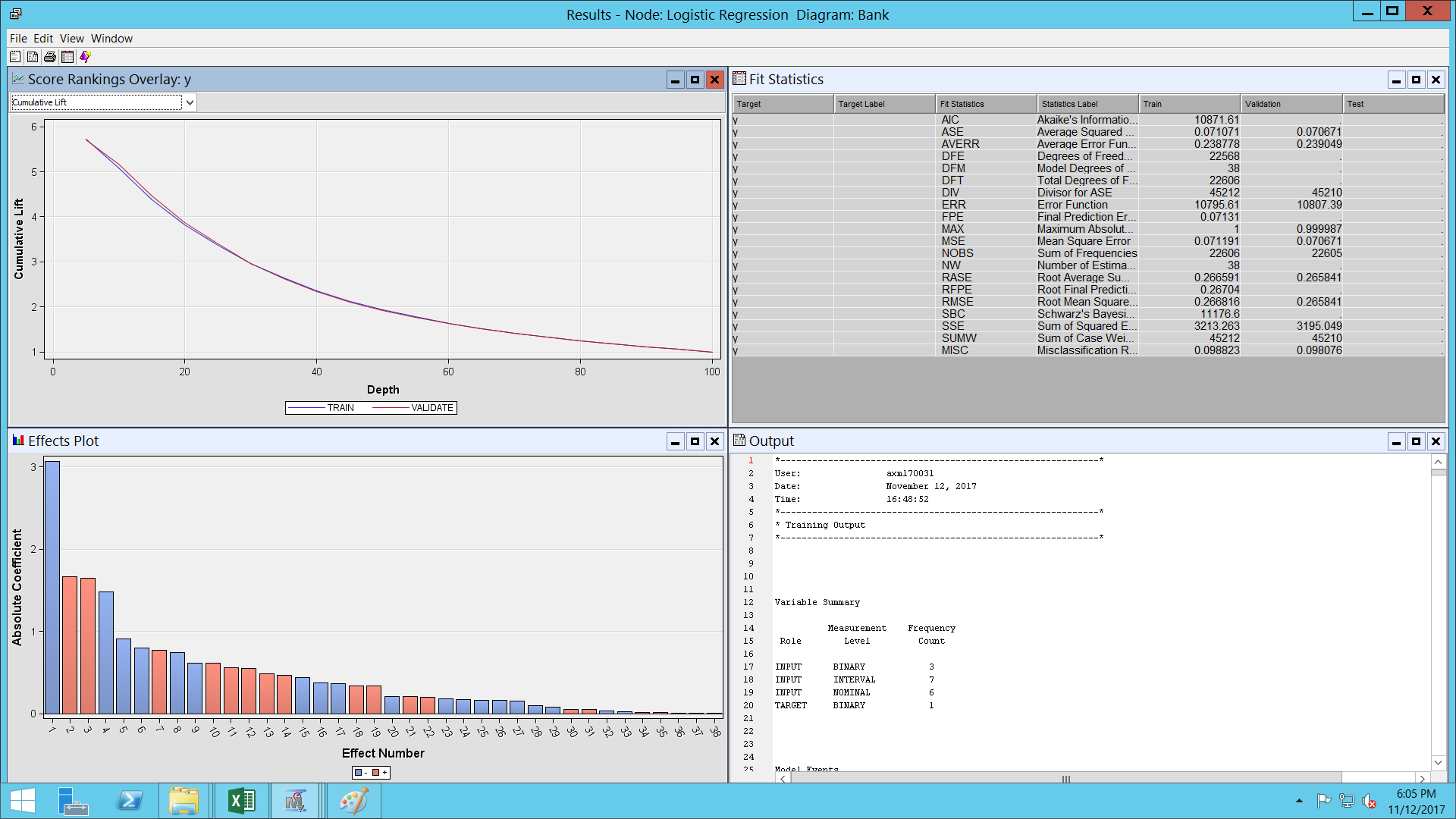
1. **Logistic Regression:**

As a second approach we used logistic regression. As we did not have the missing values in the input data we can directly use the data which was partitioned without the impute node. There are 3 ways to perform logistic Regression in Enterprise Miner:

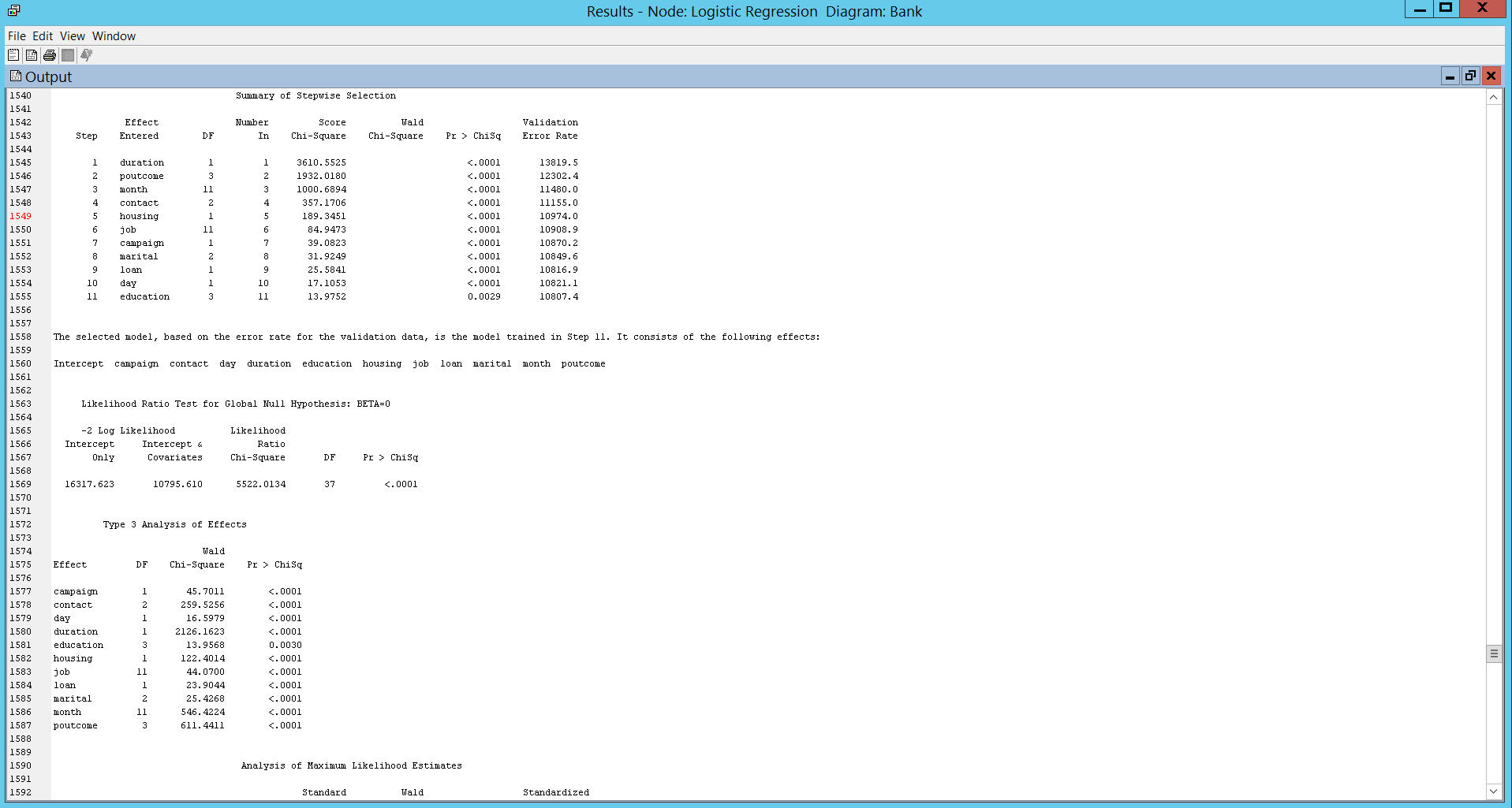
* **Forward:** We began with an empty equation no variable in the model. Predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable. Variables of greater theoretical importance (lowest **P value**) are entered first and so on until all the variables are added.
* **Backward:** After the entire model has been created considering all the variables in the data set, it starts eliminating the predictors or variables with highest P value as they are not significant or less significant in predicting the output variable, but it required threshold value to stop this elimination process.
* **Stepwise:** It is a combination of both forward and backward approach. It starts with no variable and create tree by adding variables with low p value and after the tree has been created, it starts backward approach and delete variables with high p value until the threshold (usually all p values <0.05).

We used stepwise regression because the all variables or predictors used in predicting the output in the final model are high valued and statistically significant.

**The results of the regression analysis are shown below.**

****

The output for the regression model is as below.

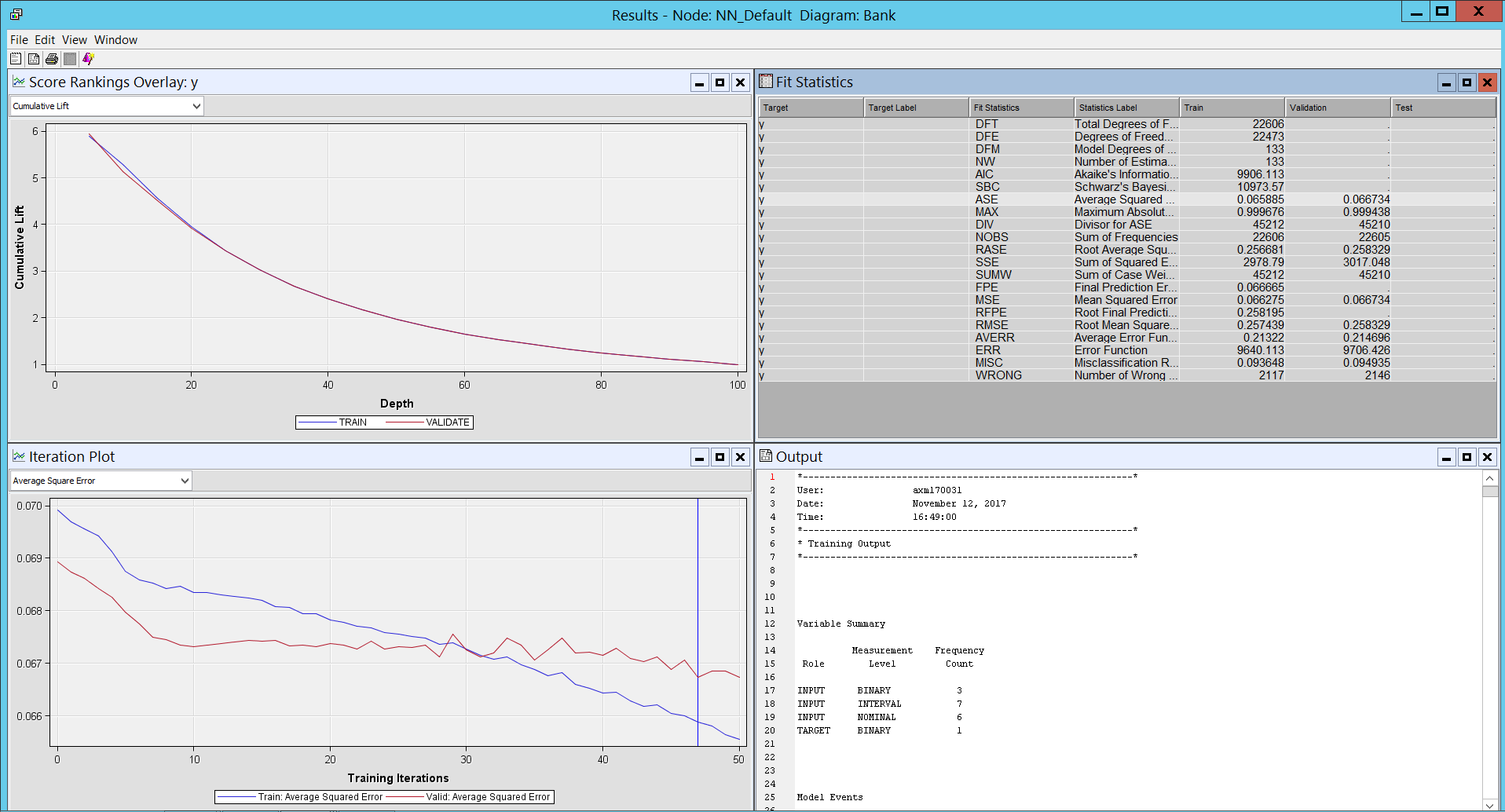


1. **Neural Network:**

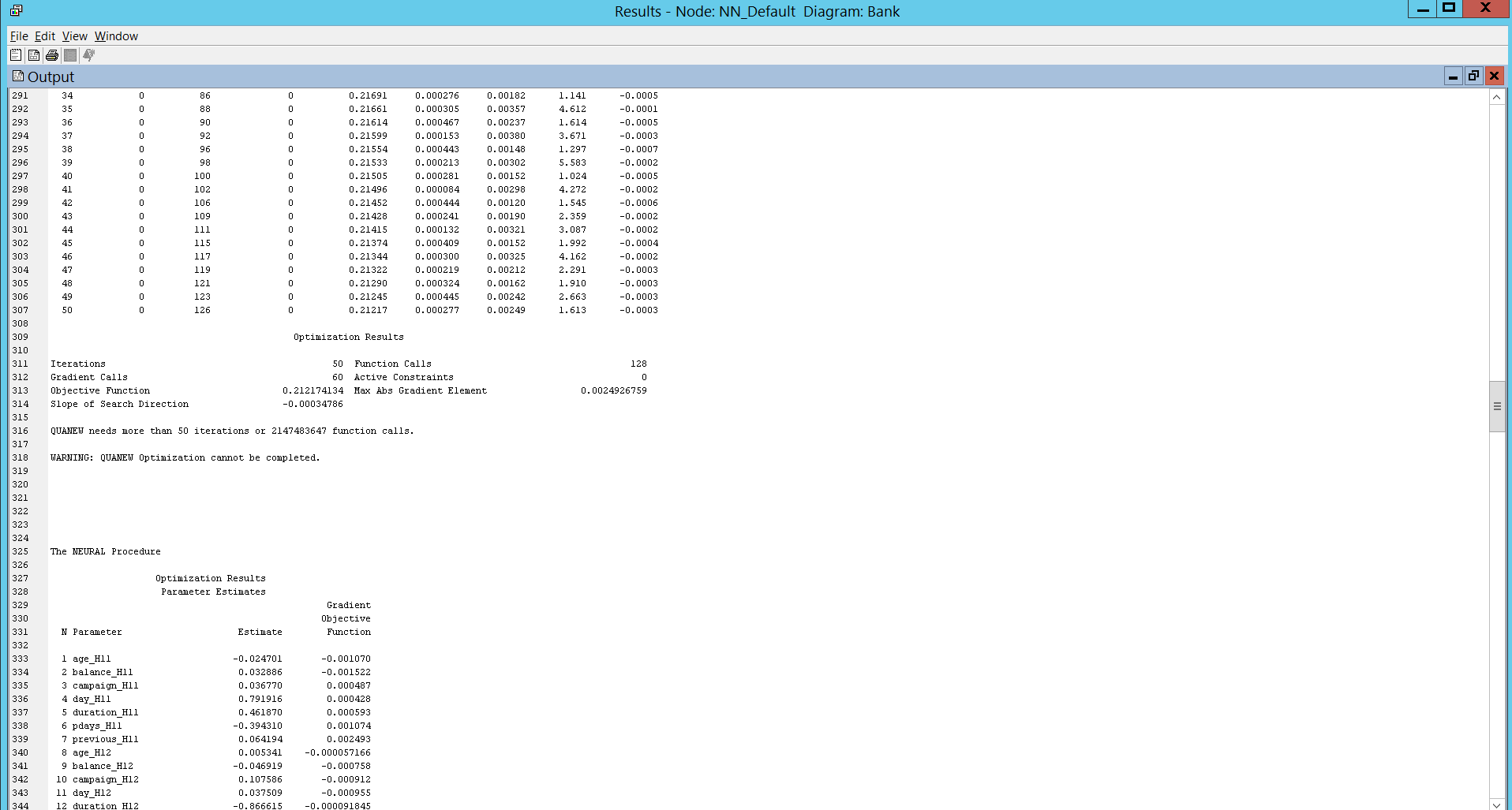
Neural Networks can handle a wide variety of nonlinear relationships between a set of predictors and a target variable, often better than regression models.

1. Neural network analysis of the dataset from data partitioning keeping default parameters.
2. Neural network analysis of the dataset from data partitioning changing number of nodes.
3. Neural network analysis of the dataset from logistic regression done from earlier step.
4. **The Neural network node is added to data partitioning node (Default Parameters) and run**.

The results are depicted below.

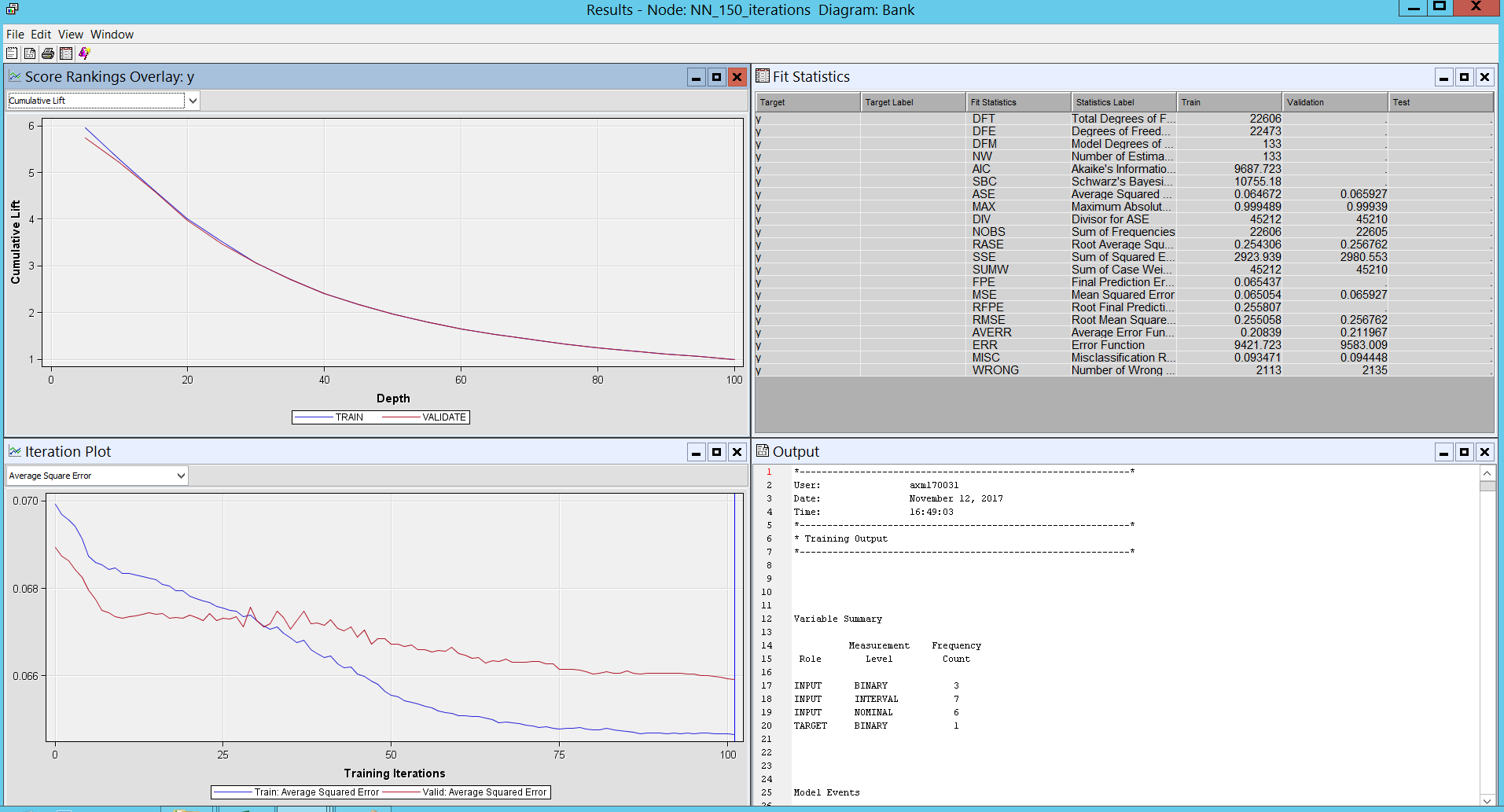


The output of the Neural network is as below. The result is that the optimization cannot be completed and requires more than 50 iterations.

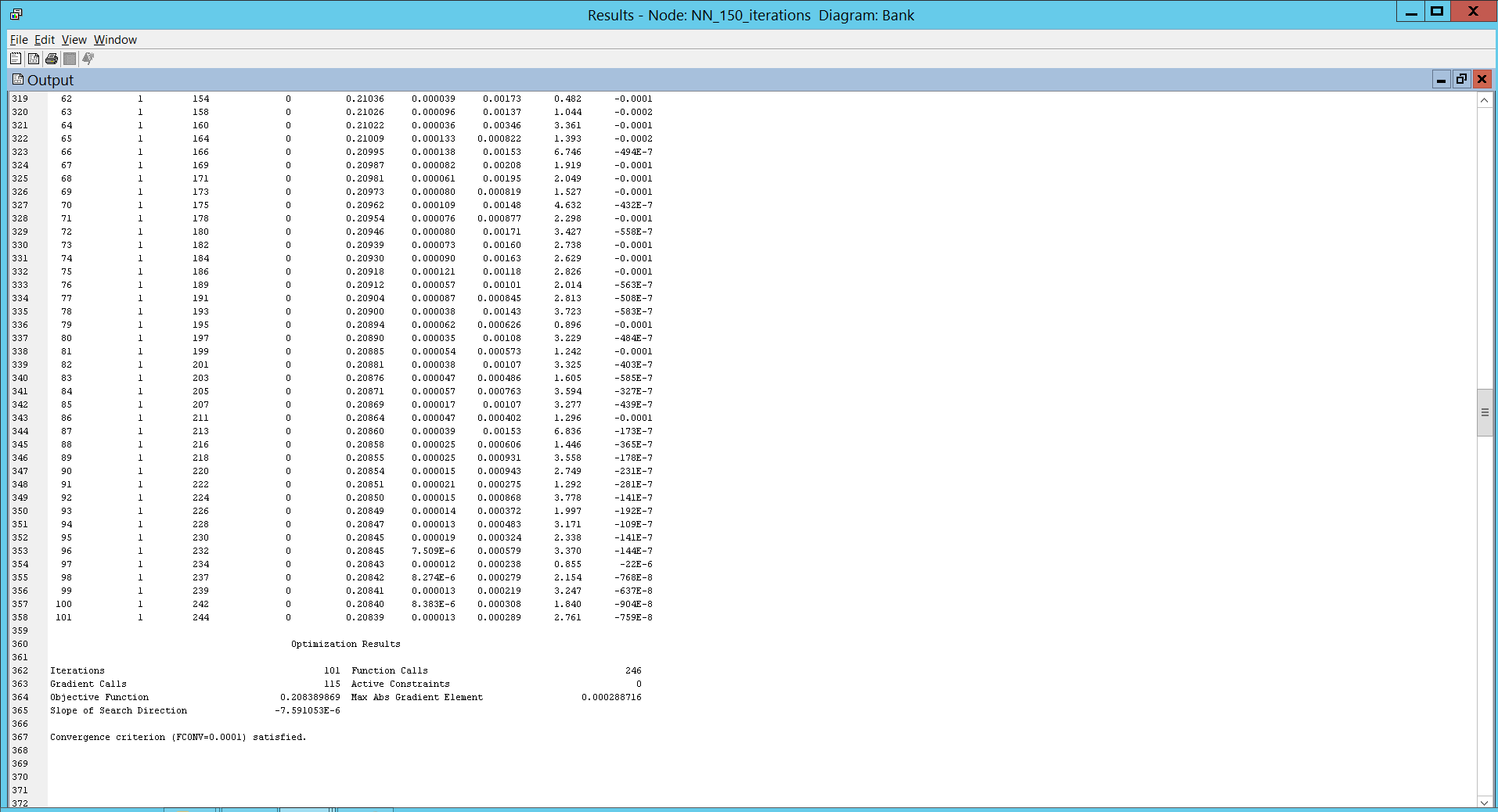


b. Neural **network analysis of the dataset from data partitioning changing number of nodes.**

Since the default network wasn’t meeting the convergence criteria. We have changed the number of nodes to 150 and have ran the model.



The Output of the networking criteria is as below. It is taking 101 iterations to meet the convergence criteria.

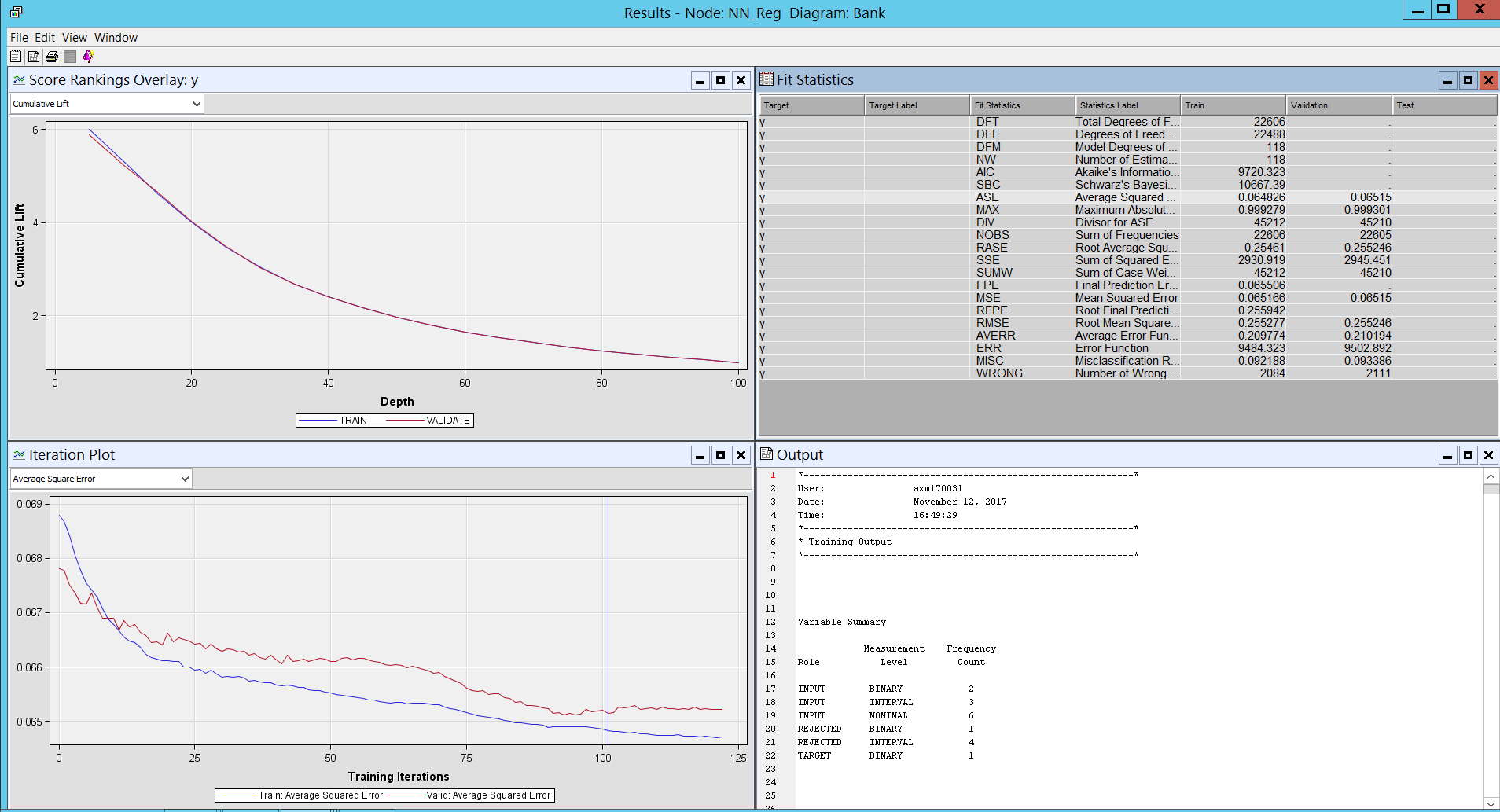


c. **The Neural network node is added to regression node and run**

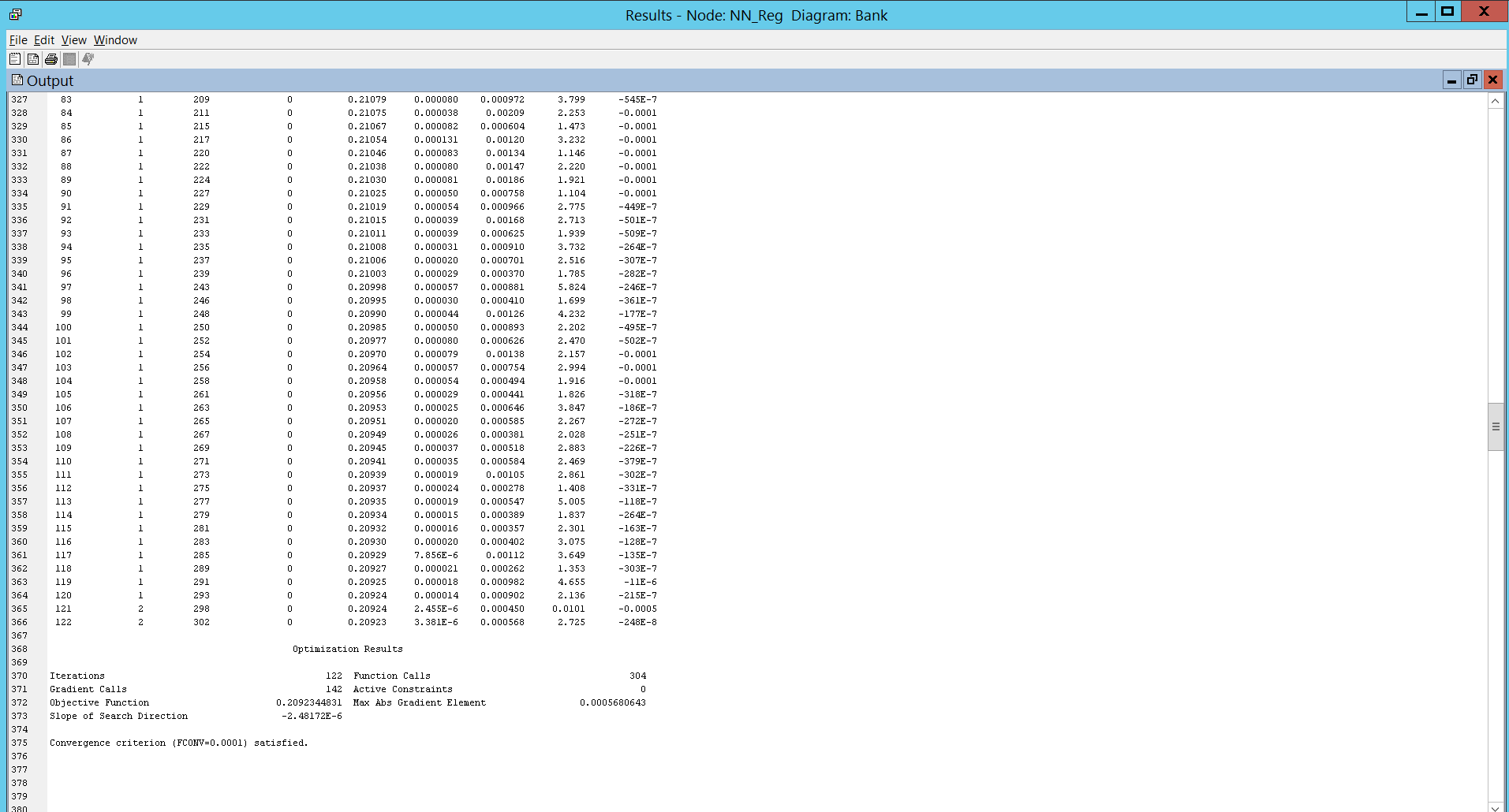
We used the output of Regression Model as an input to Neural Network as the output or regression has all statistically significant variables (stepwise regression) so the neural network will process only the significant and high valued variables in predicting the output rather than all available variables in the data set. This will make the result more accurate and will take lesser time and iterations as the number of variable decreases.

**The results are provided below.**

The Output of the neural network with regression is as below.

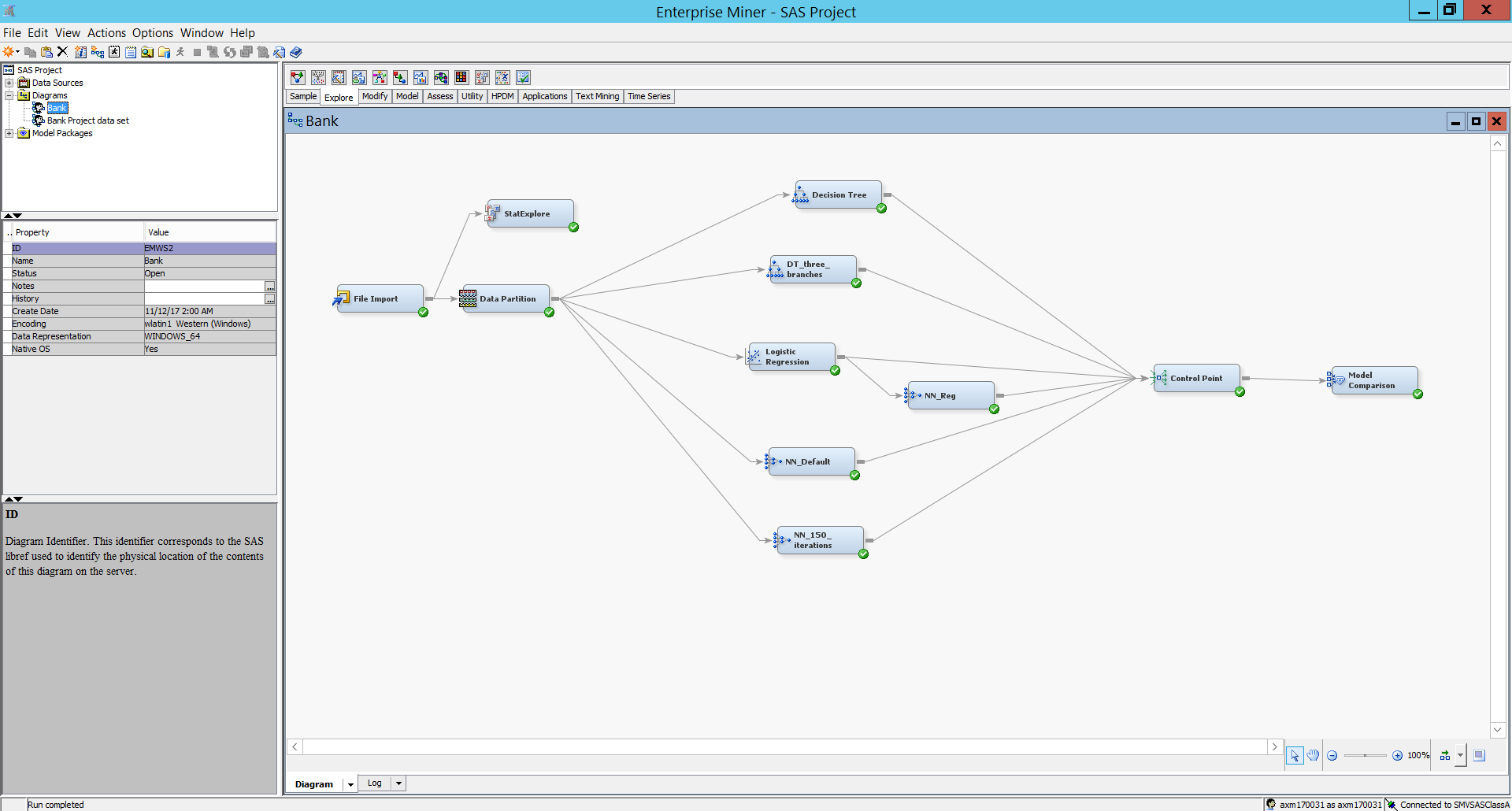


The convergence criteria were satisfied.

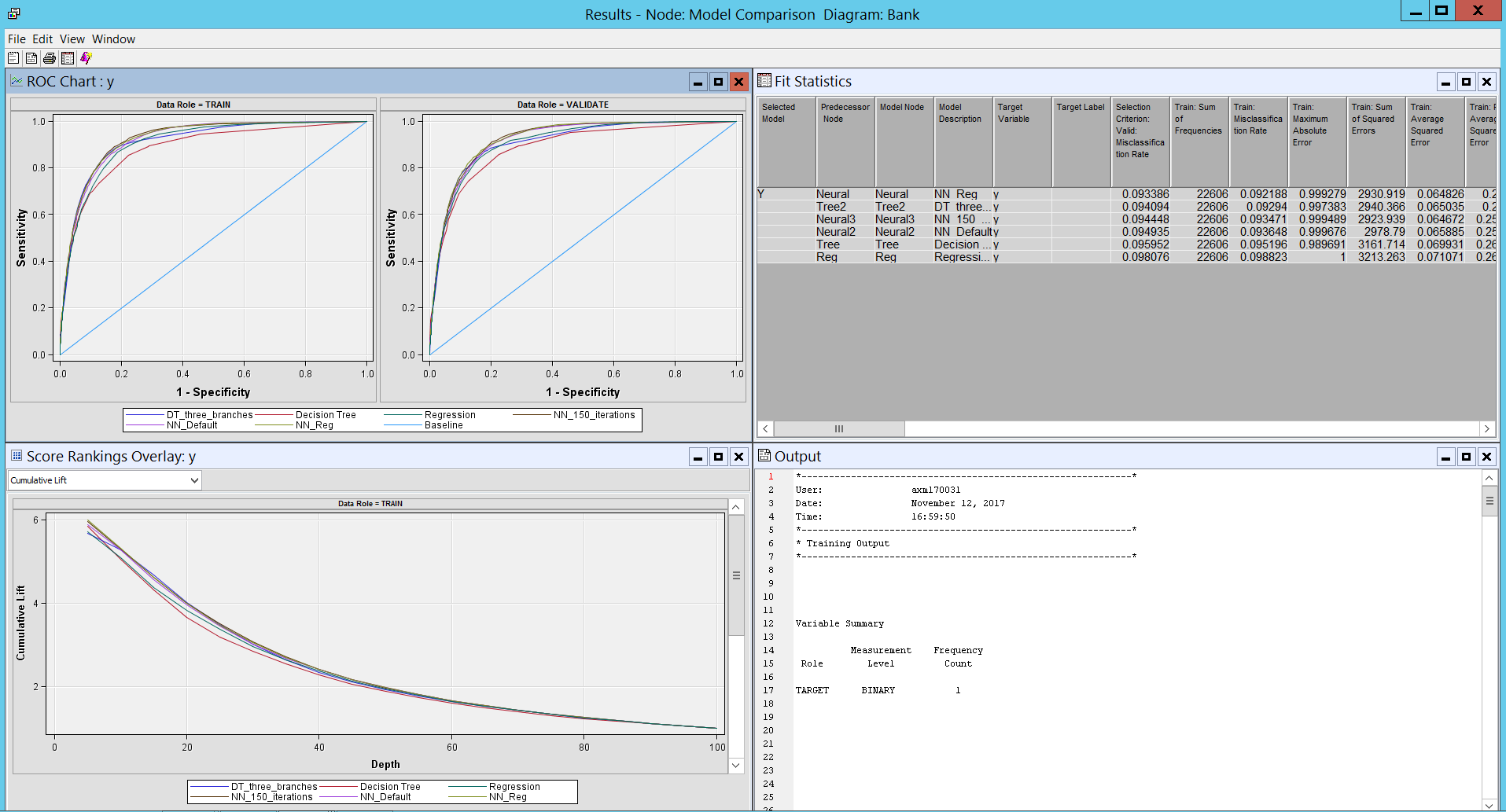


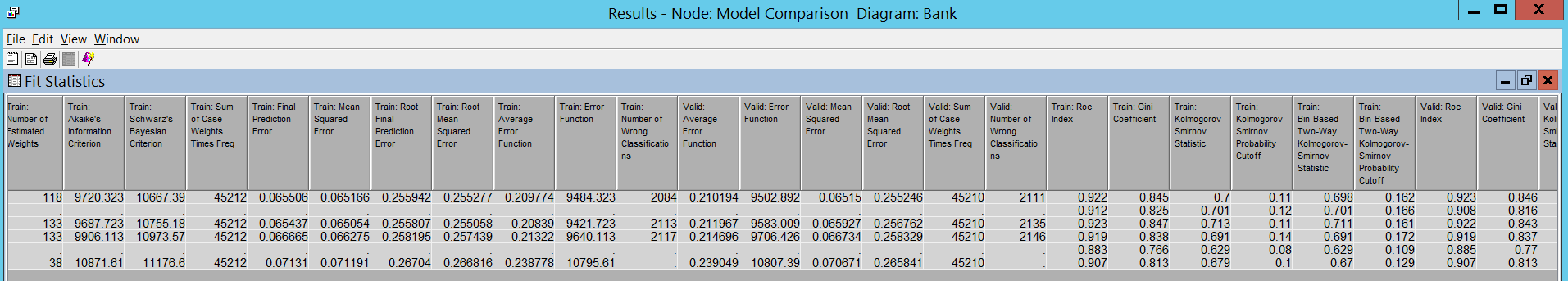
**Model Comparison:**

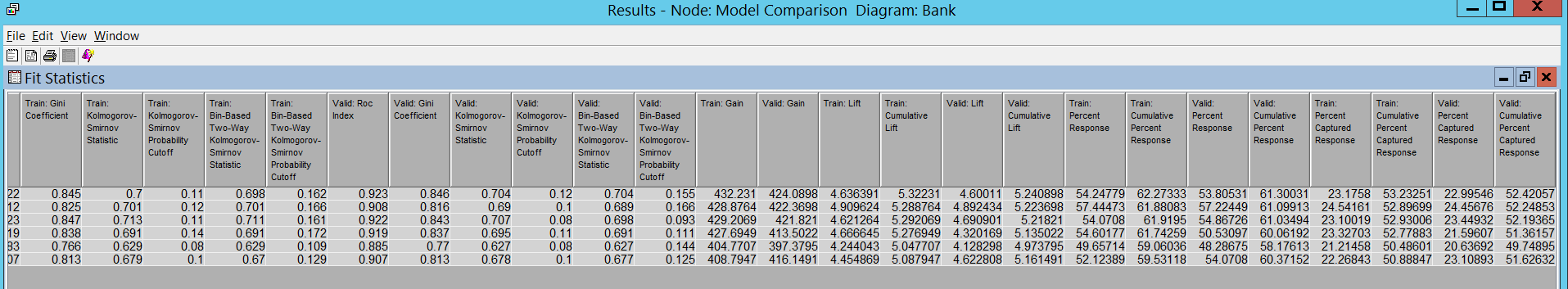
Below is the final diagram consisting of all the models we have used for our analysis. We have used control point to merge the output of all the models and used it as an input for Model Comparison.

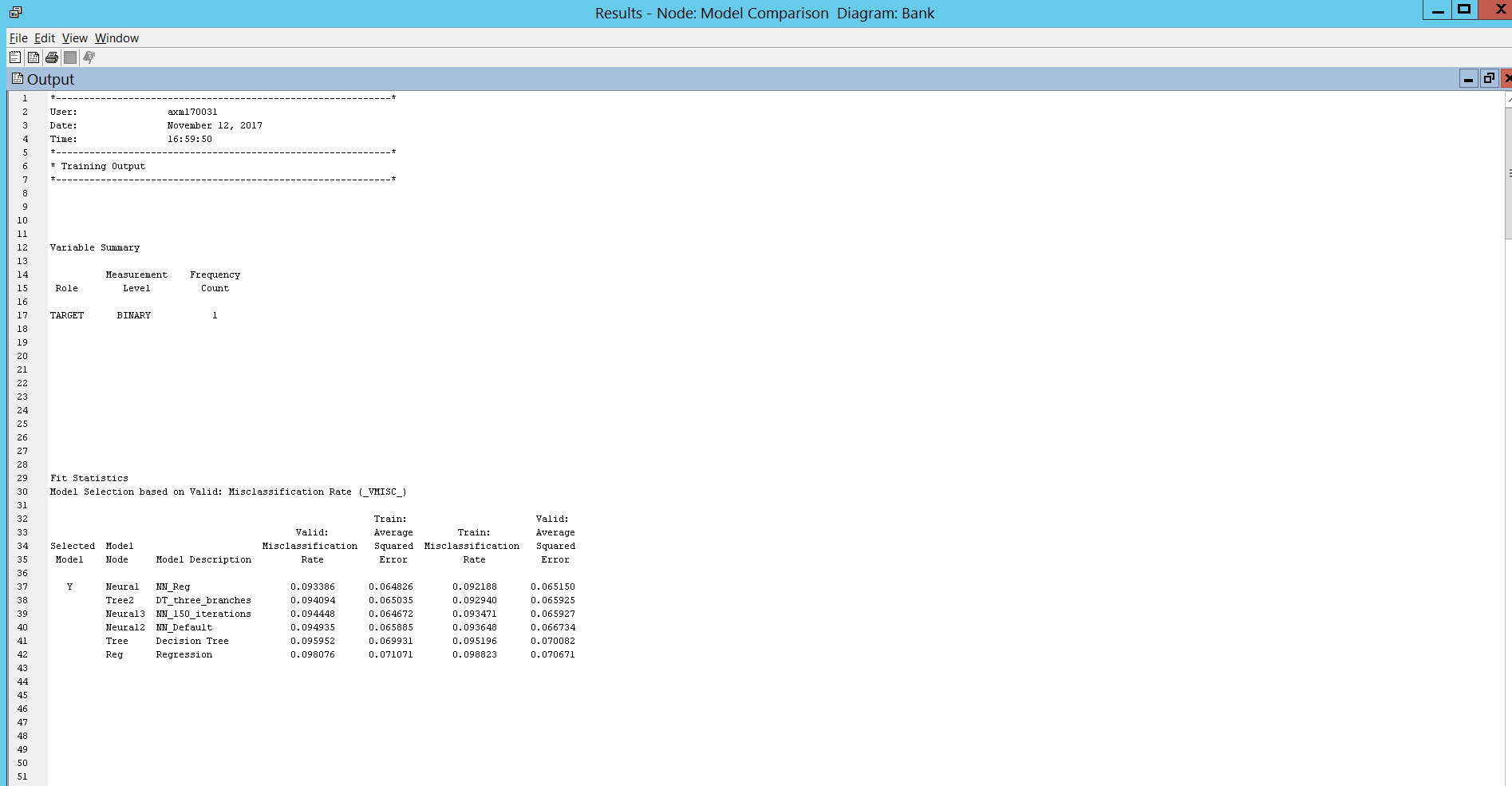


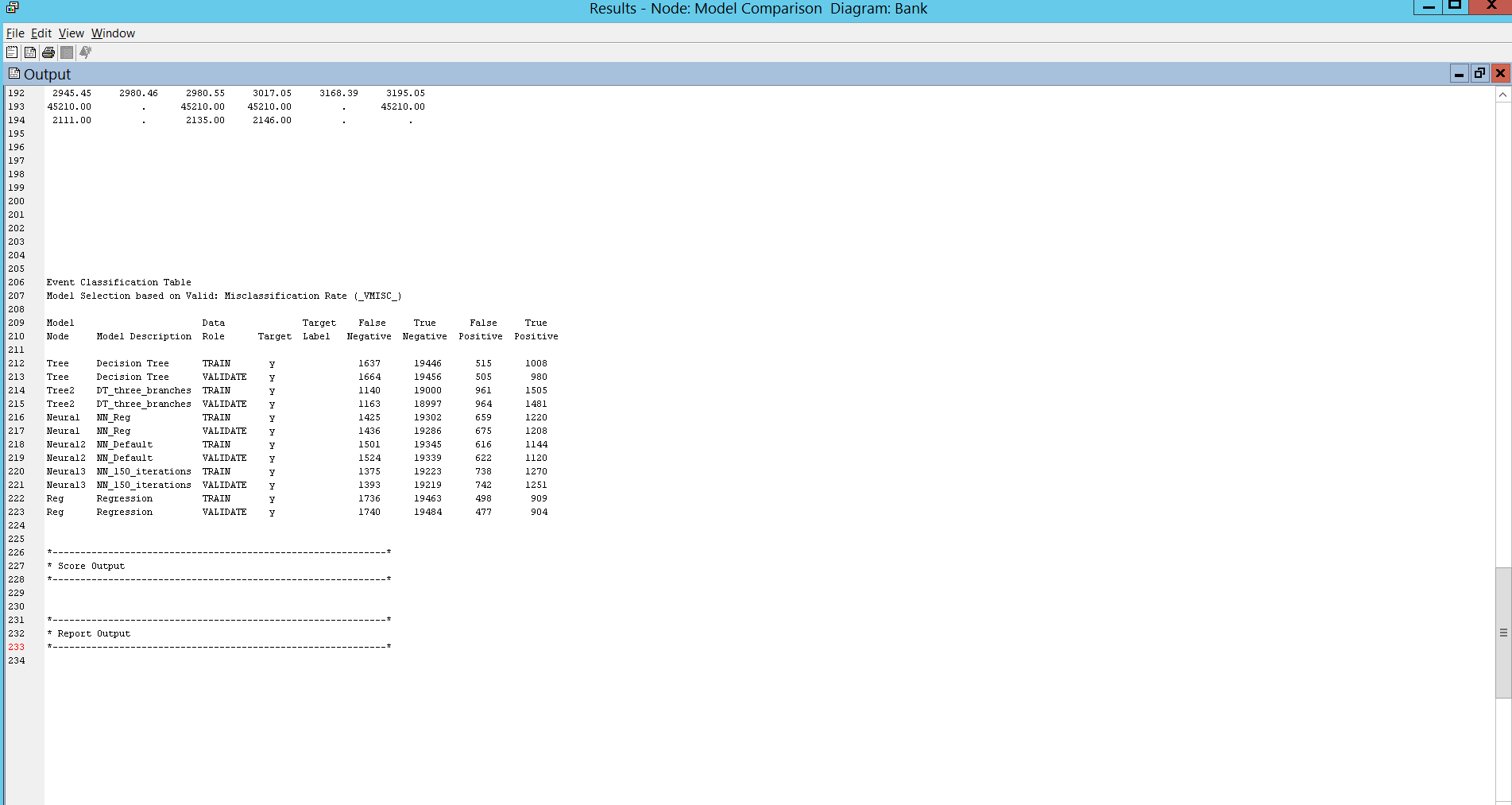
The output of the Model Comparison is as below.











We have kept the selection criteria as average square error for the Model comparison purpose. After checking the output of each node in model comparison, the neural network was selected as the best model as it has lowest average square error.

**Conclusion:**

The most influential variables in our model were duration, age and job. The recommendations provided are as follow.

* Duration: Reallocate resources from call centres to marketing research (new variables)
* Age: Focus your efforts on the 25 – 60 age group.
* Those that are active in the workforce.
* Job: Highest amount of sales went to people working in upper management

**References:**

Dataset:

* <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Data pre-processing, data modelling and comparison:

* <https://support.sas.com/en/support-home.html>
* Textbook related to SAS - AAEM02, AAEM03, AAEM04, AAEM05, AAEM06