

**Data Mining Assignment**

**Name:** Abhishek Dubey

**Student ID:** D20123718

**Full time:** MSc Data Science 2020-21

**Class:** TU59 Full Time

**Date:** 15/12/2020

**Definition of problem**

Telemarketing campaign data of Portuguese bank institution. By using it we need to identify the customers who are likely to subscribe the term deposit account based on previous marketing campaign.

Here in this project, we will use several machine learning algorithms to identify the best suit for this analysis.

Further that final model will consider as credible and valuable for telemarketing campaign managers.

Final Goal is to predict weather customer will opt term deposit- variable “y” (yes, no)

This will result in selling more term deposit account by Portuguese Bank.

**Data Exploration and Descriptive Analytics**

**\*\* We are using SAS Enterprise Miner for Analysis**

**Import Data:**

Data belongs to marking campaign of Portuguese bank.

We are using SAS Enterprise Miner for Analysis

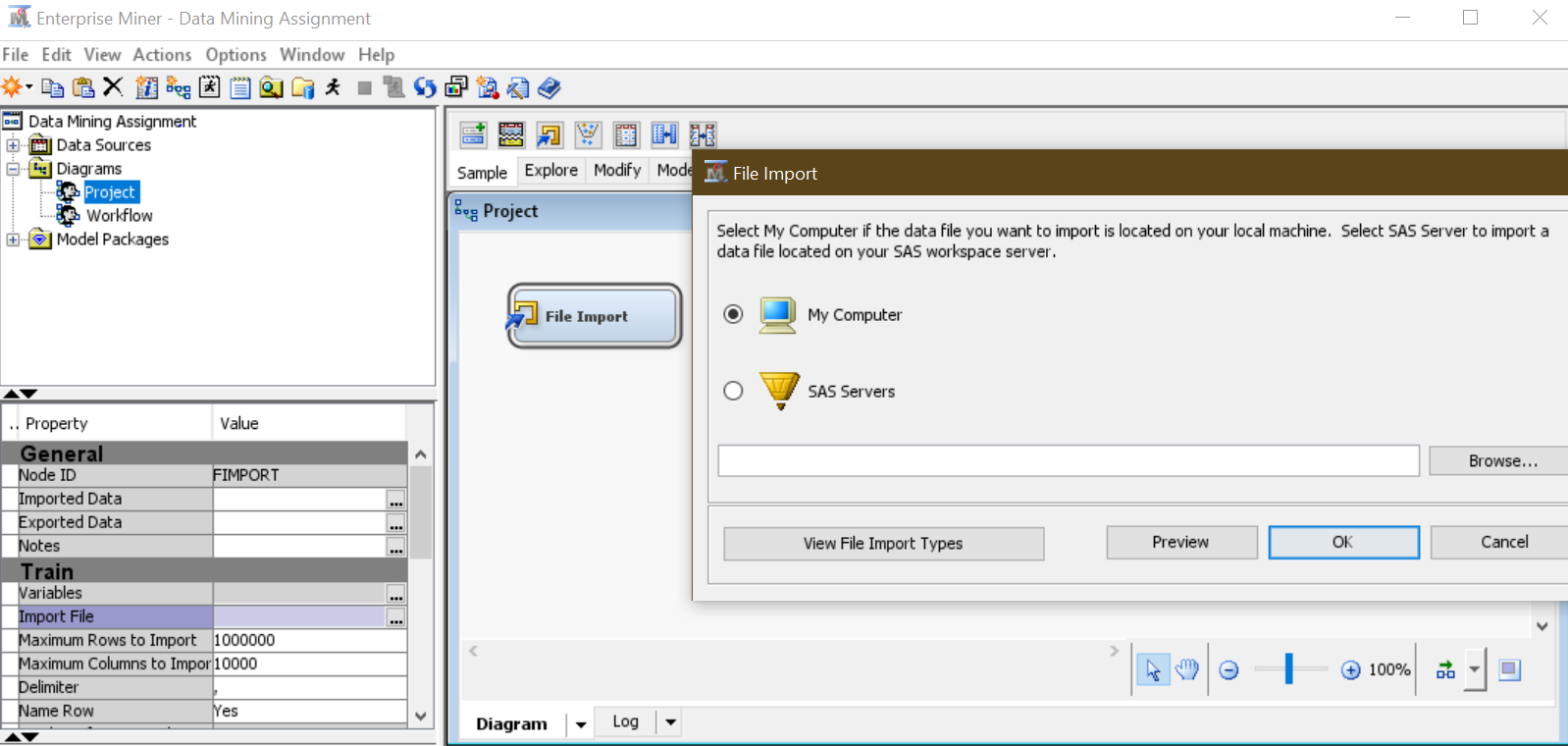
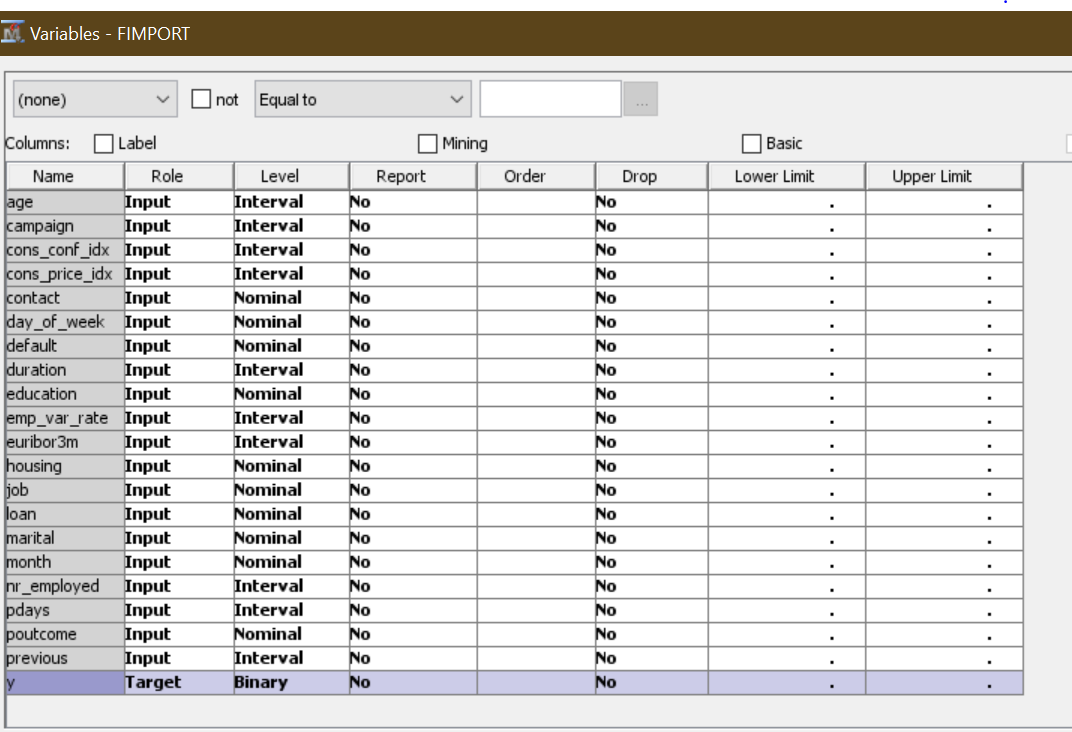


Figure: Represents how we import data in SAS

Data set contains 41188 rows (observation) and 21 columns (variables)

All Variables in Data Set



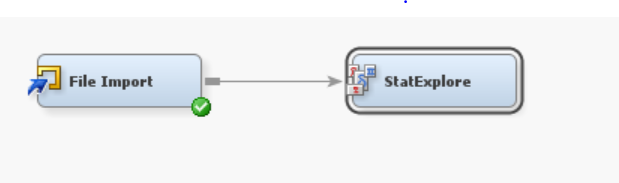
Y: Representing Target variable, Binary – yes, no

YES: represents term deposit account he/she will open

NO: represents term deposit account he/she will not open

We will now explore the data so we use **StatExplore** function from explore tab

When evaluating chi squared statistics for interval variable. Enterprise miner divide them into 5 bins so allowing interval variables to “yes”.



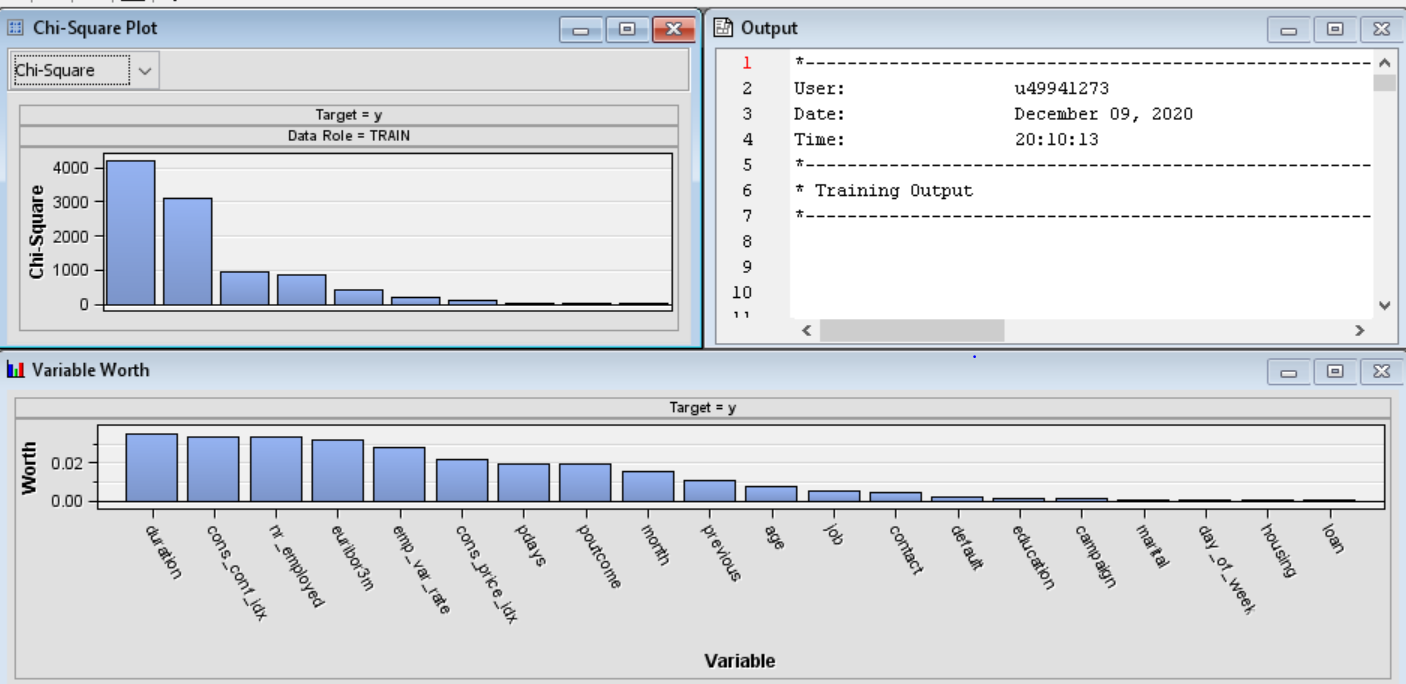
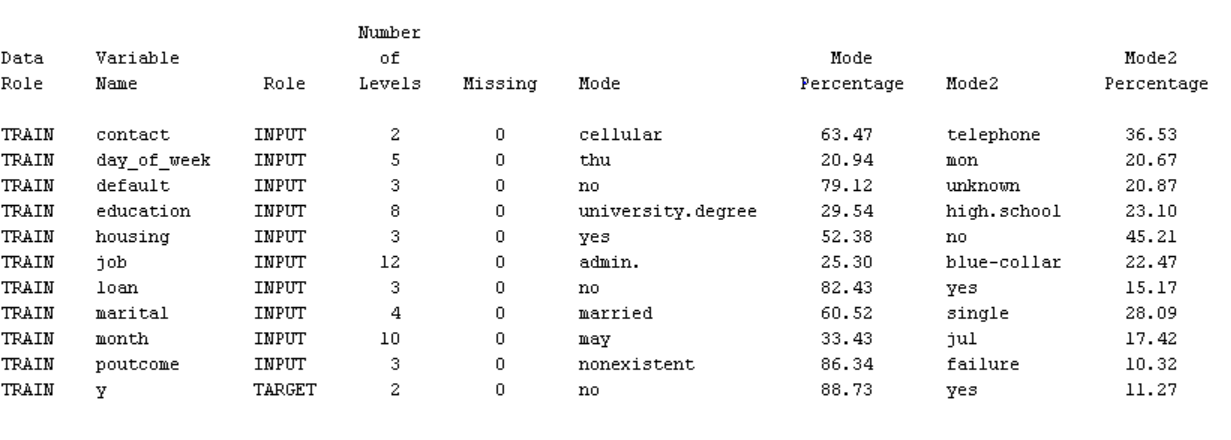
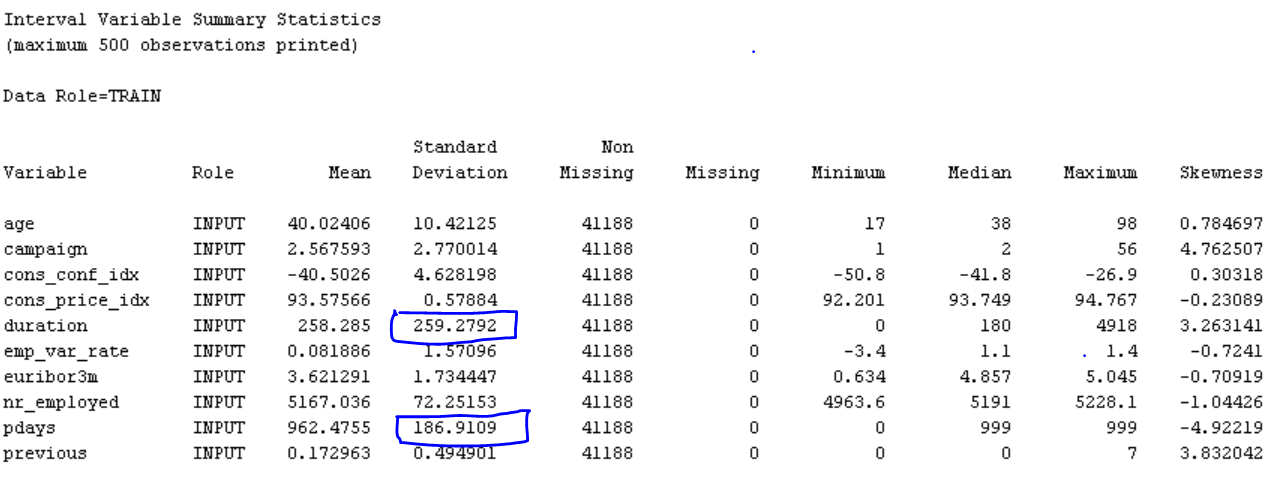


Figure: Represents the Chi-Square as per target variable y

**Class Variable Summary Statistics**, as we can check there are no missing values in dataset



**Interval Variable Summary Statistics**



As we can check there is **no missing values** present in data set.

Standard Deviation shows high more than 100 for values **duration and pdays**

Duration: duration of call-in seconds, this directly affect the variable Y. If duration is 0 then Y (target variable is also 0)

Pdays: number of days there was no contact after previous campaign. Generally, dataset contain pdays for last 30 days and if person not contacted before than data value is 999.

Which is the reason why mean of pdays are very high.

**In transformation of data, we will explore and reduce the variance in those variables with large deviation.**

**Identification of data insights from previous step**

To explore graphs in SAS Enterprise miner I am using graph explore function for analysis.

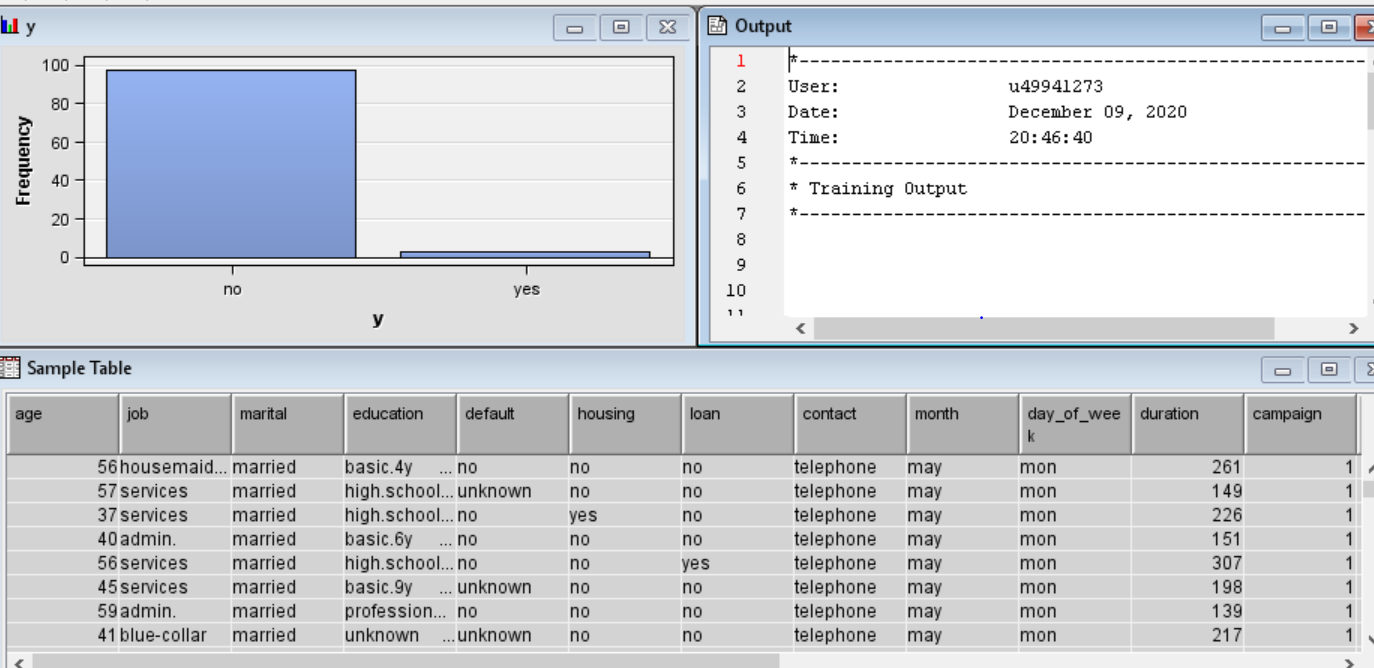
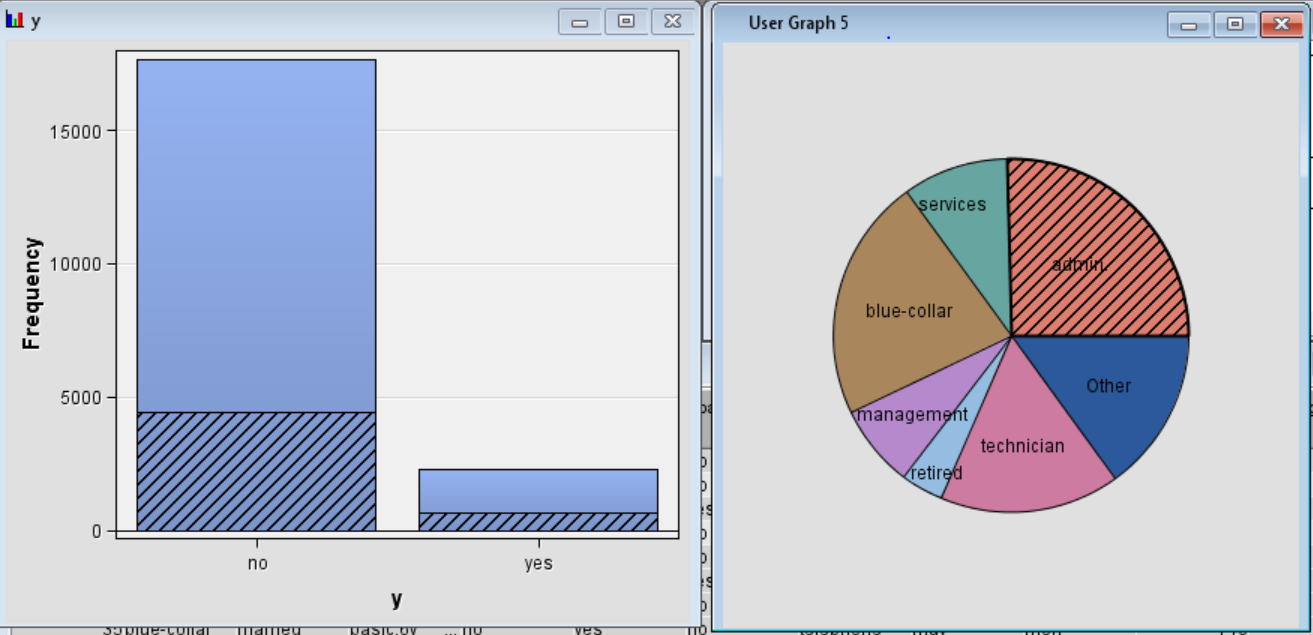


Figure: Represents the frequency of y in 100 random observations

On left: Below Graph represents the frequency in bar chart of 20000 random observations of target “y” variable

On right: Below graph represents the pie chart of 20000 random observations of job variable.

**Below insight shows that most of the admin job and most of blue-collar jobs went for “No” means don’t want to opt for term deposit, however the admin and blue-collar observations are more in number comparable to other in variable job.**



**Small part in Python for better cross analysis**

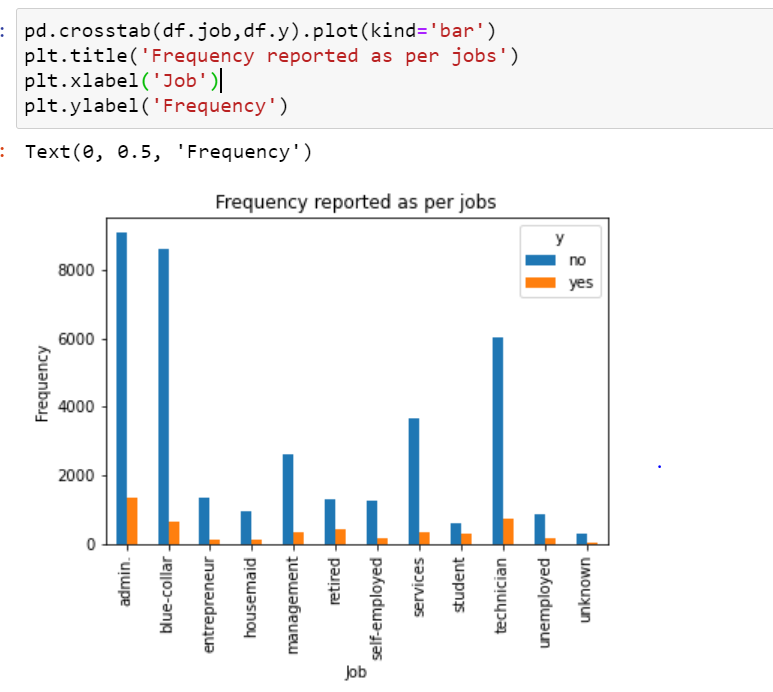


Figure: Represents the frequency of response in y variable as per jobs.

**Details of any additional data preparation (cleaning, transformations, etc), data enrichment, feature engineering, feature reduction, etc**

**\*\* This Step is taken under Python small part and rest done in SAS**

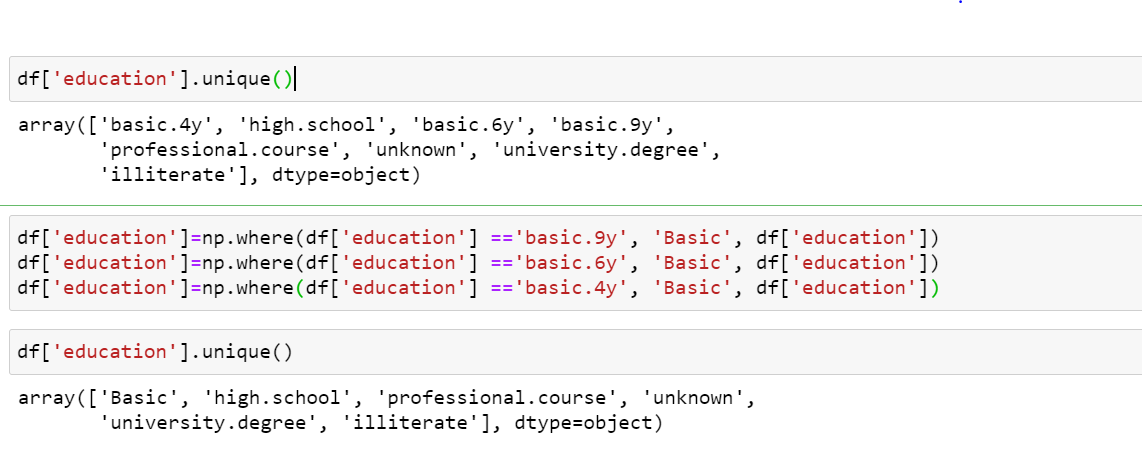
**Data Set description**



As per analysis we can see that education variable has 8 unique observations.

So, it’s better to group all basic studies together

So, for that I am grouping basic.4y, basic.6y, basic.9y to “basic”



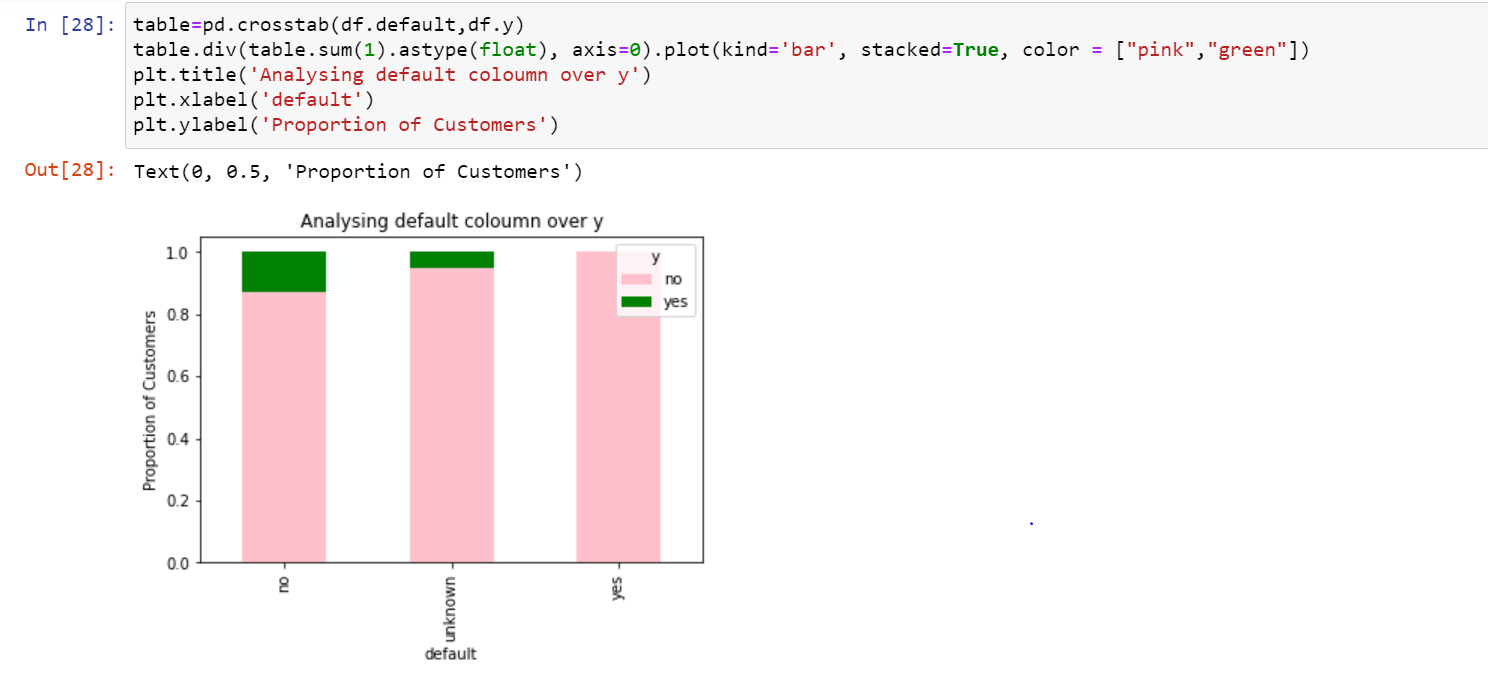
After grouping now, we have only 6 unique observations which are good for modelling

Other Non-Parametric variable is “default”

Variable contains so many unknown values which are of no use for our predicting modelling of decision tree.

As per analysis of proportion for customers in default, “unknown” values are more “no” observations of “y” variable.

Most important thing is “yes” observation of “default” variable shows 100% values of “no” observation of variable “y”. **Which means people have credit in default not going for further term deposit account in bank**



So, we are removing that variable

df1 = df.drop(['default'], axis=1)

df1.to\_csv("bank.csv")

**Checking value counts for target variable “y”**

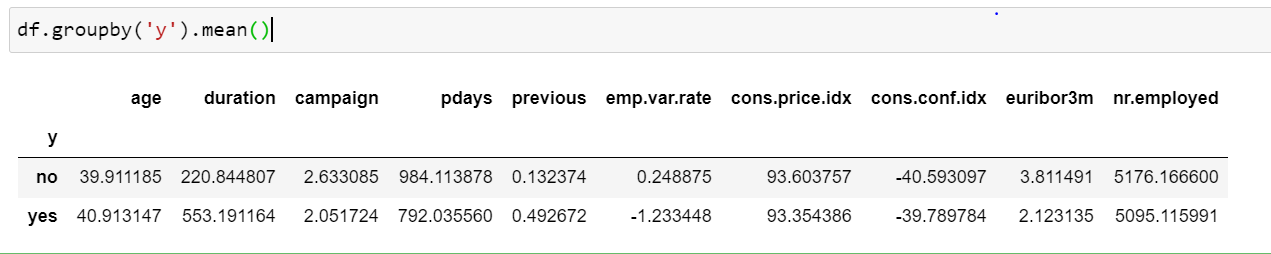
**df["y"].value\_counts()**

no 36548

yes 4640

Name: y, dtype: int64

**Means of parametric variables in dataset with respect to “y”**



**Data Partition**

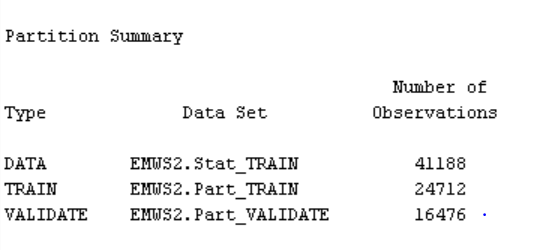
**Training data is used for model fitting**

**Validation Data is used to test the model without overfitting the data**

**Dataset is divided into 60%,40% percent randomly**

**60% Training Data**

**40% Validation Data**

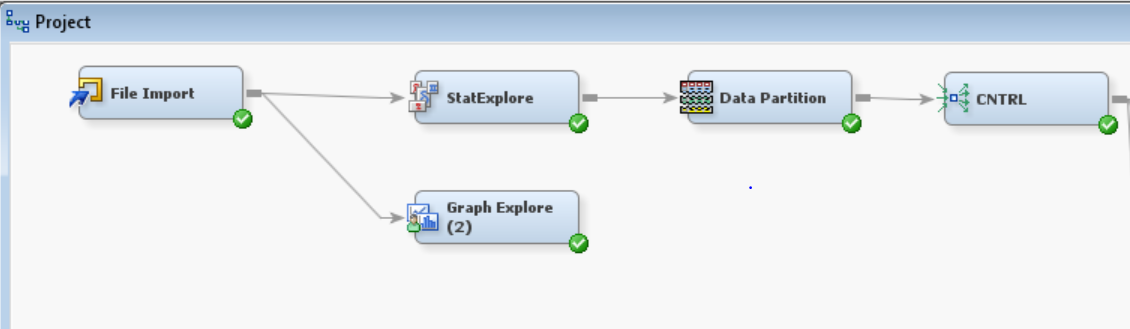


**Control Point Node:**

**It simplifies a process flow of our diagram by minimizing the connections between multiple data sources and multiple flows.**

**It works like a connector between all mining models and the data source.**

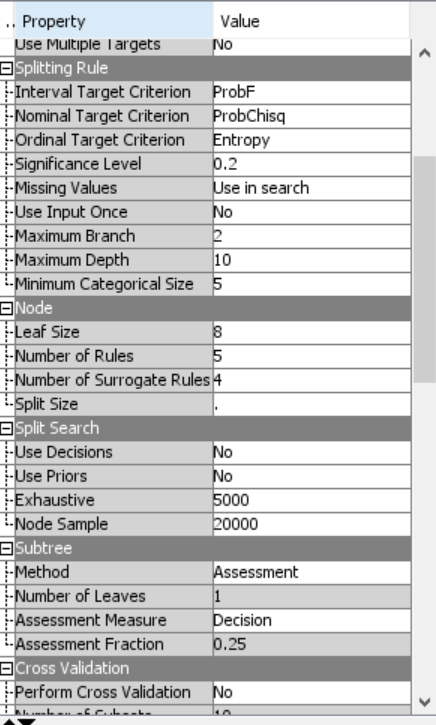
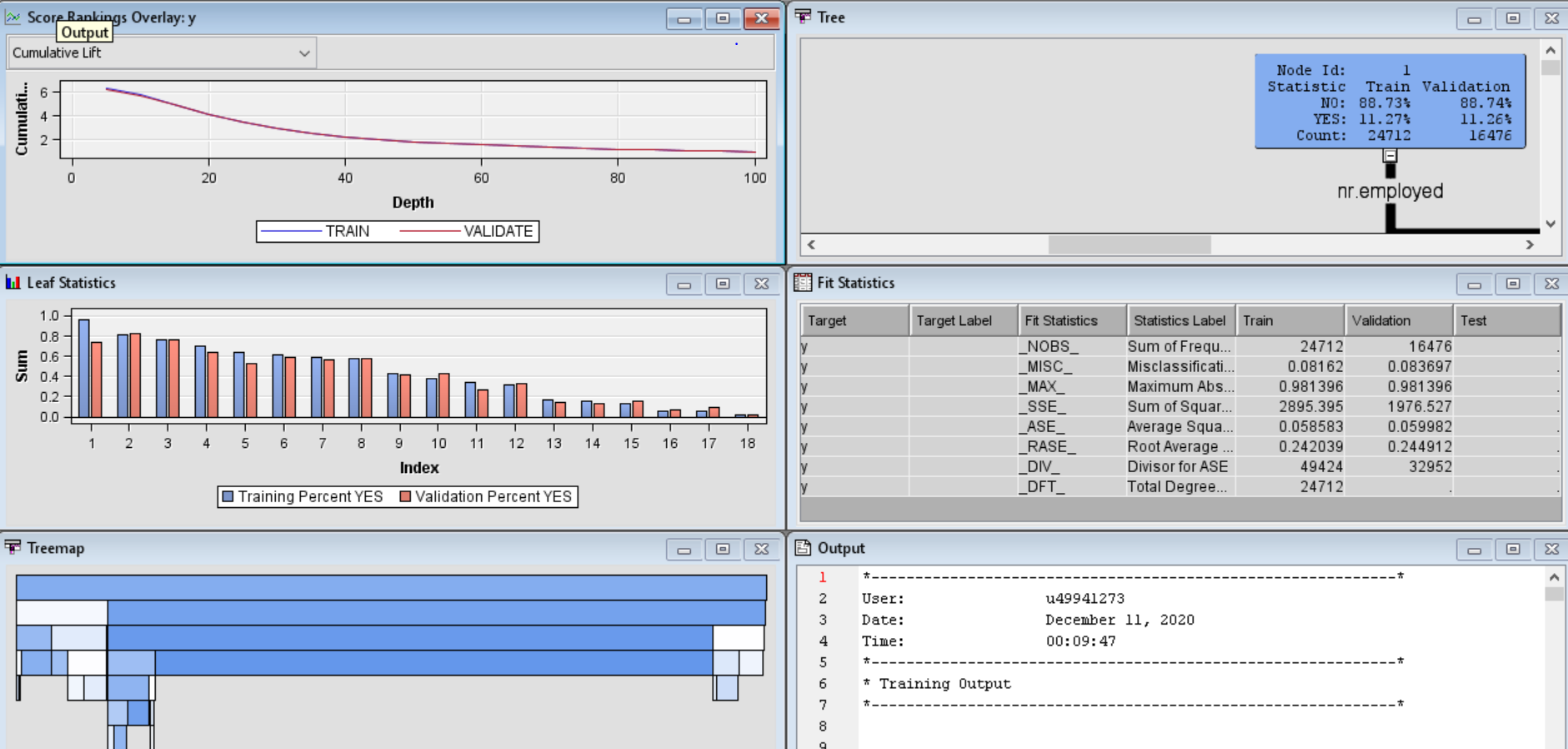
**Making manual connections are very difficult to read the flow that’s why we are using control point.**



**Developing Models**

**Decision Tree**

**Automotive Decision Tree (Self Pruning)**

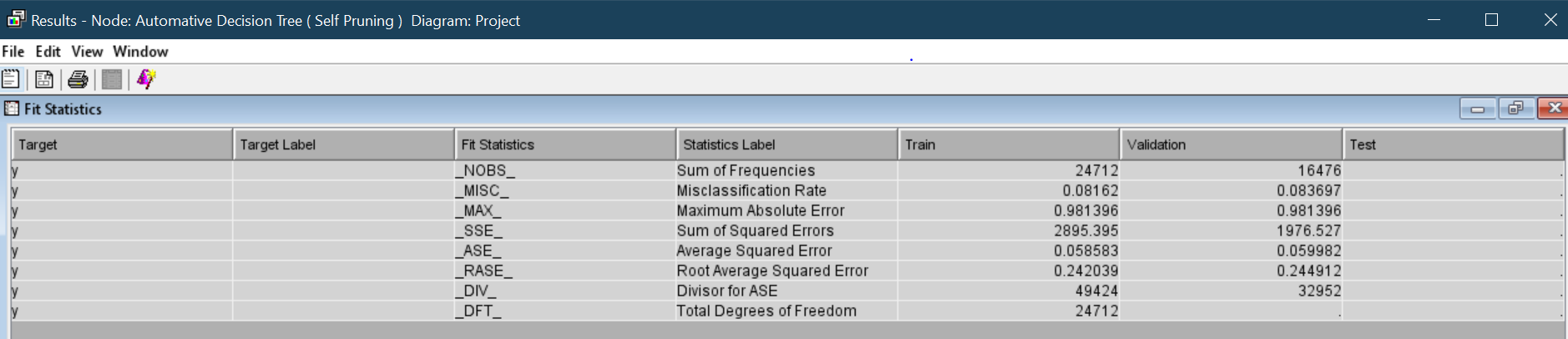
For above properties I have set maximum depth to 10 so that tree will go to 10 depths by split rules.

Node Size is set to 8 so that it should be minimum size of terminal node.

**We used automatically trained decision tree by using split rules that maximize the logworth value**

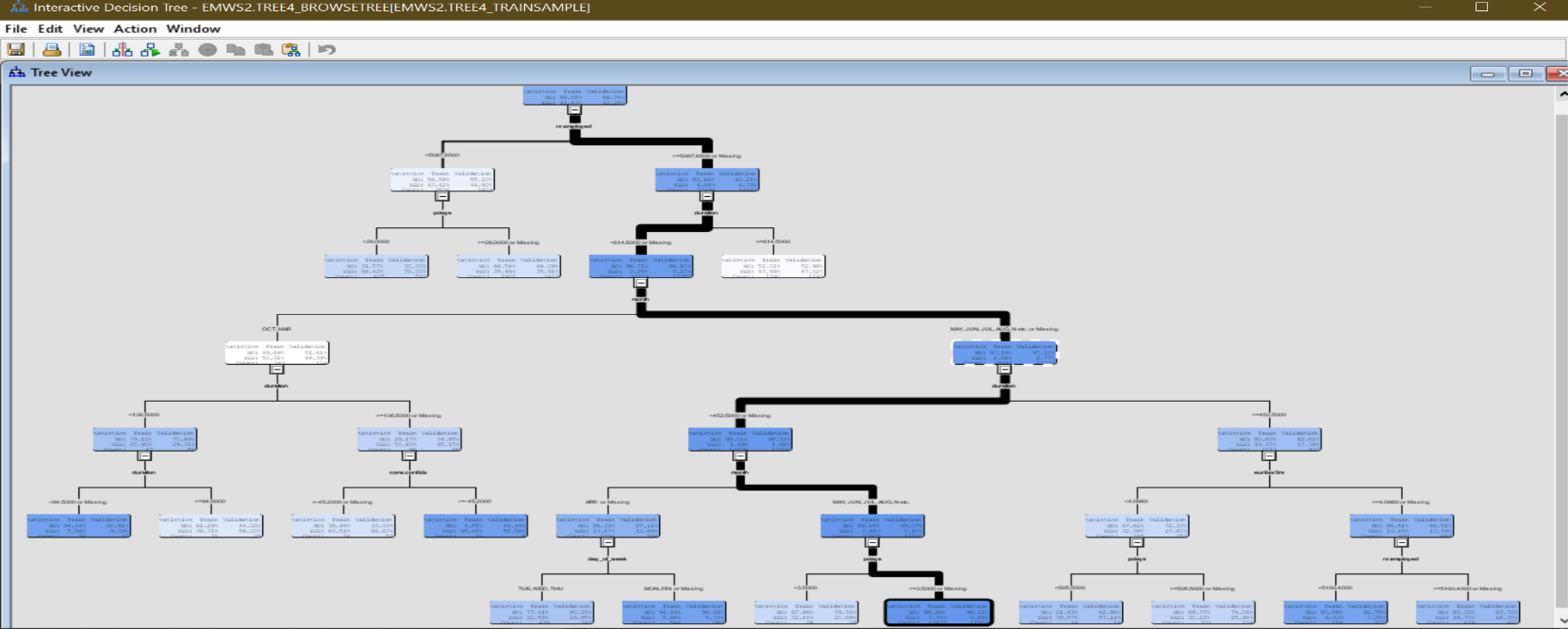
**Fit Statistics for Decision Tree.**

**As you can see root average squared error is 0.24 and Maximum Absolute Error is 0.98**

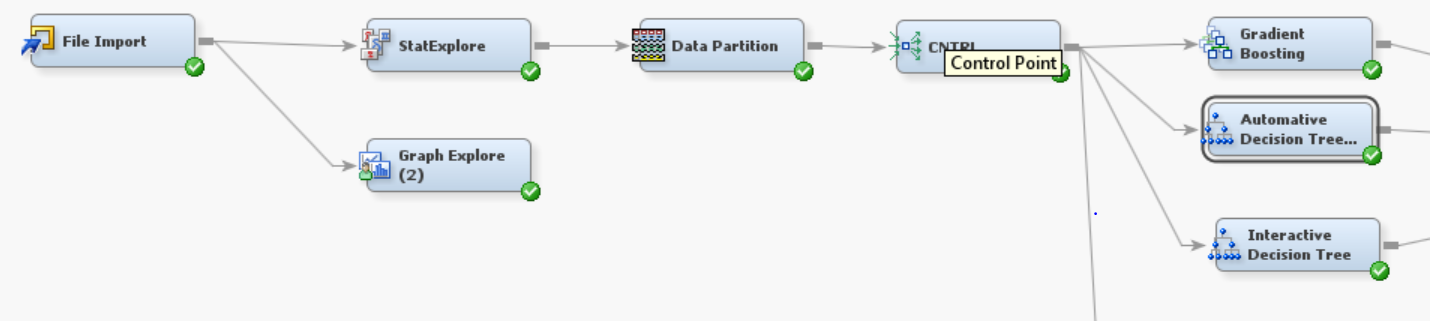


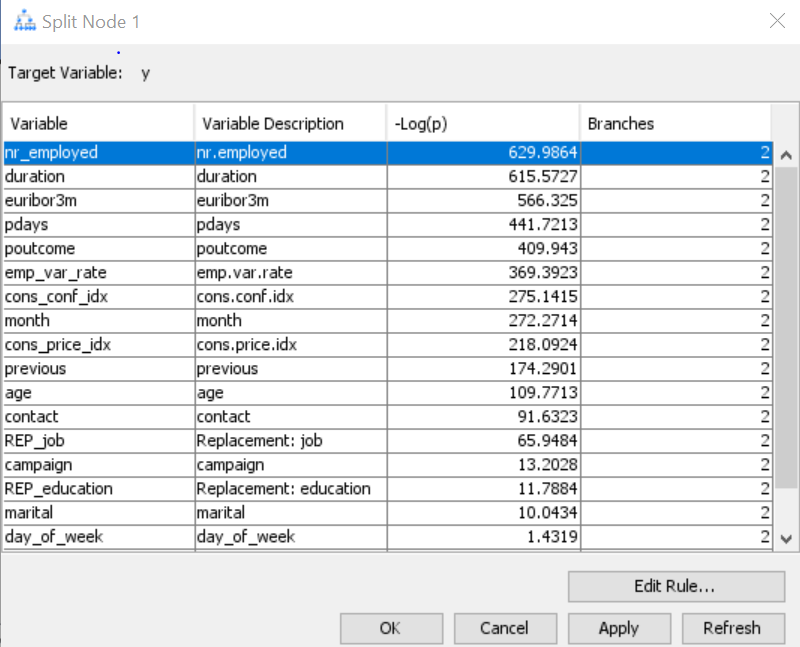
**Interactive decision tree**

**Now we make new interactive decision tree by selecting best candidate split rules by watching out logworth and splitting size from some variables.**



**Now we make new interactive decision tree by selecting best candidate split rules by watching out Log worth and splitting size from some variables.**





**Pdays : number of days past since last contact in previous campaign**

And 999 represents that customer never contacted and because of this value the standard deviation is high and mean is also high.

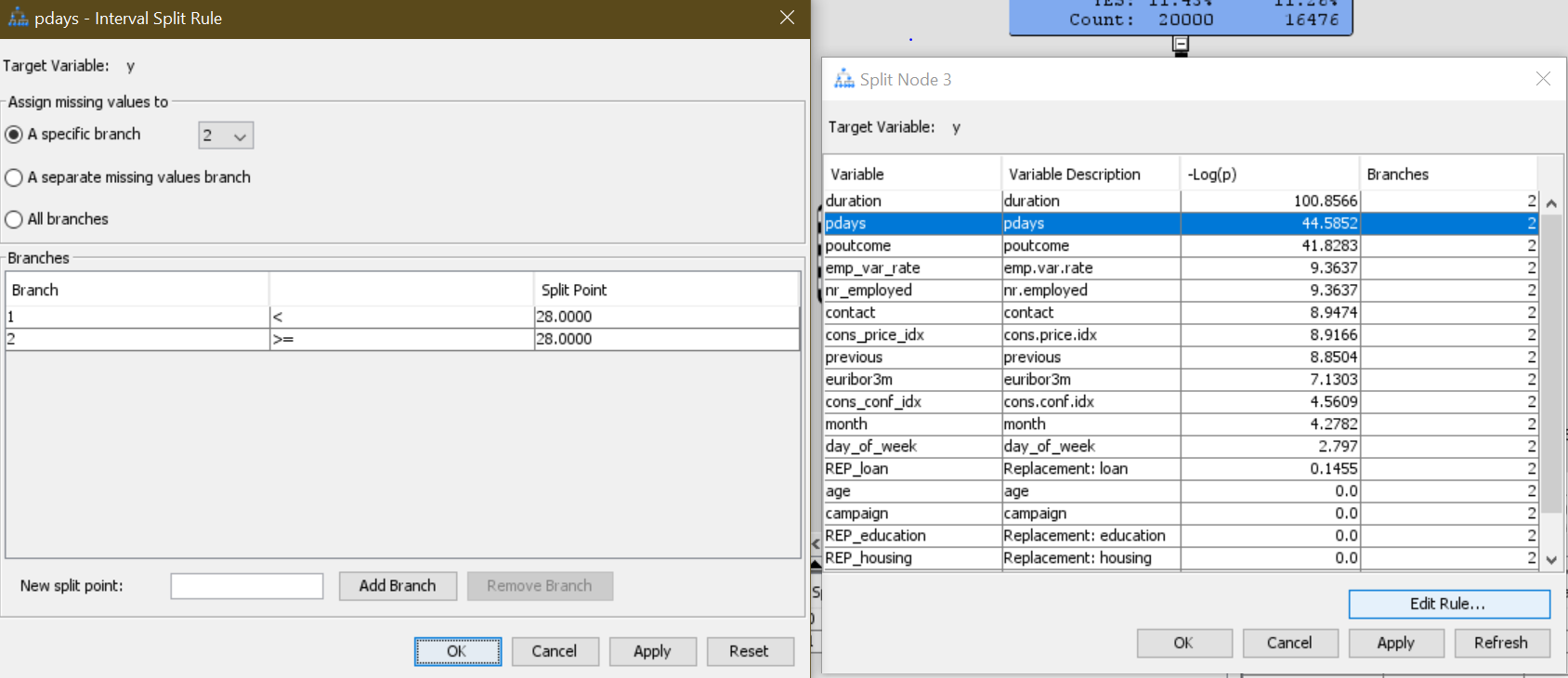
We need to rectify this problem by removing it.

We will use split point as 28

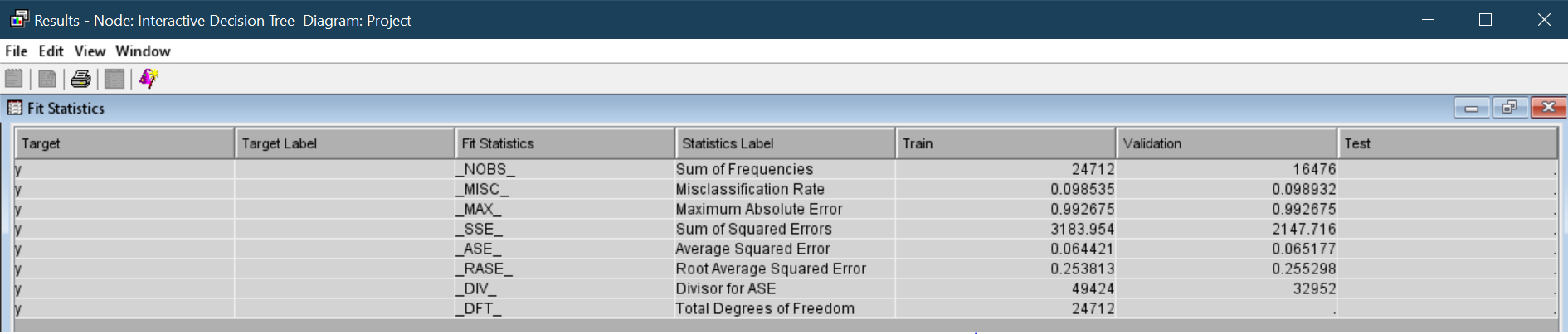
<28 means customers are contacted in last 30 days

>28 means they never contacted

Pdays has values from 0 to 27 and 999



**Fit Statistics for Interactive Decision Tree**

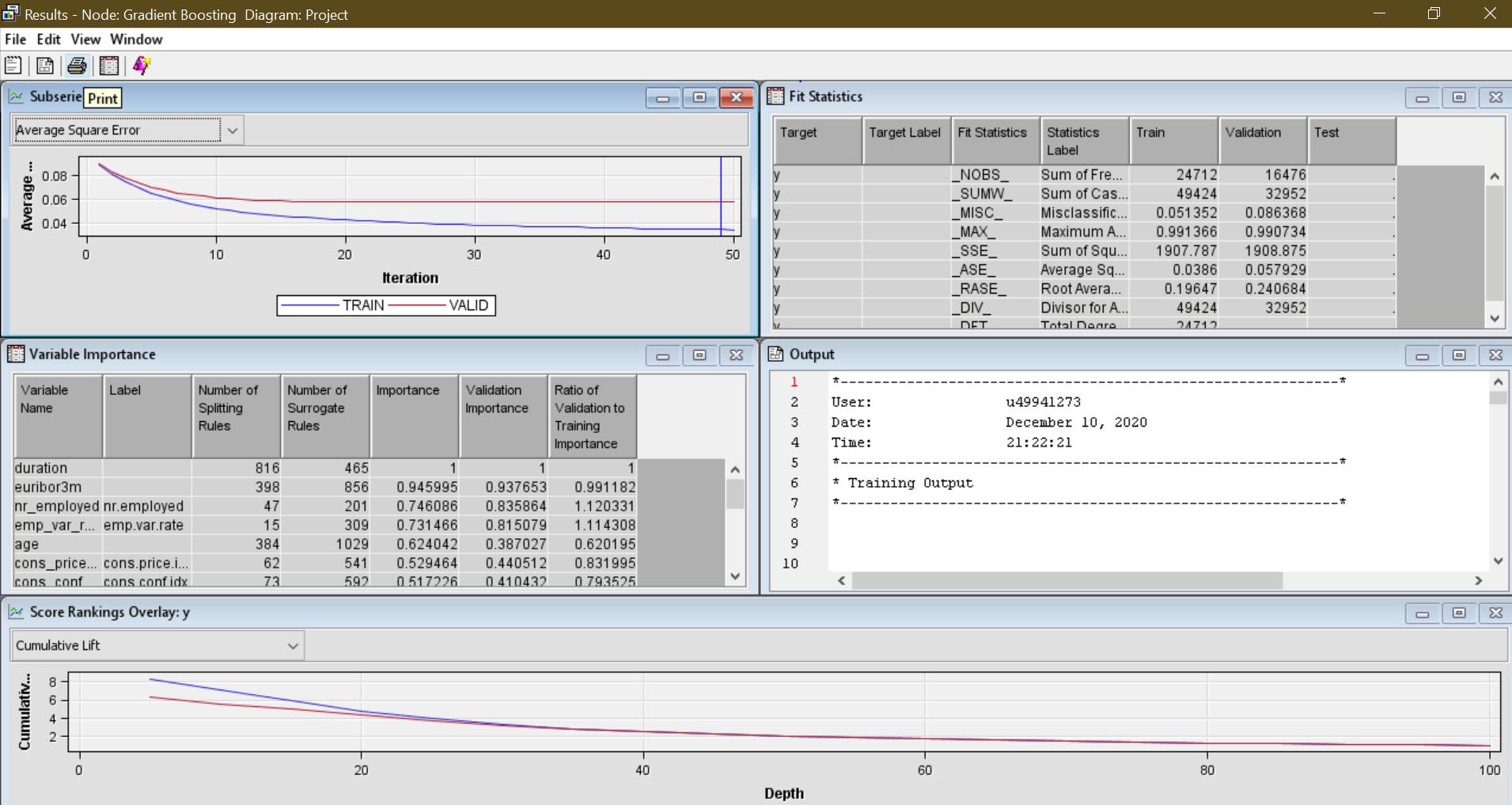


**As you can see root average squared error is 0.25 and Maximum Absolute Error is 0.99**

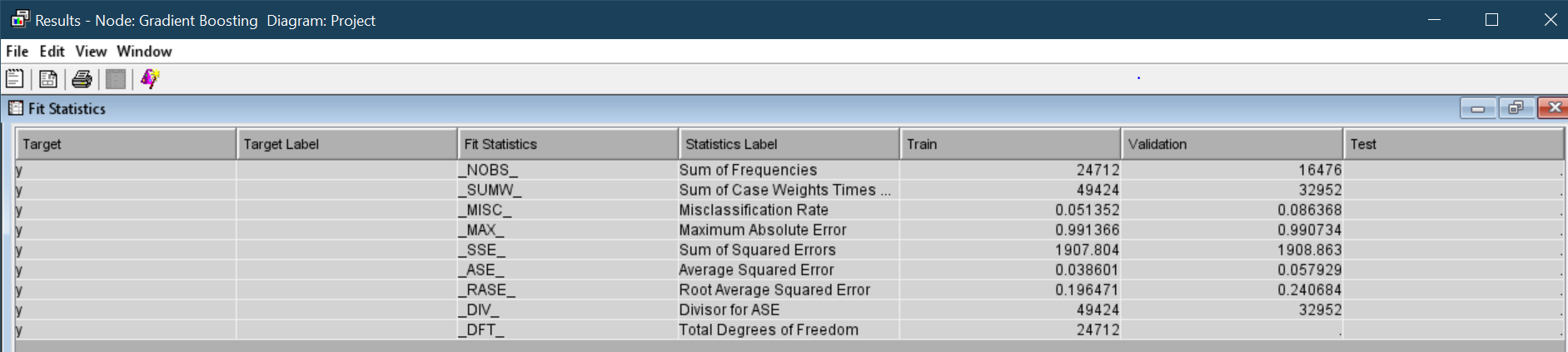
**Average Squared Error is 0.06**

**Gradient Boosting**

**We further added gradient boosting to our decision trees to enhance the efficiency and fitting of Decision Tree.**



**Fit Statistics for Gradient boosting**



**As you can see root average squared error is 0.196 and Maximum Absolute Error is 0.99**

**Average Squared Error is 0.03**

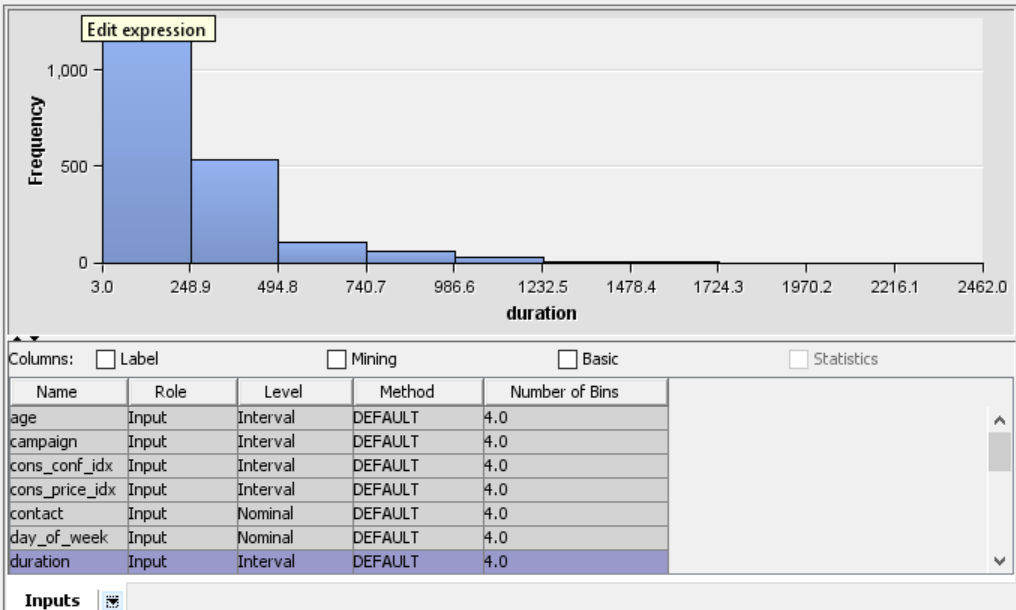
**So as per analysis of all three models Gradient Booting shows much better efficiency in terms of average squared error, root average squared error, maximum absolute error.**

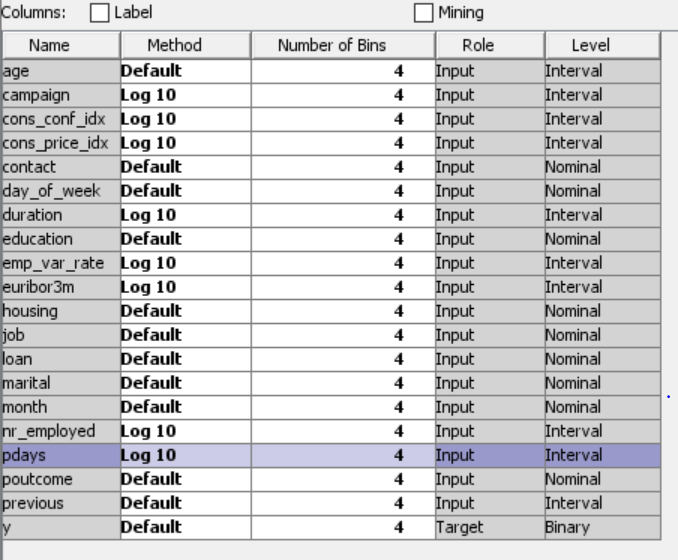
**Now we will use parametric data for predictive model.**

**Logistic Regression**

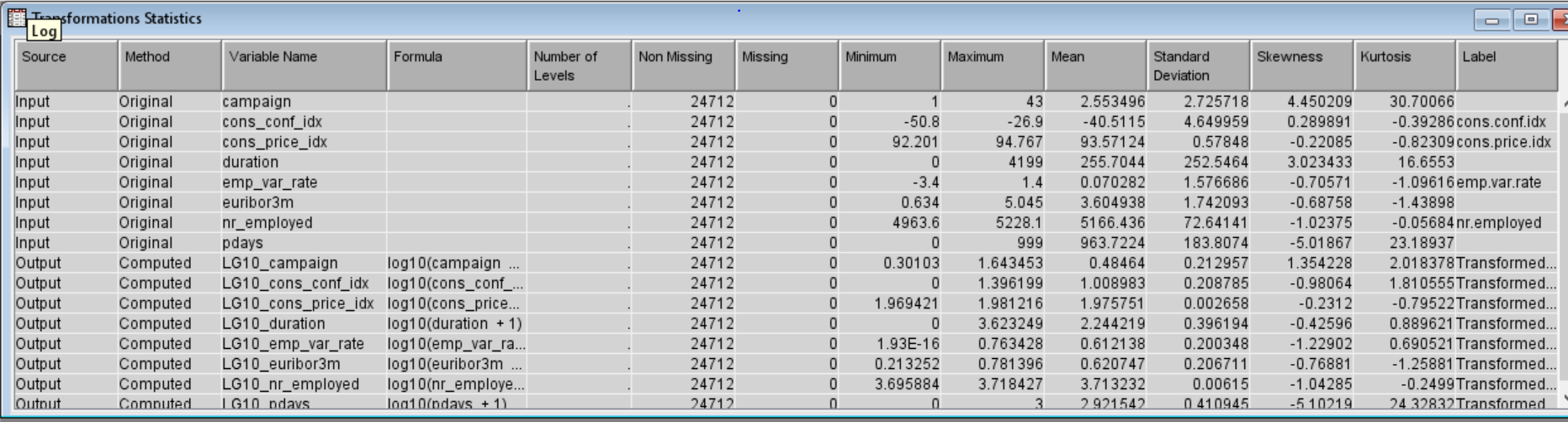
Transforming data can improve the model accuracy. It will stabilize the variance, remove non linearity, check skewness in data.

Transforming input data leads to better fit the model.

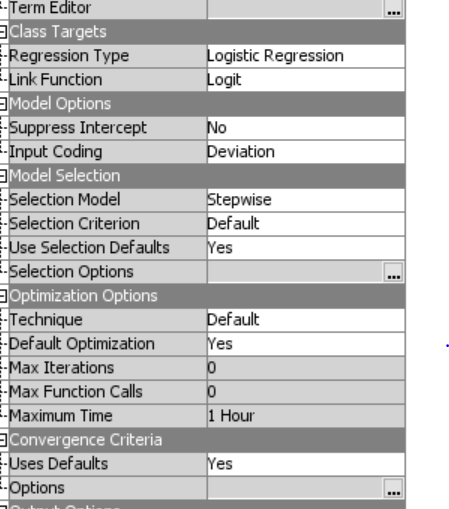
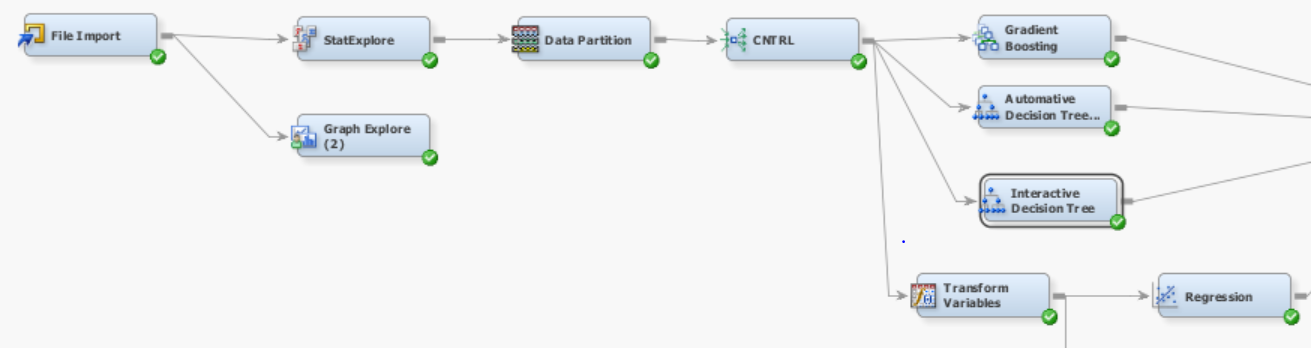




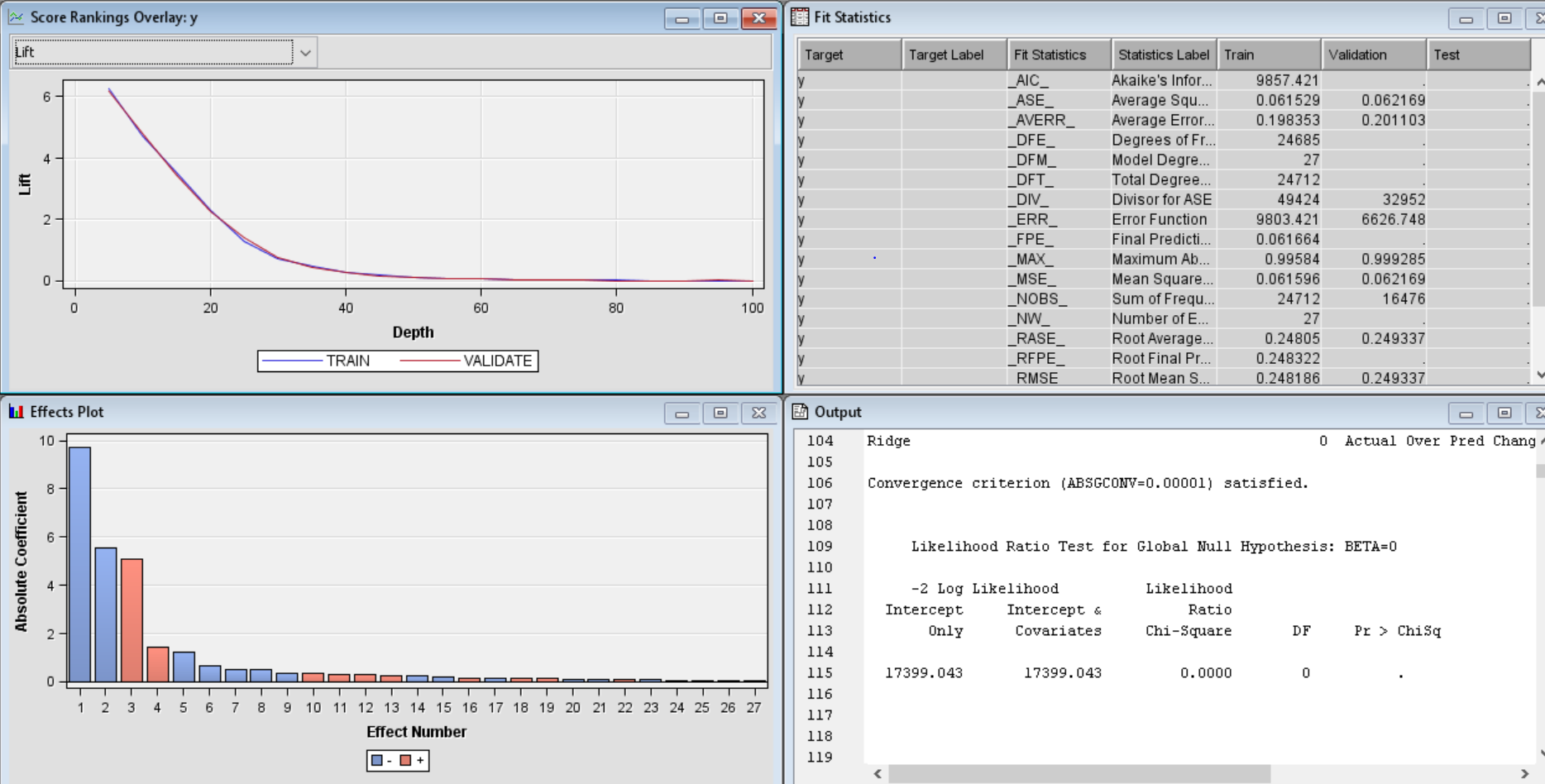
**Log10 transformation will control the skewness and kurtosis in data**



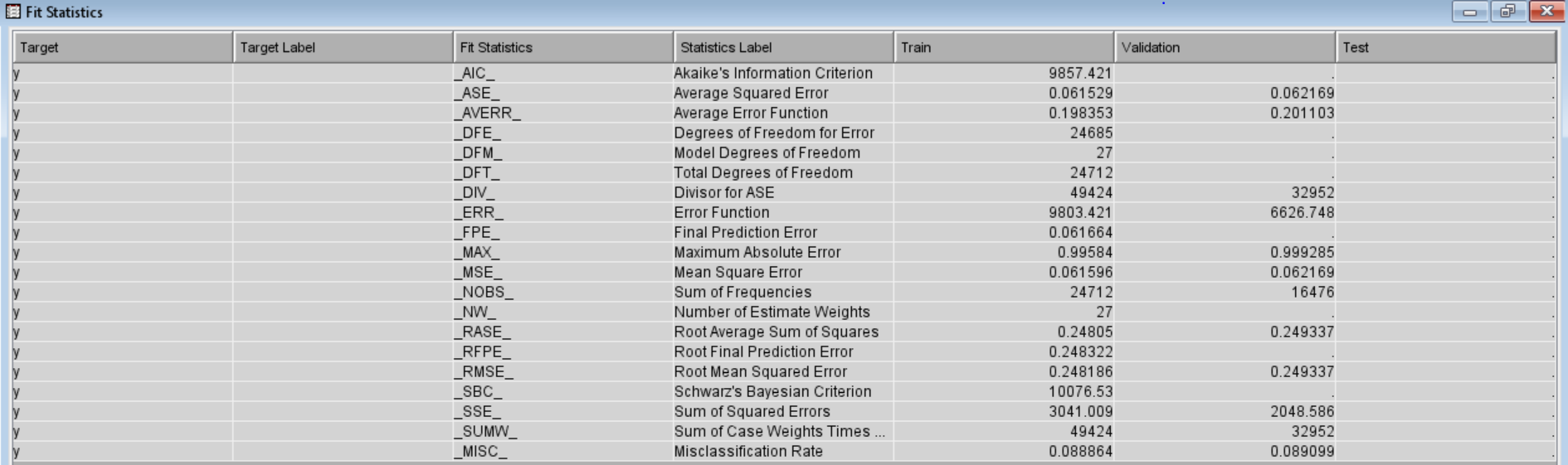
**Selection model in properties is step wise**

**Regression statistics**



**Fit Statistics for Logistic Regression Model**



**As you can see root mean squared error is 0.24 and Maximum Absolute Error is 0.99**

**Average Squared Error is 0.06**

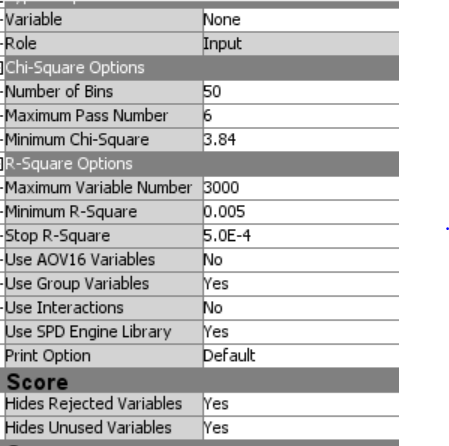
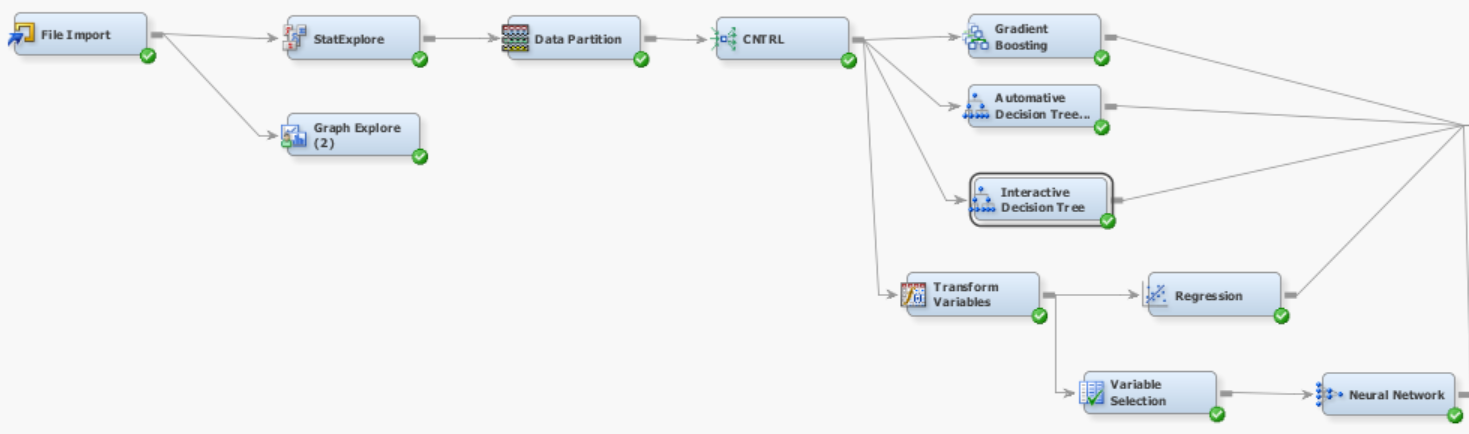
**Neural Network Model**

Neural networks are using parametric values and can handle varieties of nonlinear relation target and input variables.

They are better than Regression models

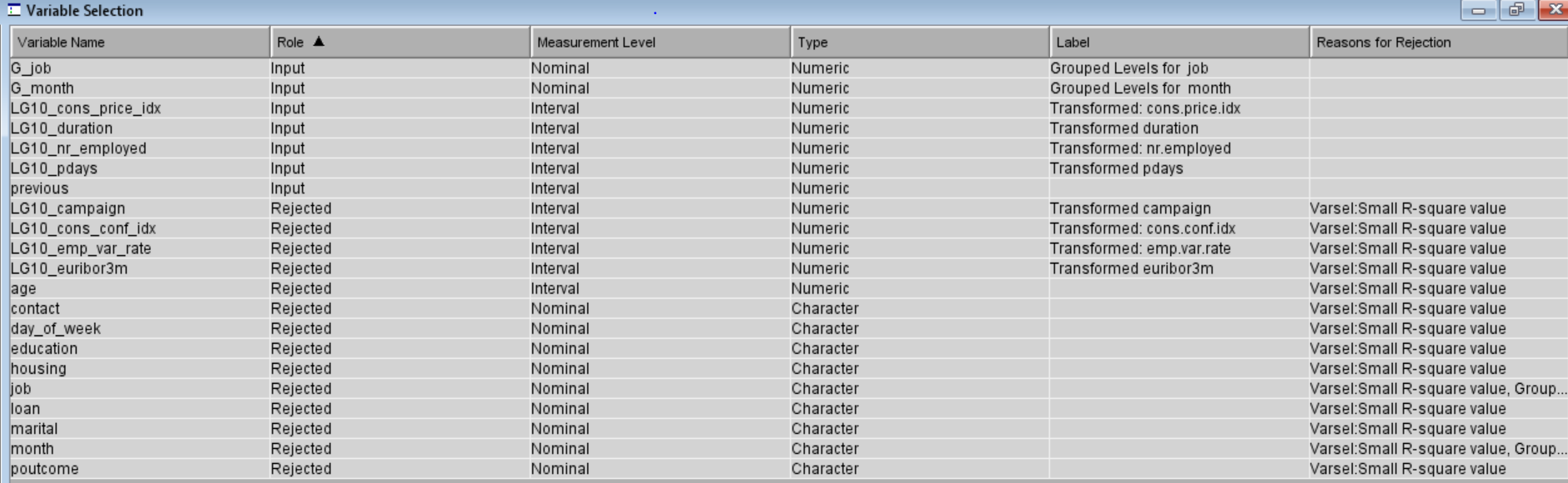
As they work on phenomena that should have more r2 value by using variable selection node

**Variable Selection**

**Variables having less than R – Square value to 0.005 has been rejected for neural network**

**Following variables gets rejected because of low r square value**



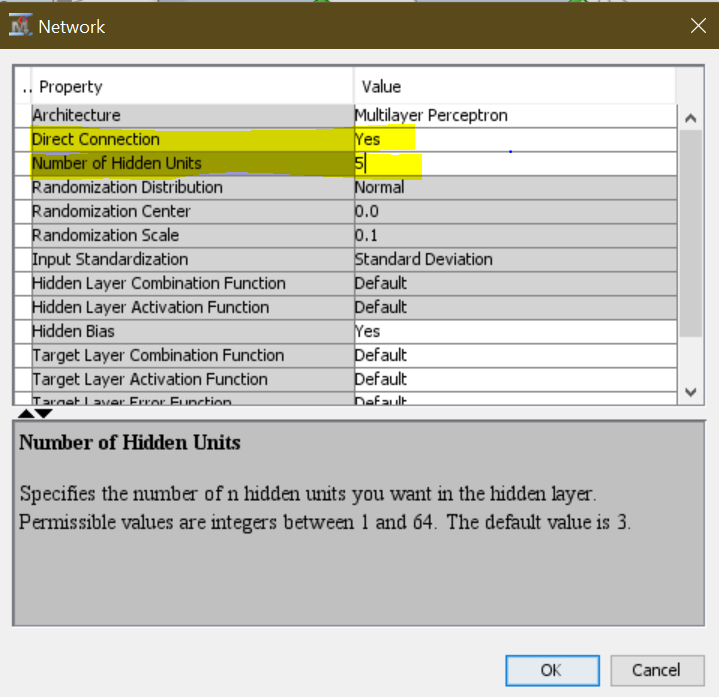
**So, we will selected variables as input for neural network**

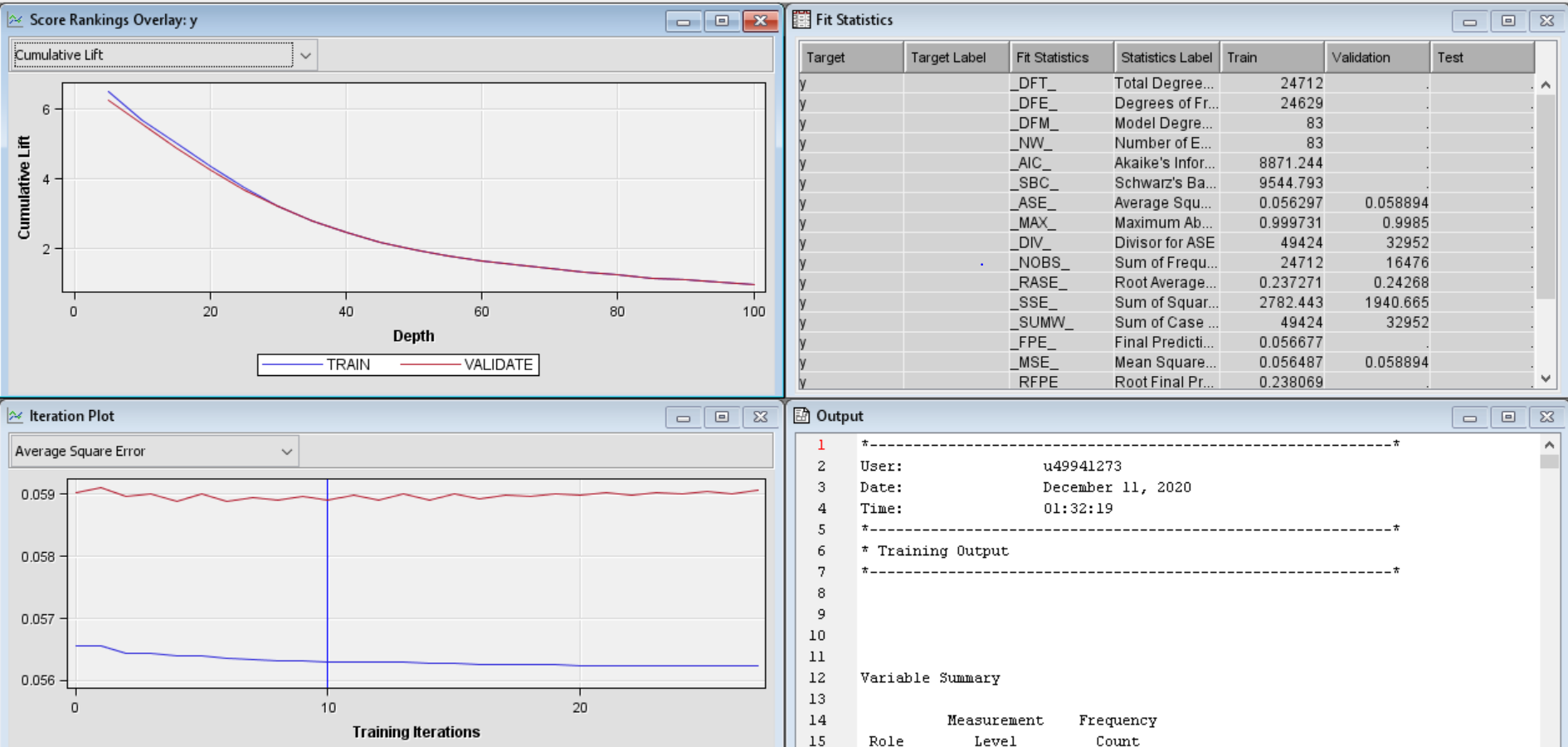
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| G\_job |  |  |  |  |  |
| G\_month |  |  |  |  |  |
| LG10\_cons\_price\_idx |  |  |  |  |  |
| LG10\_duration |  |  |  |  |  |
| LG10\_nr\_employed |  |  |  |  |  |
| LG10\_pdays  Previous |  |  |  |  |  |

**Adjusted data is now more suitable for modelling**

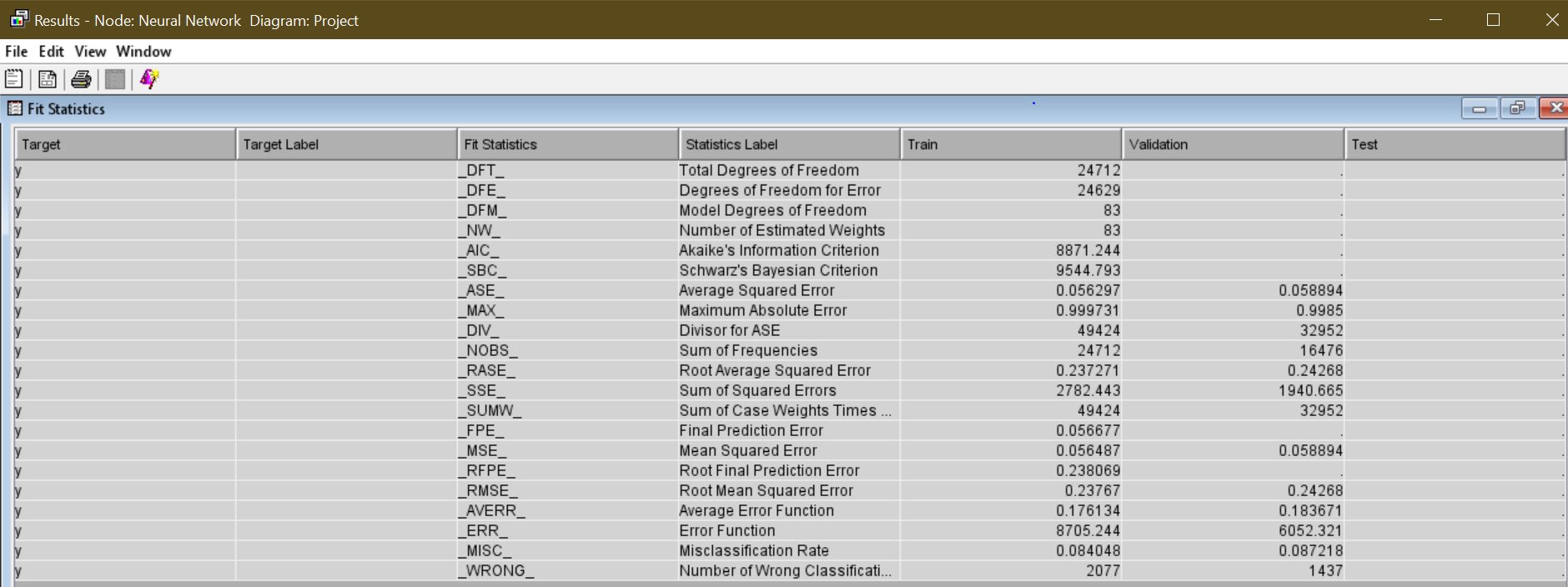
**Setting Direct Connection to yes so that Neural Network can make connections between nodes.**

**Hidden Units are set to 5 for data processing and imputation, these are hidden layers in Neural Network for decision flow**





**Fit Statistics for Neural Network**



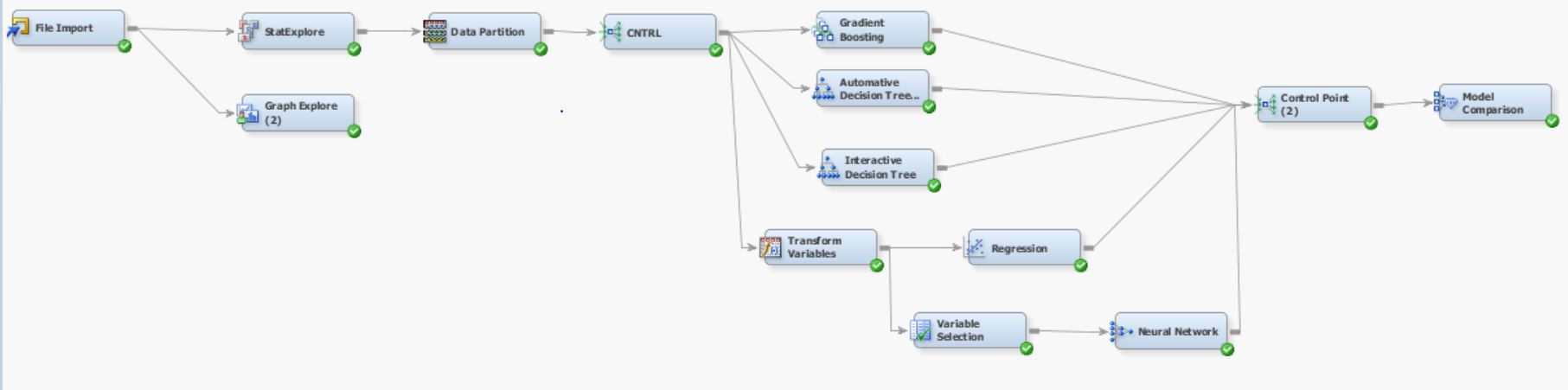
**As you can see root mean squared error is 0.23 and Maximum Absolute Error is 0.99**

**Average Squared Error is 0.056**

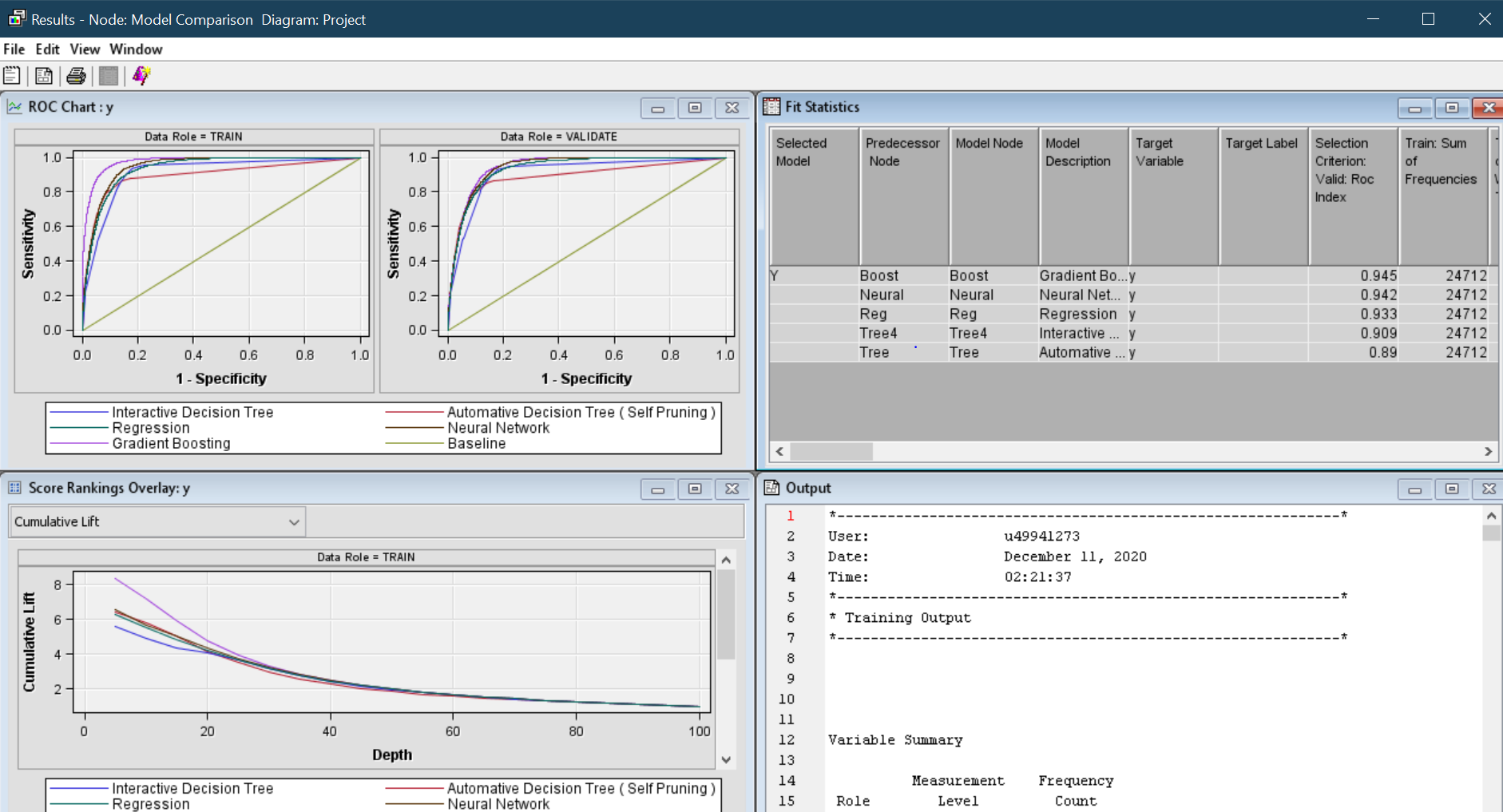
**Model Comparison**

* **Decision Tree (Self Pruning)**
* **Interactive Decision Tree**
* **Gradient Boosting – Rounding with of Decision Tree**
* **Logistic Regression (Parametric)**
* **Neural Network (Parametric)**

**Final Data Flow**

****

**Model Comparison Node**



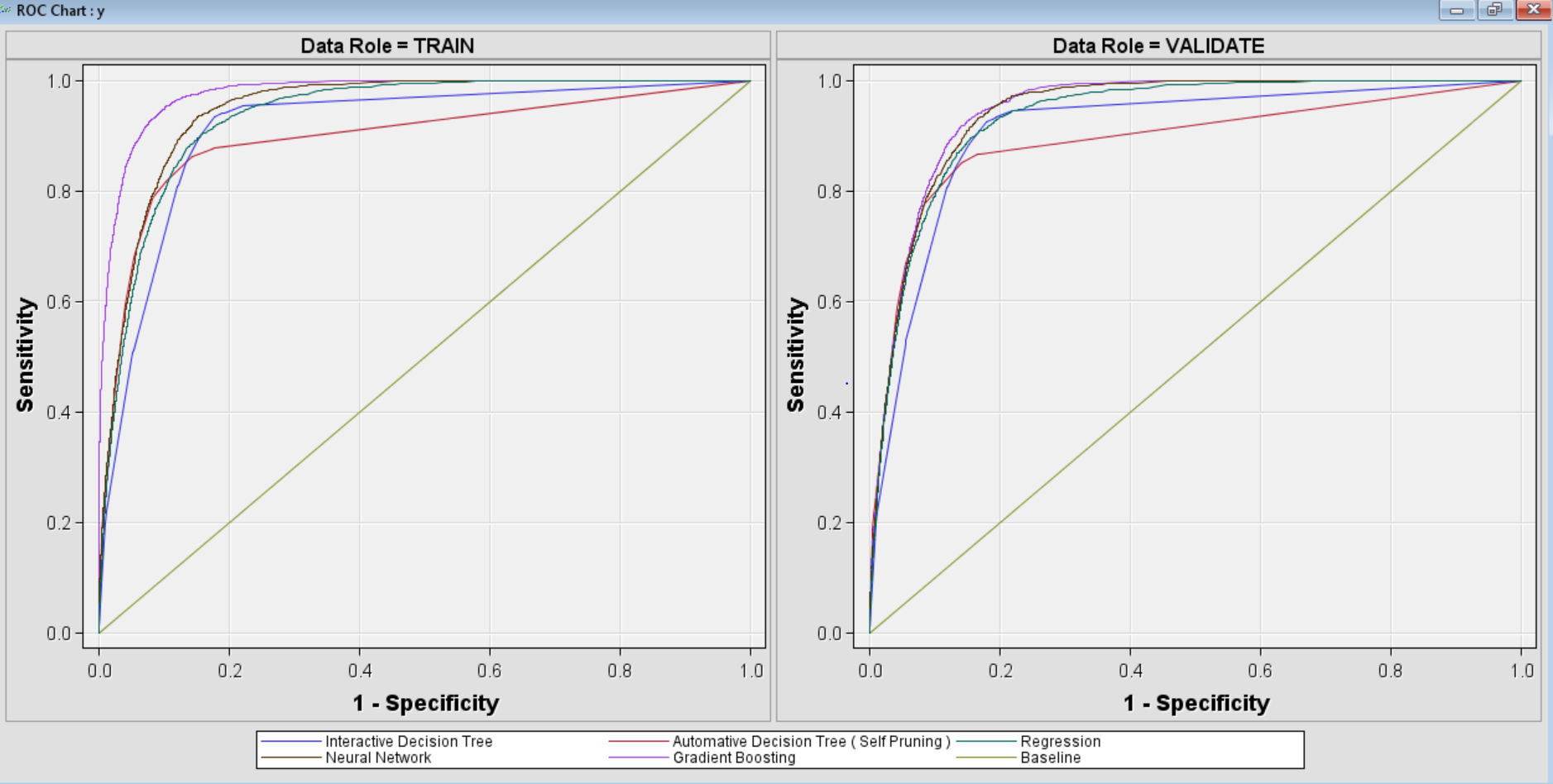
**ROC (receiver operating characteristic curve) shows the classifications threshold for our all-predictive models.**

**It has 2 parameters: True Positive Rate and False positive Rate.**

**AUC (Area Under Curve): It shows the entire area under ROC Curve, More the Area More will the AUC value**

**ROC Curve with all predictive Models**

**Clearly shows Gradient boosting shows more accuracy than others.**



**Full Statistics of all predictive models.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistics** | **Gradient Boosting** | **Neural Network** | **Logistic Regression** | **Interactive Decision Tree** | **Decision Tree** |
| Train: Roc Index | 0.977 | 0.947 | 0.935 | 0.913 | 0.896 |
| Selection Criterion: Valid: Roc Index | 0.945 | 0.942 | 0.933 | 0.909 | 0.89 |
| Train: Gini Coefficient | 0.953 | 0.894 | 0.87 | 0.826 | 0.792 |
| Train: Cumulative Lift | 7.156447613312895 | 5.684223066670502 | 5.5010926962344975 | 4.8844084258439375 | 5.799667673530911 |
| Train: Lift | 5.931987685495686 | 4.84038704603401 | 4.72548171556436 | 4.160017236686766 | 5.210653924377103 |
| Train: Root Average Squared Error | 0.19647074893488817 | 0.2372707618801991 | 0.2480503884217969 | 0.25381333546752005 | 0.24203880307426232 |
| Train: Sum of Squared Errors | 1907.8037243640604 | 2782.4434114412184 | 3041.0090585772023 | 3183.9538465229725 | 2895.3954271375505 |
| Valid: Roc Index | 0.945 | 0.942 | 0.933 | 0.909 | 0.89 |
| Valid: Gini Coefficient | 0.891 | 0.884 | 0.866 | 0.818 | 0.779 |
| Valid: Cumulative Lift | 5.585928347783977 | 5.585928347783977 | 5.48896912863247 | 4.877679947072119 | 5.7002196553329645 |
| Valid: Lift | 4.8910539438648515 | 4.912600437009631 | 4.804867971285734 | 4.1744208328073045 | 5.1211424966558265 |
| Valid: Root Average Squared Error | 0.240683601219847 | 0.24268023767596836 | 0.24933667375164256 | 0.2552980602418288 | 0.24491223889195823 |
| Valid: Sum of Squared Errors | 1908.8630919700777 | 1940.6651285369219 | 2048.5855356684688 | 2147.715784807899 | 1976.527020820928 |

As per the predictive models’ statistics designed in SAS Enterprise Miner. Data divided in 60% - 40% percent for training and validation. The accuracy of predictive models always depends upon the data splitting, more data we have for training means we can train our model efficient but still we need more validation data to check the efficiency. Validation Data is different than from Testing data as validation data can be used back for training the model and we can find best efficient predictive model for our dataset.

Here **Receiver operating characteristic index** shows the model Gradient Boosting has 0.945 and the lowest is in Decision Tree of 0.89 in validation dataset.

And if we talk about the performance on the basis of Error, **(Root Average Squared Error)** shows the model Gradient Boosting has lowest 0.240 and highest error shown in Interactive Decision Tree.

Neural Network always in second rank having **ROC index 0.942 and Root Average Squared Error of 0.242.** Or in other words there is no big difference in the efficiency of Neural network and Gradient Boosting Model.

**Cumulative LIFT** shown in Validation data is **same** for Neural Network and Gradient Boosting with **5.585928347783977**

**Gain Coefficient** shown in Training and Validation data is highest for Gradient Boosting with **0.935 and 0.891**

**Research paper Comparison:**

**Citation 1**

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing, Decision Support Systems, Elsevier, 62:22-31, June 2014 https://core.ac.uk/download/pdf/55631291.pdf

Study shows the Data mining models for telemarketing campaign. Study shows the comparison of 4 predictive models

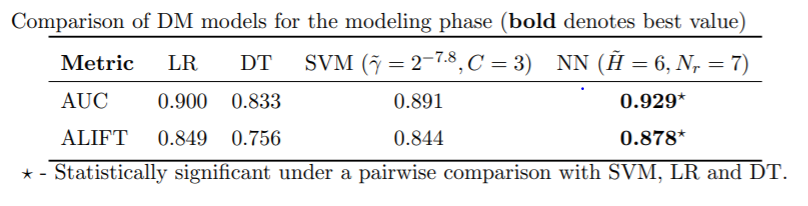
Decision Tree

Neural Network

Support Vector Machine

Logistic Regression

On the basis of Area Under Curve and ALIFT Analysis. Study finally comes to conclusion for Neural Network is the best model for the prediction of Term Deposit account for new customers on the basis of previous campaign with AUC Score of 0.929 and ALIFT Score of 0.878.



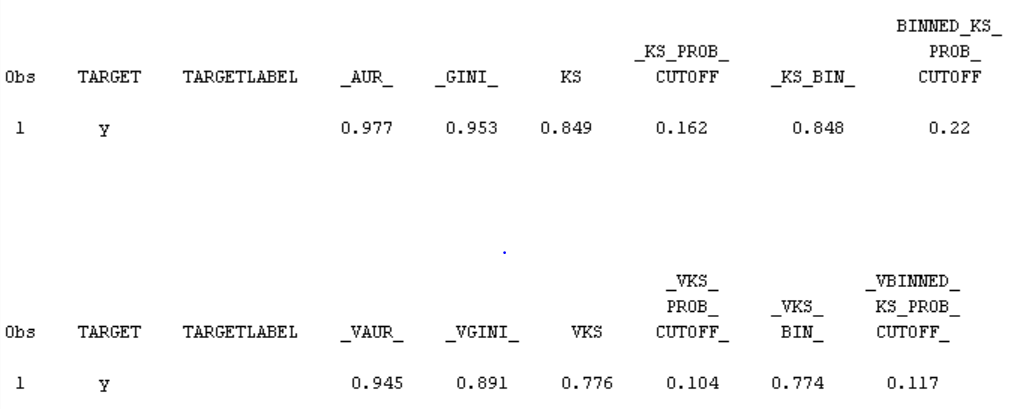
And for Analysis they used the dataset of 52944 rows from 2008 to 2013.

And our Dataset for analysis in 41,188 rows.

Study shows the dataset they used for research is unbalanced as 6557 observations for success only (12.38%)

**Our Gradient Boosting Model and Neural Network Model shows the high accuracy than the models Neural Network prepared in the study.**

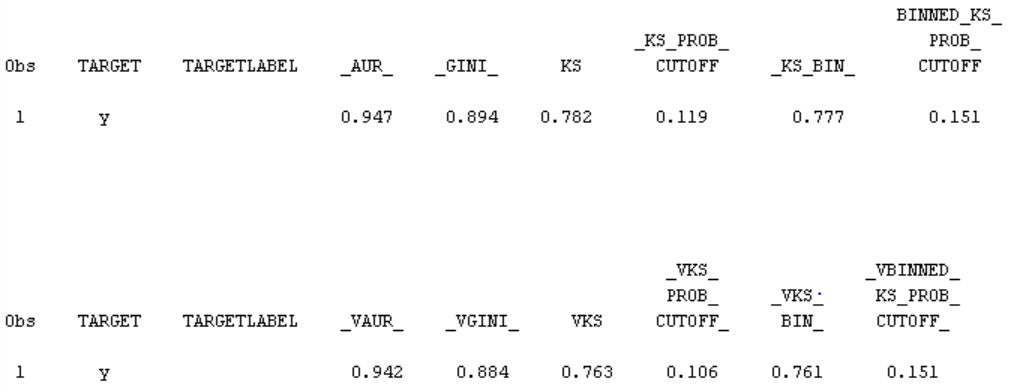
Below observation is for Gradient Boosting Model



**As you can see AUR (operating characteristic curve) shows 0.945 in the validation data. Which is much higher than the neural network used in the study of 0.929**

**Also, ROC curve and LIFT analysis shows ROC index 0.945 in Gradient Boosting**

Below observation is for Neural Network Model



**As you can see AUR (operating characteristic curve) shows 0.942 in the validation data. Which is much higher than the neural network used in the study of 0.929**

**Also, ROC curve and LIFT analysis shows ROC index 0.942 in Gradient Boosting**

**Citation 2**

[Justice Asare-Frempong](https://ieeexplore.ieee.org/author/37086262906). [Manoj Jayabalan](https://ieeexplore.ieee.org/author/37085782526). Predicting customer response to bank direct telemarketing campaign. 2017 IEEE The International Conference on Engineering Technologies and Technopreneur ship (ICE2T 2017), <https://ieeexplore.ieee.org/document/8215961>

Aim for this study is to identify the customers based on previous campaign that they will opt for term deposit account. Data used in study takes 45147 observations with 17 variables.

Study shows 4 predictive models

Multilayer Perceptron Neural Network (MLPNN), Decision Tree (C4.5), Logistic Regression and Random Forest (RF)

****

In addition to this analysis study shows that the month of September proved to be the most-fertile month for customers in relation to subscription of term deposit.

**Their Maximum Area under curve shown by random forest of 92.7% means 0.927 Area under curve.**

**Our model both gradient boosting and Neural Network shows 0.945 and 0.942 which is much higher accuracy than their model.**

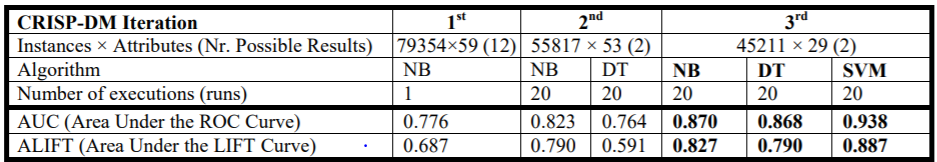
**Citation 3**

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]

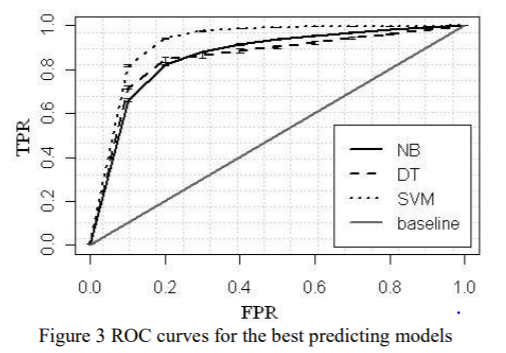
Study shows an implementation of data mining project based upon the CRISP-DM life cycle. Data set contains 45211 observations out of which 5289 are successful having 11.7 success rate.

Study shows 3 data mining models for the prediction of telemarketing campaign

Naïve Bayes (NB), Decision Trees (DT) and Support Vector Machines (SVM)



ROC curve



**Their result shows maximum Area under curve value is 0.938 by support vector machine.**

**Our model both gradient boosting and Neural Network shows 0.945 and 0.942 which is much higher accuracy than their model.**

**Citation 4**

Oluwaseun Esther Oluwabusola. Applying Business Analytics in Practice to a Bank Telemarketing Dataset. University of Strathclyde (2015). https://local.cis.strath.ac.uk/wp/extras/msctheses/papers/strath\_cis\_publication\_2714.pdf

Study shows 6 predictive models for bank telemarketing dataset

Study also shows data mining techniques which can improve the efficiency of the model

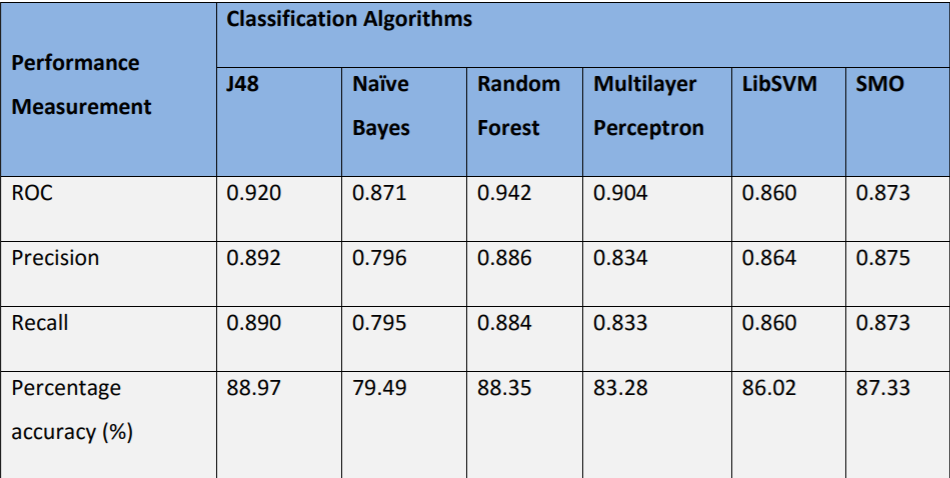
Bagging Result

Boosting Result

Classifier stacking result

After all the stacking up final accuracy is as follow:

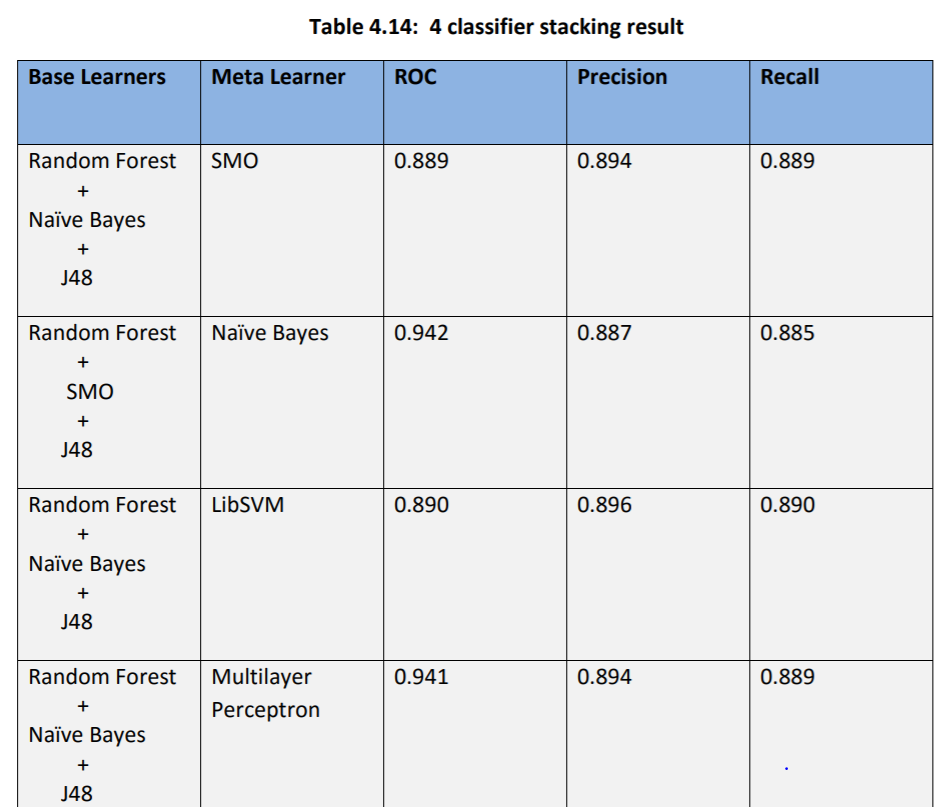
Study Result:



**Their result shows the maximum ROC value to be 0.942 for Random Forest which is equal to the ROC value of our model Neural Network 0.942 but out gradient boost model shows much higher accuracy of ROC value 0.945**

**Stacking results are quite impressive. Using base learners as Random Forest + SMO+J48 and meta learner as Naïve Bayes results shows ROC value of 0.942**

**Check below table:**



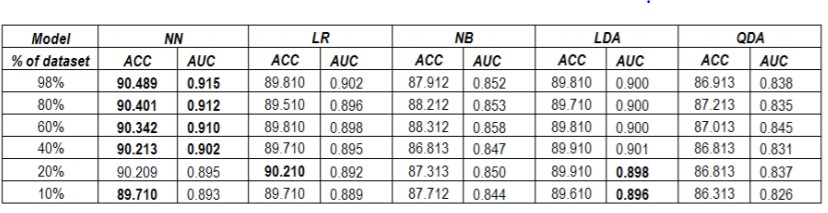
**Citation 5**

Nachev, A. (2015). Application of data mining techniques for direct marketing. Computational Models for Business and Engineering Domains. Available at: <http://www.foibg.com/ibs_isc/ibs-30/ibs-30-p09.pdf>.

Study shows the cross-validation and multiple runs for the partitioning of train and test sets (70% and 30%) for the direct marketing response task. Dataset contains 45211 observations and 17 variables.

Aim of this study is to compare different models on the basis of dataset percentage use for telemarketing research.

Study have taken Neural Networks (NN), Logistic Regression, Naïve Bayes, Linear and Quadratic Discriminant Analysis (QDA)



**Results on the basis of the data saturation maximum Area under curve achieved by Neural Network of about 0.915.**

**Our model both gradient boosting and Neural Network shows 0.945 and 0.942 which is much higher accuracy than their model.**

Below image shows the ROC curve for different models. Leading 0.915 neural network.

