Big Data & Predictive Analytics-MKTG 746

Group Project

Term Deposit Subscription Predictive Analytics

Group 3 Members:

Chinenye Iwunze 301247045

Kunfayat Adekanmbi 301195456

Shola Fabiyi 301192414

Saad Uddin Mir 301297765

Siddhant Shah 301298415

Submitted to Professor Resmi Ann Thomas

Date: 19/08/2023

Table of Contents

1. Executive Summary and Introduction
2. File Import
    2.1. Data Source
    2.2. Data Dictionary
    2.3. File Import
    2.4. Data Leakage
3. Data Wrangling
    3.1. Data Filter
    3.2. Data Partition
4. Decision Tree
    4.1. Maximal Tree
    4.2. Two-split ASE Tree
    4.3. Three-split ASE Tree
    4.4. Four-split ASE Tree
    4.5. Two-split Misclassification Tree
    4.6. Decision Tree Summary
5. Regression
    5.1. Data Massaging
    5.1.1. Data Imputation
    5.1.2. Data Replacement
    5.2. Full Regression
    5.3. Forward Regression
    5.4. Backward Regression
    5.5. Stepwise Regression
    5.6. Regression Summary
6. Neural Networks (NN)
    6.1. Neural Networks connected from Impute Node
    6.2.0 Three Hidden Units 50 Iterations NN
    6.2.1. Three Hidden Units 100 Iterations NN
    6.2.2. Four Hidden Units 50 Iterations NN
    6.2.3. Four Hidden Units 100 Iterations NN
    6.2.4. Five Hidden Units 50 Iterations NN
    6.3. Neural Network Summary
7. Model Comparison
    7.1 Diagram of Model Comparison
    7.2 Settings of Model Comparison Node
8. Conclusion
    8.1. Summary
    8.2. Recommendations for the business
    8.3 Recommendations for the model
9. References

1. Executive Summary

A bank aims to enhance its valuation through a campaign for its deposit program. However, up until now the method used to identify potential customers interested in opening term deposit accounts still relies on manual processes. The bank desires to pinpoint customers who truly hold the potential to engage with the deposit program campaign. By identifying these prospective customers, the bank can better manage its budget for marketing campaigns and enhance time efficiency.

A predictive study was performed to identify the variables that had the strongest influence on whether or not a client will want to subscribe to a term deposit. The Statistical Analysis System (SAS) was employed as the environment through which predictive analysis will be performed. The probability(decision) trees, several regression models, and neural networks were run after the data was imported and wrangled. A node called Model Comparison was then used to choose the most efficient model based on set parameters.

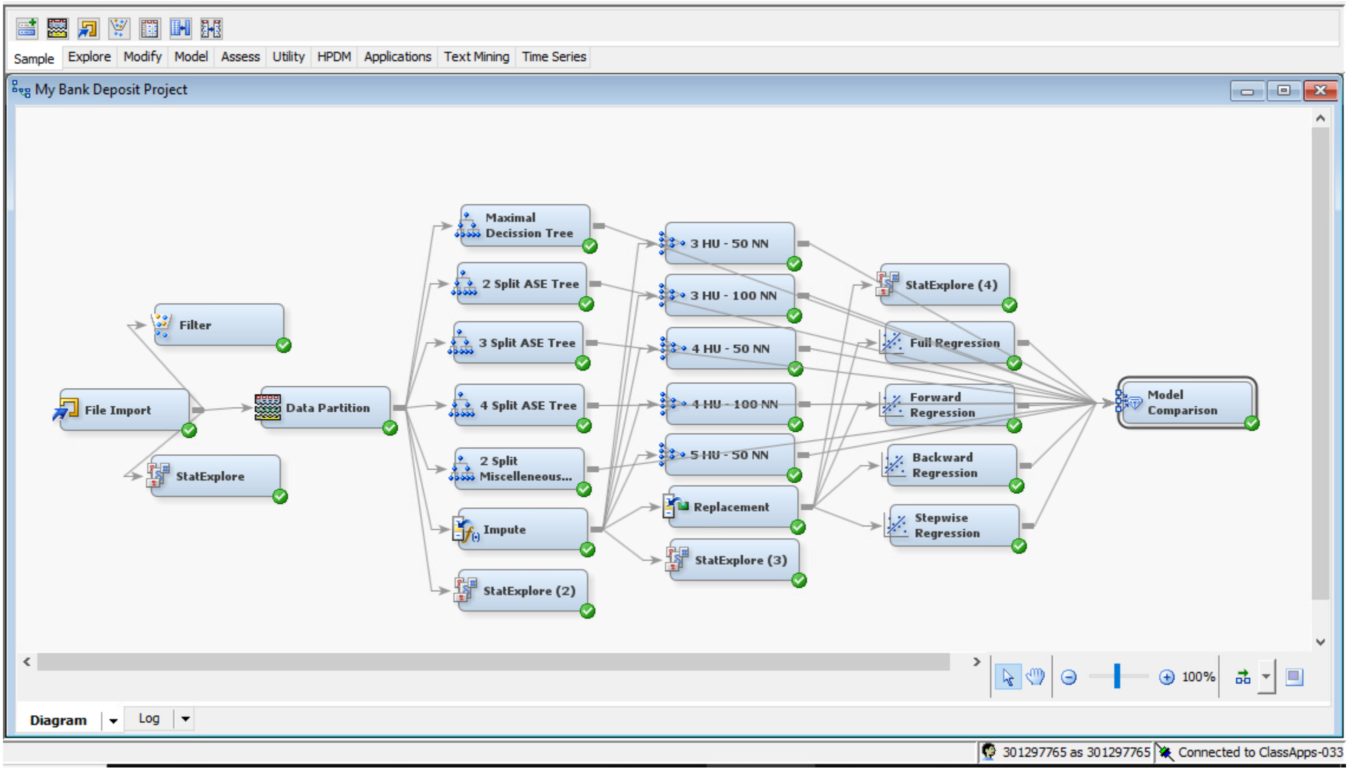


Figure 1A. The Comprehensive Predictive Analysis Model

The model comparison using Average Square Error was used at the end to determine the best model to proceed with and to get insights into input variables which helps us to target customers who are most likely to proceed with bank deposits.

Going through various models, we were able to determine variables like "Duration, Month, Previous campaigns, loan" which have a huge impact in identifying customers who are most likely to buy deposits. and recommend the following strategies to increase the potential for customers to engage in deposit offerings.

- Offer campaigns with longer durations to customers, as evidenced by the impact of longer durations on customer engagement with deposits.
- Increase deposit campaign offerings throughout the first quarter (January, February, and March), the second three months (July, August, and September), and the last three months (October, November, and December).
- Evaluate the success of each campaign; the analysis indicates that customers who were successfully engaged in a previous deposit campaign are more likely to engage in subsequent campaigns.
- Focus deposit campaigns on customers who do not have home loan installments, as the analysis suggests that customers without home loan instillments are more likely to engage with deposit offerings.

Introduction

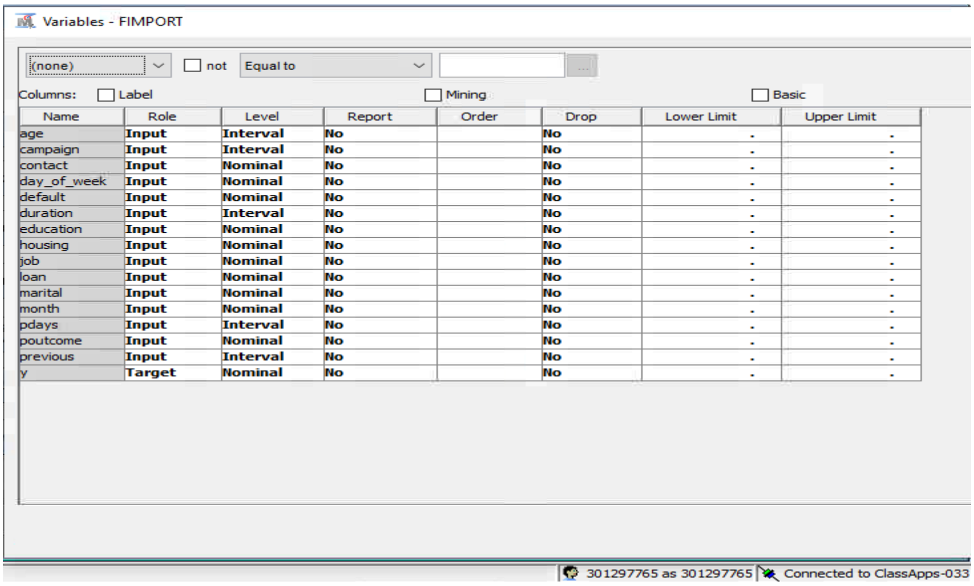
Banks exist to offer customers monetary services while also making additional revenue. Therefore, banks dedicate substantial capacities and efforts to acquiring funds. Banks can accomplish this by engaging in physical cross-selling initiatives and deliver services. As a financial institution, a bank must avoid losing deposit customers, as such losses could diminish the bank's assets. Furthermore, the bank must actively seek other customers to open deposit accounts. As the deposit balance grows, it also increases the potential loan amount that can be extended to customers. Consequently, the bank earns profits through the interest on these loans. Thus, it can be asserted that an increase in deposit balances also augments the bank's profits.

Given the situation, we started to assess which elements in the data set might contribute to a high amount of term deposit transactions. Our team discovered a data set that was the outcome of the cross-selling marketing campaign by a bank to provide term deposits.

2. File Import

We have gone through dataset to make sure that there are no blank spaces, or duplicate entries before importing the dataset into SAS Enterprise Miner, The CSV file was then loaded using the File Import node into SAS Enterprise Miner.

The level of each variable was chosen to match the suitable data type by using the module known as the Edit Variable of the File Import node. The target variable was "Y" because the goal of the analysis was to determine whether a consumer would decide to subscribe to a bank term deposit or not. This fits to a binary variable where "Yes" meant the customer has subscribed to a term deposit. "No" meant that the customer had not subscribed to a term deposit.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
age	Input	Interval	No		No	.	.
campaign	Input	Interval	No		No	.	.
contact	Input	Nominal	No		No	.	.
day_of_week	Input	Nominal	No		No	.	.
default	Input	Nominal	No		No	.	.
duration	Input	Interval	No		No	.	.
education	Input	Nominal	No		No	.	.
housing	Input	Nominal	No		No	.	.
job	Input	Nominal	No		No	.	.
loan	Input	Nominal	No		No	.	.
marital	Input	Nominal	No		No	.	.
month	Input	Nominal	No		No	.	.
pdays	Input	Interval	No		No	.	.
poutcome	Input	Nominal	No		No	.	.
previous	Input	Interval	No		No	.	.
y	Target	Nominal	No		No	.	.

Fig, 2.21 File Import showing Variables.

2.1 Data Source

This data is publicly available for research at Kaggle.com.
https://www.kaggle.com/datasets/aslanahmedov/predict-term-deposit?resource=download

From May 2008 through November 2010, the bank made phone calls to potential buyers. More than one contact with the same client was frequently required to determine whether a client will place an order. The complete data collection, bankadditional-full.csv, was employed.

There are 41,188 observations and 21 Variables in the Data Set. The target response (y) is a binary response that indicates whether the client has signed up for a term deposit. 'Yes,' represented that the client has signed up for a term deposit. 'No' indicates that the client did not sign up for a term deposit.

2.2 Data Dictionary

Data dictionary provides a short idea on the variables in the data set. It assigns the variable name, its category, its description, and the type of variable. The variables are broken into 3 categories: Numerical, Binary and categorical.

Variable	Description	Variable Type	Variable Category
age	Customer's age at the time of the call	Numeric	Input
Job	Customer's type of job - 'administrative.','bluecollar','business person ','housemaid','management','retired','selfemployed','services','student','technician','unemployed','unknown')	Categorical	Input
marital	Clients Marital Status at time of call - 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed	Categorical	Input
Education	Clients educational background at time of call - 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'	Categorical	Input
default	Does customer have credit in default? - 'no','yes','unknown'	Binary	Input
housing	Does customer have a house loan? - 'no','yes','unknown'	Categorical	Input
loan	Does customer have a personal loan? - 'no','yes','unknown'	Binary	Input
contact	Communication type with client – 'cellular','telephone'	Categorical	Input
day	Last contact day of week with the customer - 'mon','tue','wed','thu','fri'	Numeric	Input
month	Last contact month of year with the customer - 'jan', 'feb', 'mar', ..., 'nov', 'dec'	Categorical	Input
duration	Last time called in seconds to Customer. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.	Numeric	Input
campaign	Number of contacts performed during this campaign for this customer (includes last contact)	Numeric	Input
pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)	Numeric	Input
previous	Number of contacts performed before this campaign and for this client	Numeric	Input

poutcome	Outcome of the previous marketing campaign - 'failure','nonexistent','success'	Categorical	Input
y	Has the customer subscribed a term deposit? - 'yes','no'	Binary	Target/R esponse

3. Data Wrangling

The data was analyzed using Decision Tree, Neural Network, and Regression models. Because the goal answer (y) has 88.73% negative replies and 11.27% positive ones. It is advisable to employ all yes responses and the same number of no responses when modelling the data. This makes sure that we completely understand the variables in the model influence 'yes' and 'no' responses. A model would predict that 'no' variables made a difference if there were too many 'no' observations.

To overcome this imbalance in the target variable, we simplified and stratified input variables such as job, education, default, marital, housing, month and loan to guarantee that the sample as closely resembles the unfiltered data as feasible. This strategy increases the model's ability to recognize which variables influence target response (y). Following that, the data was divided 50/50 into validation and training data sets.

3.1 Data Filter

A node known as StatExplore was connected to the File Import node to further analyse the data set in order to determine whether they were redundant or unimportant.

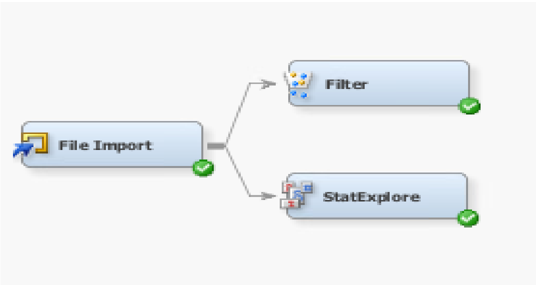
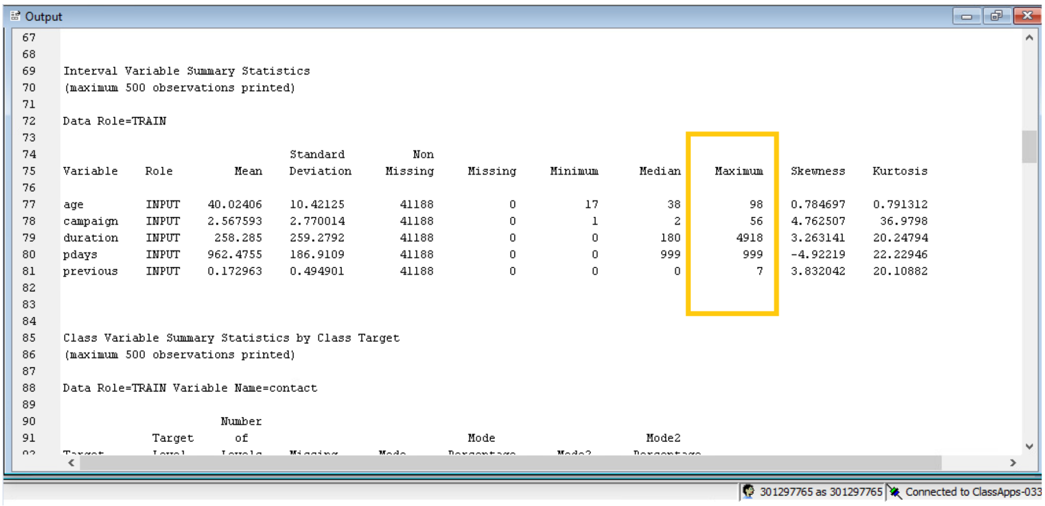


Figure 3. 1.1 Stat Explore and Filter Nodes Connected



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

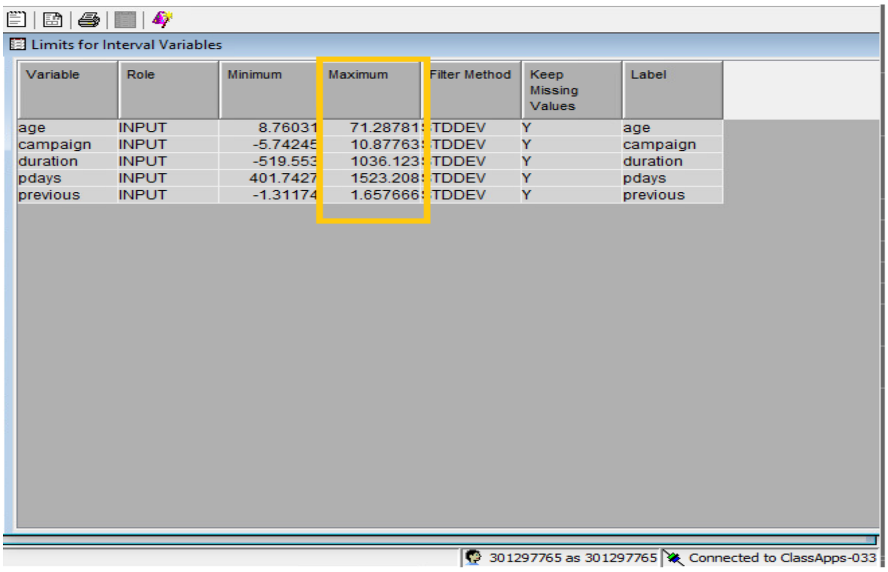
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	40.02406	10.42125	41188	0	17	38	98	0.784697	0.791312
campaign	INPUT	2.567593	2.770014	41188	0	1	2	56	4.762507	36.9798
duration	INPUT	258.285	259.2792	41188	0	0	180	4918	3.263141	20.24794
pdays	INPUT	962.4755	186.9109	41188	0	0	999	999	-4.92219	22.22946
previous	INPUT	0.172963	0.494901	41188	0	0	0	7	3.832042	20.10882

Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=contact

Figure 3. 1.2 Output after Stat Explore and Filter Nodes Connected

The StatExplore node shows that the age, campaign, duration, pdays, and previous variables contained some extreme values that were almost impossible to happen.



Limits for Interval Variables

Variable	Role	Minimum	Maximum	Filter Method	Keep Missing Values	Label
age	INPUT	8.76031	71.28781!	TDDEV	Y	age
campaign	INPUT	-5.74245	10.87763!	TDDEV	Y	campaign
duration	INPUT	-519.553	1036.123!	TDDEV	Y	duration
pdays	INPUT	401.7427	1523.208!	TDDEV	Y	pdays
previous	INPUT	-1.31174	1.657666!	TDDEV	Y	previous

301297765 as 301297765 Connected to ClassApps-033

Results from the Filter Node

The maximum value for age was 98, however, people who are 98 are less likely to take such financial decisions and moreover, it exceeds life expectancy limits and there may be very few examples. So, this has been replaced by 71 as shown in the above filter node output along with other variable changes.

The figure above shows the maximum values selected by the filter for the input variable.

3.2 Data Partition

The model's performance was adjusted using the Data Partition node to prevent either overfitting or underfitting.

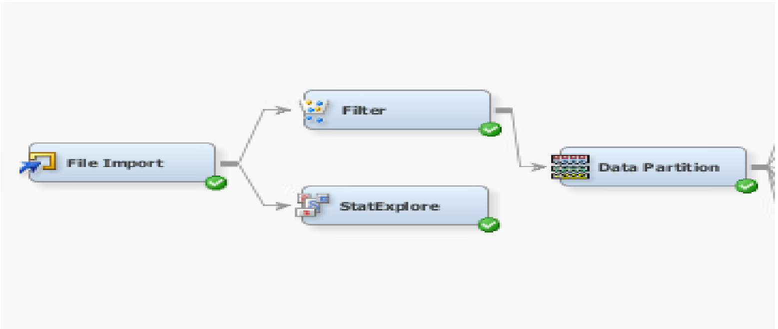
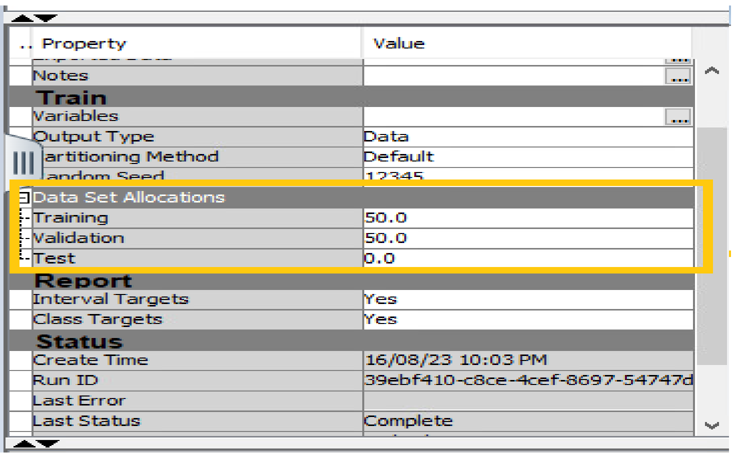


Figure 3.2.1. Connection of Data Partition Node

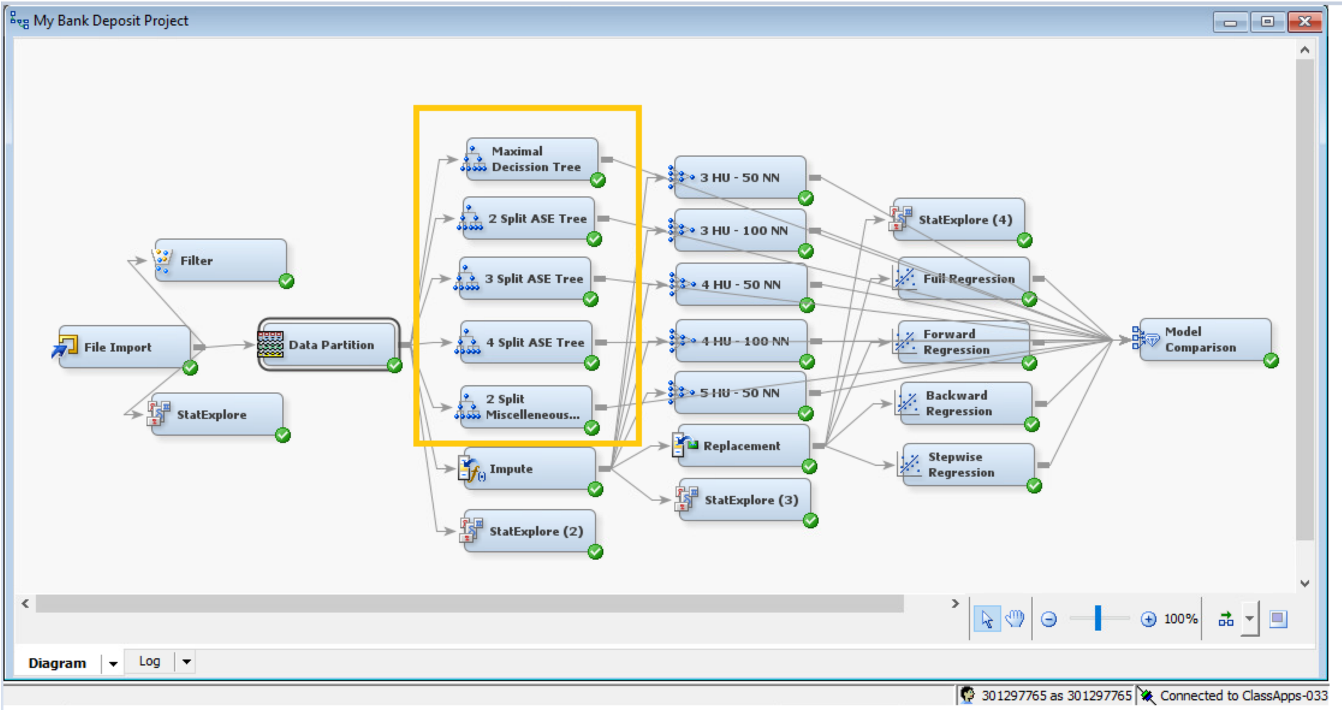
In the Property Panel, under Data Set Allocations, the validation value was modified to 50.0, the training value was also modified to 50.0, and the test value was set to 0.



Property	Value
Notes	
Train	
Variables	
Output Type	Data
artitioning Method	Default
andom Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	16/08/23 10:03 PM
Run ID	39ebf410-c8ce-4cef-8697-54747d
Last Error	
Last Status	Complete

Figure 3.2.2 Data Set Allocations in the Property Panel

4. Decision Tree



4.1 Maximal Tree

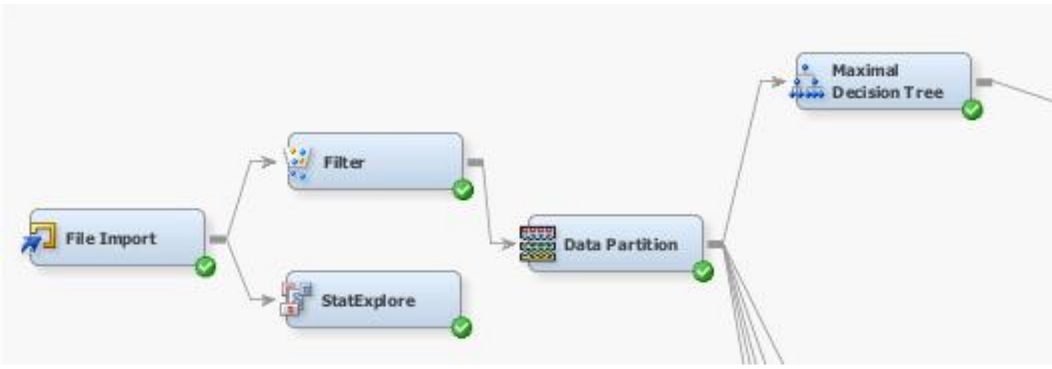


Figure 4.1.1. Connection of the Maximal Tree Model

The Maximal tree utilizing an interactive training method and Average Square Error for evaluation yielded the following outcomes.

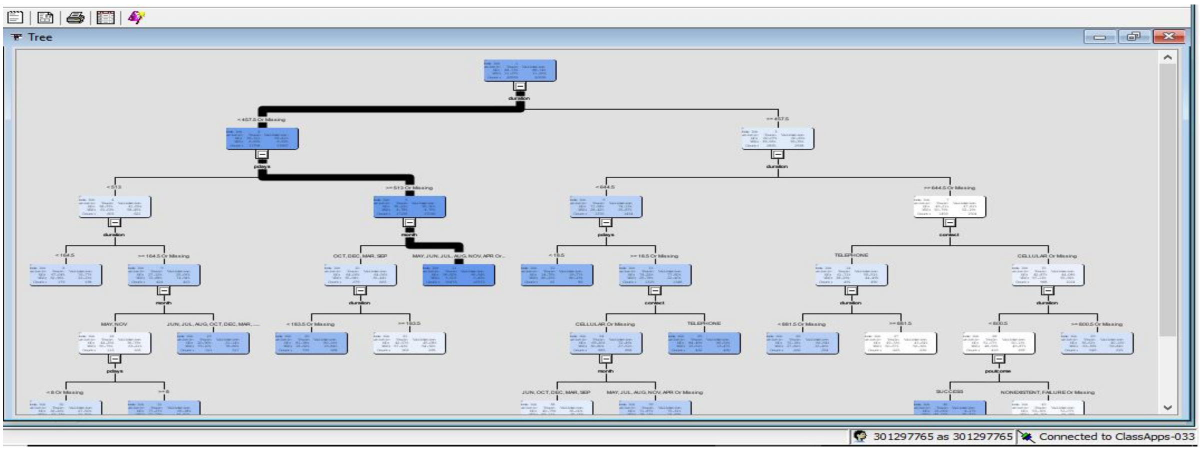


Figure 4.1.2. Maximal Decision Tree Result

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_NOBS_	Sum of Frequen...	20593	20595	.
y	y	_MISC_	Misclassificatio...	0.086	0.083515	.
y	y	_MAX_	Maximum Absol...	0.987718	0.987718	.
y	y	_SSE_	Sum of Squared...	2499.778	2441.953	.
y	y	_ASE_	Average Square...	0.060695	0.059285	.
y	y	_RASE_	Root Average S...	0.246363	0.243485	.
y	y	_DIV_	Divisor for ASE	41186	41190	.
y	y	_DFT_	Total Degrees o...	20593	.	.

Figure 4.1.3. Statistics result of Maximal Tree model.

The variable 'duration' shows up two times in the tree according to the interactive result after the training node. This suggests that the decision tree model may be fine-tuned to reach a better ASE (0.059285). Consequently, decision tree models were executed with 2, 3, and 4 splits afterward. Also, the criterion was altered to misclassification to find the optimal model.

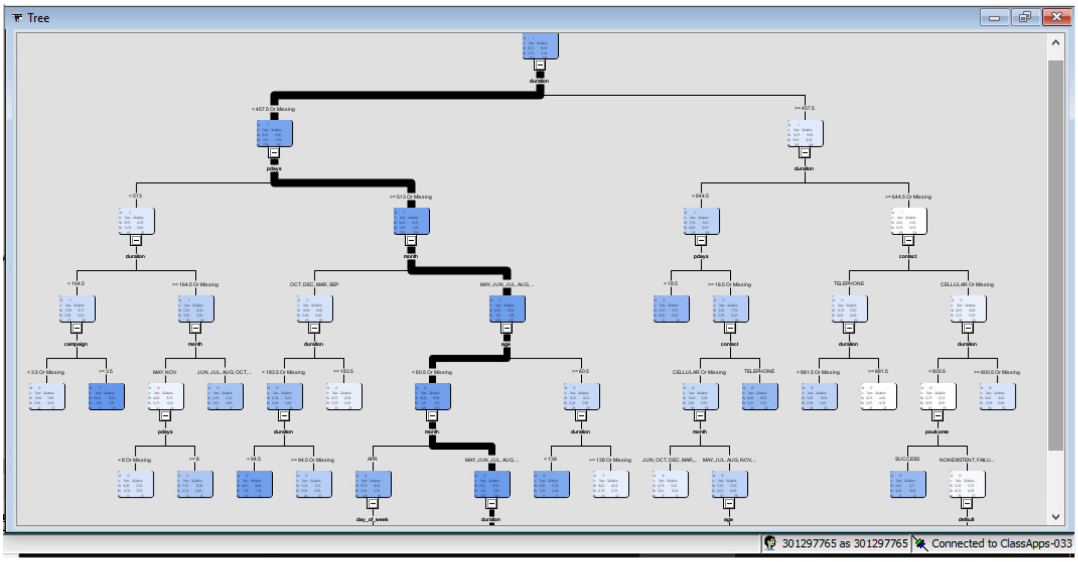
4.2. 2-Split ASE Tree

Property	Value
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment

Figure 4.2.1 Setting of 2-split ASE Tree

The 2-split ASE decision tree had a validation ASE of 0.057833, which was lower than the 0.059285 of the largest or Maximal tree. Therefore, it was a better decision tree compared to the maximal tree.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_NOBS_	Sum of Frequen...	20593	20595	.
y	y	_MISC_	Misclassificatio...	0.086	0.083515	.
y	y	_MAX_	Maximum Absol...	0.996065	1	.
y	y	_SSE_	Sum of Squared...	2393.212	2382.157	.
y	y	_ASE_	Average Square...	0.058107	0.057833	.
y	y	_RASE_	Root Average S...	0.241055	0.240486	.
y	y	_DIV_	Divisor for ASE	41186	41190	.
y	y	_DFT_	Total Degrees o...	20593	.	.



4.3 3-split ASE Tree

Property	Value
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

Figure 4.3.1 Setting of 3-split ASE Tree

The 3-split ASE decision tree's validation ASE is 0.056496, which is less than both the validation ASE of the Maximal tree and the 2-split ASE. Consequently, it was a more favorable decision tree compared to the other two.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_NOBS_	Sum of Frequencies	20593	20595	
y	y	_MISC_	Misclassification Rate	0.083232	0.083807	
y	y	_MAX_	Maximum Absolute Error	0.999519	1	
y	y	_SSE_	Sum of Squared Errors	2318.819	2327.054	
y	y	_ASE_	Average Squared Error	0.056301	0.056496	
y	y	_RASE_	Root Average Squared Error	0.237279	0.237688	
y	y	_DIV_	Divisor for ASE	41186	41190	
y	y	_DFT_	Total Degrees of Freedom	20593	.	

Figure 4.3.2 Fit Statistics of 3-split ASE Tree

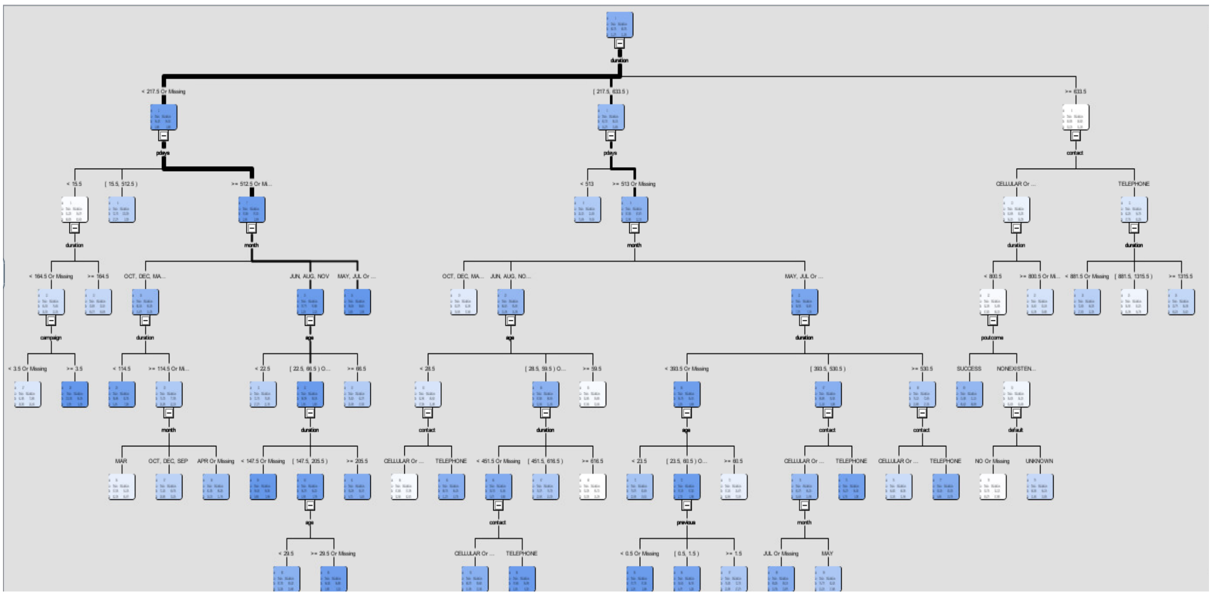


Figure 4.3.3 Result of 3-split ASE Tree

4.4 4-split ASE Tree

The 4-split ASE tree yielded improved outcomes compared to the 3-split ASE tree, with a validation ASE of 0.04739. Therefore, it surpassed the other three decision tree models used in this study. Consequently, we ceased increasing the maximum branch length and recognized the 4-split ASE decision tree model as the best ASE decision tree model.

Property	Value
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	4
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No

Figure 4.4.1 Setting of 4-split ASE Tree

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_NOBS_	Sum of Frequen...	18415	18418	.
y	y	_MISC_	Misclassificatio...	0.06511	0.068954	.
y	y	_MAX_	Maximum Absol...	0.99904	1	.
y	y	_SSE_	Sum of Squared...	1614.491	1751.139	.
y	y	_ASE_	Average Square...	0.043836	0.047539	.
y	y	_RASE_	Root Average S...	0.209371	0.218034	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_DFT_	Total Degrees o...	18415	.	.

Figure 4.4.2 Fit statistics of 4-split ASE Tree

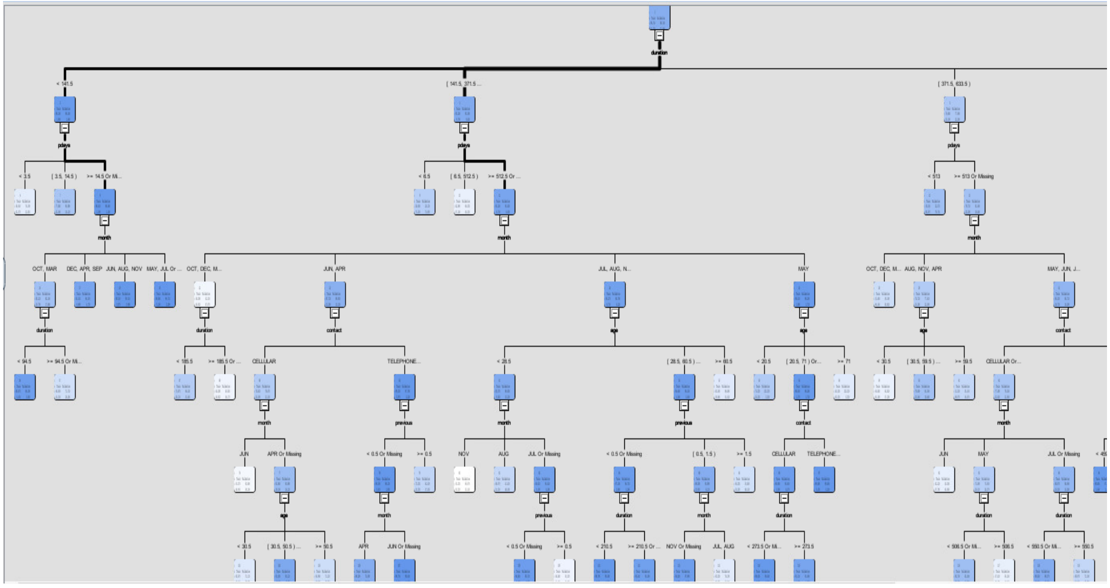


Figure 4.4.3 Result of 4-split ASE Tree

4.5 2-split Misclassification Tree

Property	Value
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
plit Search	
se Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	

Figure 4.5.1 Setting of 2-split Misclassification Tree

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_NOBS_	Sum of Frequen...	18415	18418	.
y	y	_MISC_	Misclassificatio...	0.066468	0.069932	.
y	y	_MAX_	Maximum Absol...	0.988197	0.988197	.
y	y	_SSE_	Sum of Squared...	1805.411	1859.911	.
y	y	_ASE_	Average Square...	0.04902	0.050492	.
y	y	_RASE_	Root Average S...	0.221405	0.224703	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_DFT_	Total Degrees o...	18415	.	.

Output

```
1  *------------------------------------------------------------*
2  User:                    301297765
```

Figure 4.5.2 Fit Statistics of 2-Split Misclassification Tree

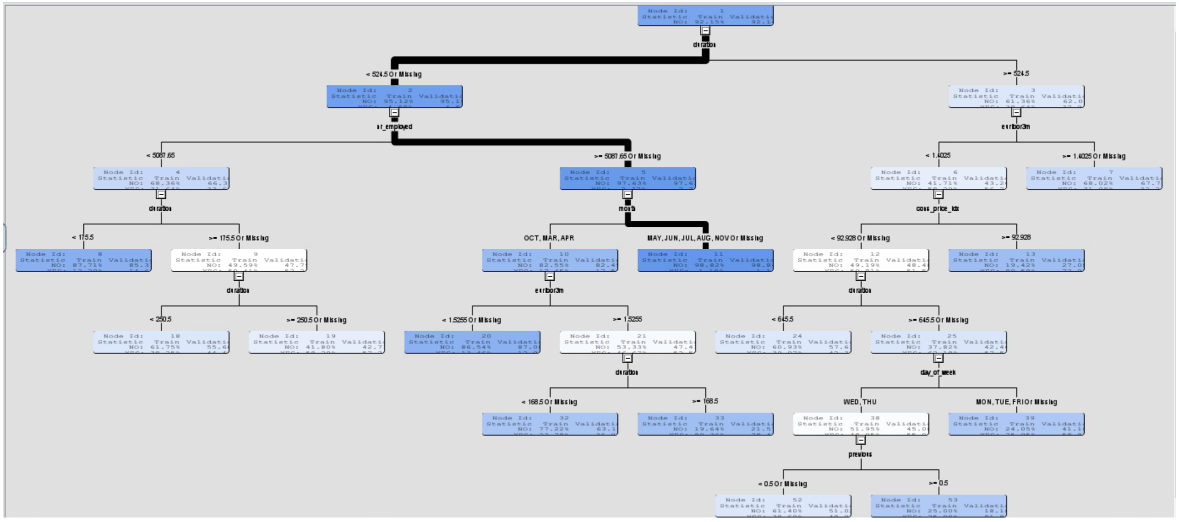


Figure 4.5.3 Result of 2-Split Misclassification Tree

The projected outcome percentages for leaves 1, 2, and 3 are more than the observed result percentages, indicating that the train model outperformed the validation model.

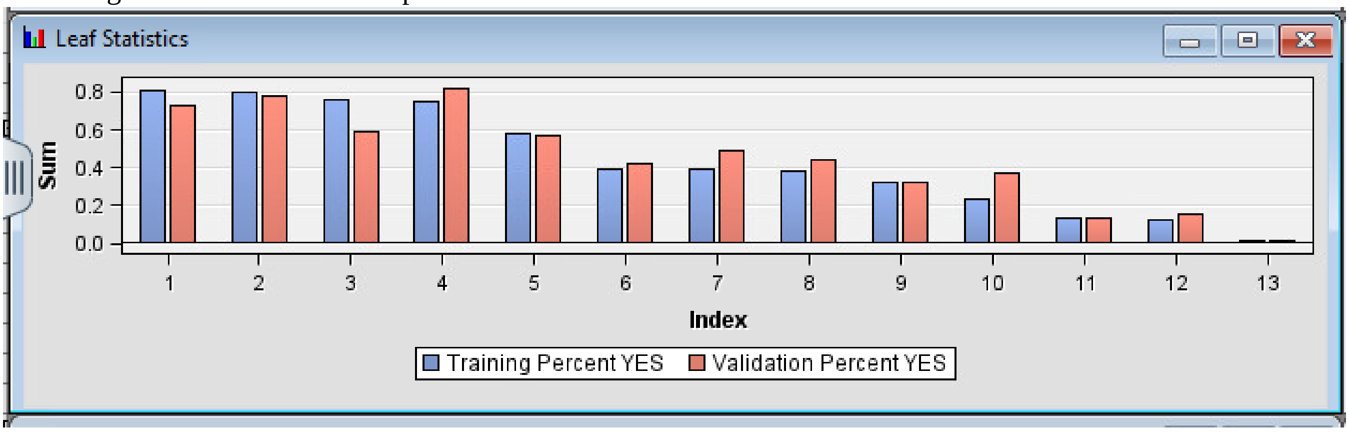


Figure 4.5.4 Leaf Statistics of 2-split Misclassification Tree

4.6 Decision Tree Summary

According to the above result, we observe that the best decision tree model was the 4-split decision tree. As 2 ways of assessment by ASE or misclassification gave the same result, we would be quite assured to go with the 4-split decision tree with lowest Average Squared Error.

Decision or Probability Tree	ASE
Decision Tree with 4 splits, ASE	0.043836
Decision Tree with 2 splits, Misclassification	0.049022
Decision Tree with 3 splits, ASE	0.056301
Decision Tree with 2 splits, ASE	0.058107
Decision Tree with Maximal, ASE	0.060695

Figure 4.6.1 Summary of Table of Decision Tree

Based on the decision tree models that were tested, the following observations were made:

- Campaigns with longer durations to customers, Longer the duration of the call, the better the chances of getting customers to sign up for a deposit.
- Maximize deposit campaign offerings during Quarter 1 (January, February, March), Quarter 3 (July, August, September), and Quarter 4 (October, November, December).
- The analysis indicates that customers who were successfully engaged in a previous deposit campaign are more likely to engage in subsequent campaigns.
- Customers who do not have home loan instalments, as the analysis suggests that customers without home loan instalments are more likely to engage with deposit offerings.

5. Logistic Regression

5.1 Data Massaging

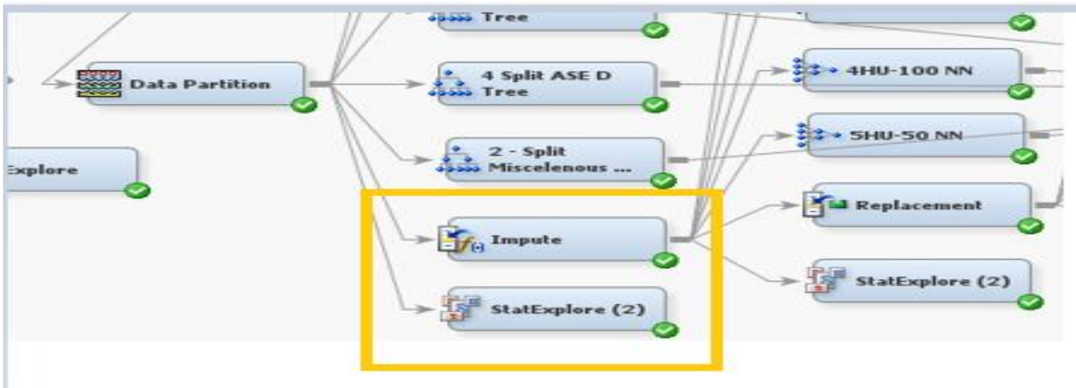


Figure 5.1.1. The flow of Data Massaging

5.1.1. Data Imputation

Following the decision tree, another predictive model would be utilized to select the best model to predict Bank deposit consumers, which is regression analysis. Before running regression, it was necessary to confirm that the data met the requirements for testing regression models. The first requirement to be addressed was that the dataset has to be free of missing data. Therefore, the dataset will pass through an impute node as shown above.

After the Data Partition node, values that are missing in the dataset are filled by an Impute node. To handle or replace missing values of class variables in the dataset, the impute node employed the mode or the highest frequency observation. However, it replaced missing values of interval variables with the variable's average or mean, as illustrated in the figure below.

```
Output




Limits and Replacement Values for Interval Variables

                                                     Lower                     Upper
                Replace                 Lower     Replacement      Upper     Replacemen
Variable        Variable                limit        Value         Limit        Value

age             REP_age                 8.734        8.734         71.18        71.18
campaign        REP_campaign           -5.674       -5.674         10.79        10.79
duration        REP_duration         -496.358     -496.358       1006.96      1006.96
pdays           REP_pdays             413.824      413.824       1513.98      1513.98
previous        REP_previous           -1.311       -1.311          1.66         1.66


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*




Replacement Counts

Obs     Variable      Label         Role      Train     Validation

 1      age           age           INPUT      172         197
 2      campaign      campaign      INPUT      459         410
 3      duration      duration      INPUT      437         507
 4      pdays         pdays         INPUT      728         787
 5      previous      previous      INPUT      525         539

```

Figure 5.1.1.1 Output of Impute Node

In the flow, we introduced a replacement node and connected to the Impute node. This replacement node sorted out the interval variable outliers with values greater than or less than three standard deviations from the mean. Figure depicts results of the dataset replacement.

Total Replacement Counts

Variable	Label	Role	Train	Validation
age	age	INPUT	172	197
campaign	campaign	INPUT	459	410
duration	duration	INPUT	437	507
pdays	pdays	INPUT	728	787
previous	previous	INPUT	525	539

Output

```
*------------------------------------------------------------*
User:                    301297765
Date:                    August 18, 2023
Time:                    14:40:58
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

             Measurement    Frequency
 Role           Level         Count

INPUT        INTERVAL           5
```

Interval Variables

Variable	Replace Variable	Limits Method	Lower limit	Upper Limit	Label	Replacement Method	Lower Replacement Value	Upper Replacement Value
age	REP_age	STDDEV	8.734434	71.18233	age	COMPUTED	8.734434	71.18233
campaign	REP_campaign	STDDEV	-5.67384	10.78528	campaign	COMPUTED	-5.67384	10.78528
duration	REP_duration	STDDEV	-496.358	1006.962	duration	COMPUTED	-496.358	1006.962
pdays	REP_pdays	STDDEV	413.8244	1513.975	pdays	COMPUTED	413.8244	1513.975
previous	REP_previous	STDDEV	-1.31071	1.661703	previous	COMPUTED	-1.31071	1.661703

Figure 5.1.1.2. Replacement Count

As seen in the Total Replacement Counts section above, some outliers were replaced. The heavily skewed values in the variables were replaced by this process.

After the Replacement, a StatExplore node was attached to the Replacement node to check the data skewness.

5.2 Full Regression

Similar to how forward and backward regression worked to choose or remove the variables, full regression did not work in this way. To find significant variables, one had to manually examine the p-value. The information below indicates that variables are significant (p-value 0.05) under the condition "Pr > ChiSq" as follows:

```

            Type 3 Analysis of Effects

                                  Wald
Effect            DF        Chi-Square    Pr > ChiSq

REP_age            1            0.0030        0.9564
REP_campaign       1           14.7498        0.0001
REP_duration       1         2396.4417        <.0001
REP_pdays          1           23.5443        <.0001
REP_previous       1            1.4107        0.2349
contact            1          195.1356        <.0001
day_of_week        4            9.3193        0.0536
default            2           50.1351        <.0001
education          7           12.9237        0.0740
housing            2            0.5222        0.7702
job               11           88.1736        <.0001
loan               1            2.8029        0.0941
marital            3            3.7067        0.2949
month              9          648.4031        <.0001
poutcome           2            7.5393        0.0231


                                              Analysis of Maximum Likelihood Estimates
```

Figure 5.2.1. The output of Type 3 Analysis of Effects

The odd ratios were noted to determine the degree to which various parameters were correlated with the likelihood that an event would occur. In this situation, the ratio shows that the percentage likelihood of receiving a bank deposit would alter depending on whether the unit of the specific variables increased or decreased.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	AIC	Akaike's Information Criterion	6095.834	.	.
y	y	_ASE_	Average Squared Error	0.049865	0.051774	.
y	y	_AVERR_	Average Error Function	0.163069	0.166567	.
y	y	_DFE_	Degrees of Freedom for Error	18370	.	.
y	y	_DFM_	Model Degrees of Freedom	45	.	.
y	y	_DFT_	Total Degrees of Freedom	18415	.	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_ERR_	Error Function	6005.834	6135.658	.
y	y	_FPE_	Final Prediction Error	0.05011	.	.
y	y	_MAX_	Maximum Absolute Error	0.995469	0.997213	.
y	y	_MSE_	Mean Square Error	0.049987	0.051774	.
y	y	_NOBS_	Sum of Frequencies	18415	18418	.
y	y	_NW_	Number of Estimate Weights	45	.	.
y	y	_RASE_	Root Average Sum of Squares	0.223305	0.227539	.
y	y	_RFPE_	Root Final Prediction Error	0.223852	.	.
y	y	_RMSE_	Root Mean Squared Error	0.223579	0.227539	.
y	y	_SBC_	Schwarz's Bayesian Criterion	6447.775	.	.
y	y	_SSE_	Sum of Squared Errors	1836.537	1907.143	.
y	y	_SUMW_	Sum of Case Weights Times Freq	36830	36836	.
y	y	_MISC_	Misclassification Rate	0.07016	0.074655	.

Figure 5.2.2. Fit statistics of Type 3 Analysis

Odds Ratio Estimates

Effect		y	Point Estimate
REP_age		yes	1.000
REP_campaign		yes	0.936
REP_duration		yes	1.005
REP_pdays		yes	0.997
REP_previous		yes	1.305
contact	cellular vs telephone	yes	3.340
day_of_week	fri vs wed	yes	0.956
day_of_week	mon vs wed	yes	0.845
day_of_week	thu vs wed	yes	0.959
day_of_week	tue vs wed	yes	1.102
default	no vs yes	yes	6.697
default	unknown vs yes	yes	3.537
education	basic.4y vs unknown	yes	0.899
education	basic.6y vs unknown	yes	0.875
education	basic.9y vs unknown	yes	0.729
education	high.school vs unknown	yes	0.830
education	illiterate vs unknown	yes	1.963
education	professional.course vs unknown	yes	0.706
education	university.degree vs unknown	yes	0.930
housing	no vs yes	yes	1.012
housing	unknown vs yes	yes	0.872
job	admin. vs unknown	yes	0.621
job	blue-collar vs unknown	yes	0.435
job	entrepreneur vs unknown	yes	0.418
job	housemaid vs unknown	yes	0.612

Figure 5.2.3. Output - Odds Ratio Estimates (1)

job	management vs unknown	yes	0.475
job	retired vs unknown	yes	1.128
job	self-employed vs unknown	yes	0.648
job	services vs unknown	yes	0.469
job	student vs unknown	yes	1.212
job	technician vs unknown	yes	0.634
job	unemployed vs unknown	yes	0.605
loan	no vs yes	yes	1.144
loan	unknown vs yes	yes	.
marital	divorced vs unknown	yes	1.363
marital	married vs unknown	yes	1.316
marital	single vs unknown	yes	1.501
month	apr vs sep	yes	0.637
month	aug vs sep	yes	0.269
month	dec vs sep	yes	1.801
month	jul vs sep	yes	0.201
month	jun vs sep	yes	0.688
month	mar vs sep	yes	3.416
month	may vs sep	yes	0.241
month	nov vs sep	yes	0.253
month	oct vs sep	yes	1.716
poutcome	failure vs success	yes	0.404
poutcome	nonexistent vs success	yes	0.554

```
*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*
```

2023-05-10 -

Figure 5.2.4. Output – Odds Ratio Estimates (2)

5.3. Forward Regression

After running this model, it was discovered that the significant variables seen in the forward regression were chosen using the p-value indicated below "Pr > ChiSq", and there are 7 variables included in this.

```
NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

                         Summary of Forward Selection

             Effect                  Number        Score                     Validation
    Step     Entered          DF         In   Chi-Square    Pr > ChiSq       Error Rate

       1     REP_duration      1          1    3563.3713        <.0001          11647.1
       2     REP_pdays         1          2    2101.9078        <.0001          10319.5
       3     month             9          3    1111.6893        <.0001           9526.7
       4     contact           1          4     251.9339        <.0001           9317.2
       5     job              11          5     128.1704        <.0001           9178.7
       6     default           2          6      54.1201        <.0001           9106.1
       7     REP_campaign      1          7      15.4872        <.0001           9073.0
       8     poutcome          2          8       6.6864        0.0353           9063.2


The selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:

Intercept  REP_campaign  REP_duration  REP_pdays  contact  default  job  month  poutcome
```

Figure 5.3.1. Output - Summary of Forward Selection

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	AIC	Akaike's Information Criterion	6135.526	.	.
y	y	_ASE_	Average Squared Error	0.050514	0.051874	.
y	y	_AVERR_	Average Error Function	0.165179	0.16706	.
y	y	_DFE_	Degrees of Freedom for Error	18389	.	.
y	y	_DFM_	Model Degrees of Freedom	26	.	.
y	y	_DFT_	Total Degrees of Freedom	18415	.	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_ERR_	Error Function	6083.526	6153.828	.
y	y	_FPE_	Final Prediction Error	0.050657	.	.
y	y	_MAX_	Maximum Absolute Error	0.991973	0.996631	.
y	y	_MSE_	Mean Square Error	0.050585	0.051874	.
y	y	_NOBS_	Sum of Frequencies	18415	18418	.
y	y	_NW_	Number of Estimate Weights	26	.	.
y	y	_RASE_	Root Average Sum of Squares	0.224753	0.227759	.
y	y	_RFPE_	Root Final Prediction Error	0.225071	.	.
y	y	_RMSE_	Root Mean Squared Error	0.224912	0.227759	.
y	y	_SBC_	Schwarz's Bayesian Criterion	6338.87	.	.
y	y	_SSE_	Sum of Squared Errors	1860.429	1910.83	.
y	y	_SUMW_	Sum of Case Weights Times Freq	36830	36836	.
y	y	_MISC_	Misclassification Rate	0.071301	0.074221	.

Figure 5.3.2. Fit Statistics of Forward Regression

However, the model selection step number is equal to 5, which suggests five variables are enough to run this forward regression model, according to the Iteration Plot with the Average Square Error (ASE).

```
              Type 3 Analysis of Effects

                                   Wald
Effect                 DF    Chi-Square    Pr > ChiSq

REP_duration            1     1898.0512        <.0001


                                     Analysis of Maximum Likelihood Estimates

                                                      Standard        Wald                  Standardized
Parameter                      y     DF    Estimate      Error    Chi-Square    Pr > ChiSq      Estimate    Exp(Est)

Intercept                      yes    1     -4.2845     0.0610       4933.09        <.0001                     0.014
REP_duration                   yes    1     0.00567   0.000130       1898.05        <.0001        0.5685       1.006

                         Odds Ratio Estimates

                                                          Point
Effect                                            y    Estimate

REP_duration                                      yes     1.006


Step  2: Effect REP_nr_employed entered.
```

Figure 5.3.3. Output – Odds Ratio Estimates

5.4. Backward Regression

12 variables were taken out of the procedure after the backward regression was run on the data. All the variables that were excluded had high p-values, as can be seen under "Pr > ChiSq".

```
NOTE: No (additional) effects met the 0.05 significance level for removal from the model.


                       Summary of Backward Elimination

              Effect                  Number        Wald                   Validation
     Step     Removed          DF         In    Chi-Square    Pr > ChiSq    Error Rate

        1     REP_age           1         14        0.0030        0.9564        9048.8
        2     housing           2         13        0.5234        0.7698        9048.3
        3     marital           3         12        4.1150        0.2493        9053.0
        4     REP_previous      1         11        1.4893        0.2223        9062.3
        5     loan              2         10        3.7295        0.1549        9058.0
        6     education         7          9       13.7338        0.0561        9059.8
        7     day_of_week       4          8        9.0519        0.0598        9063.2


The selected model, based on the error rate for the validation data, is the model trained in Step 2. It consists of the fol

Intercept  REP_campaign  REP_duration  REP_pdays  REP_previous  contact  day_of_week  default  education  job  loan  marita
```

Figure 5.4.1. Output - Summary of Backward Elimination

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	AIC	Akaike's Information Criterion	6095.834	.	.
y	y	_ASE_	Average Squared Error	0.049865	0.051774	.
y	y	_AVERR_	Average Error Function	0.163069	0.166567	.
y	y	_DFE_	Degrees of Freedom for Error	18370	.	.
y	y	_DFM_	Model Degrees of Freedom	45	.	.
y	y	_DFT_	Total Degrees of Freedom	18415	.	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_ERR_	Error Function	6005.834	6135.658	.
y	y	_FPE_	Final Prediction Error	0.05011	.	.
y	y	_MAX_	Maximum Absolute Error	0.995469	0.997213	.
y	y	_MSE_	Mean Square Error	0.049987	0.051774	.
y	y	_NOBS_	Sum of Frequencies	18415	18418	.
y	y	_NW_	Number of Estimate Weights	45	.	.
y	y	_RASE_	Root Average Sum of Squares	0.223305	0.227539	.
y	y	_RFPE_	Root Final Prediction Error	0.223852	.	.
y	y	_RMSE_	Root Mean Squared Error	0.223579	0.227539	.
y	y	_SBC_	Schwarz's Bayesian Criterion	6447.775	.	.
y	y	_SSE_	Sum of Squared Errors	1836.537	1907.143	.
y	y	_SUMW_	Sum of Case Weights Times Freq	36830	36836	.
y	y	_MISC_	Misclassification Rate	0.07016	0.074655	.

Figure 5.4.2. Fit Statistics – Backward Regression

```
                                Odds Ratio Estimates

                                                                         Point
Effect                                                          y     Estimate

REP_campaign                                                    yes      0.936
REP_duration                                                    yes      1.005
REP_pdays                                                       yes      0.997
REP_previous                                                    yes      1.305
contact        cellular vs telephone                            yes      3.337
day_of_week    fri vs wed                                       yes      0.957
day_of_week    mon vs wed                                       yes      0.846
day_of_week    thu vs wed                                       yes      0.960
day_of_week    tue vs wed                                       yes      1.102
default        no vs yes                                        yes      6.720
default        unknown vs yes                                   yes      3.552
education      basic.4y vs unknown                              yes      0.899
education      basic.6y vs unknown                              yes      0.874
education      basic.9y vs unknown                              yes      0.729
education      high.school vs unknown                           yes      0.830
education      illiterate vs unknown                            yes      1.963
education      professional.course vs unknown                   yes      0.705
education      university.degree vs unknown                     yes      0.930
job            admin. vs unknown                                yes      0.620
job            blue-collar vs unknown                           yes      0.435
job            entrepreneur vs unknown                          yes      0.418
job            housemaid vs unknown                             yes      0.613
job            management vs unknown                            yes      0.475
job            retired vs unknown                               yes      1.132
job            self-employed vs unknown                         yes      0.648
job            services vs unknown                              yes      0.469
job            student vs unknown                               yes      1.210
job            technician vs unknown                            yes      0.633
job            unemployed vs unknown                            yes      0.605
```

```
loan        no vs yes                          yes     1.145
loan        unknown vs yes                     yes     0.928
marital     divorced vs unknown                yes     1.362
marital     married vs unknown                 yes     1.315
marital     single vs unknown                  yes     1.498
month       apr vs sep                         yes     0.637
month       aug vs sep                         yes     0.269
month       dec vs sep                         yes     1.801
month       jul vs sep                         yes     0.201
month       jun vs sep                         yes     0.688
month       mar vs sep                         yes     3.411
month       may vs sep                         yes     0.241
month       nov vs sep                         yes     0.253
month       oct vs sep                         yes     1.716
poutcome    failure vs success                 yes     0.405
poutcome    nonexistent vs success             yes     0.555


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*
```

Figure 5.4.3 Output – Odds Ratio Estimates

5.5. Stepwise Regression

Given the significant variables are the same, stepwise regression yields exactly the same results as forwarding regression.

```


                         Summary of Stepwise Selection

            Effect                  Number        Score          Wald                    Validation
    Step    Entered          DF         In     Chi-Square    Chi-Square    Pr > ChiSq    Error Rate

       1    REP_duration      1          1      3563.3713                      <.0001       11647.1
       2    REP_pdays         1          2      2101.9078                      <.0001       10319.5
       3    month             9          3      1111.6893                      <.0001        9526.7
       4    contact           1          4       251.9339                      <.0001        9317.2
       5    job              11          5       128.1704                      <.0001        9178.7
       6    default           2          6        54.1201                      <.0001        9106.1
       7    REP_campaign      1          7        15.4872                      <.0001        9073.0
       8    poutcome          2          8         6.6864                      0.0353        9063.2

```

Figure 5.5.1. Summary of Stepwise Selection

The odds ratio estimates point estimate is shown below with the outcome as follows:

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y	y	_AIC_	Akaike's Information Criterion	6135.526		
y	y	_ASE_	Average Squared Error	0.050514	0.051874	.
y	y	_AVERR_	Average Error Function	0.165179	0.16706	.
y	y	_DFE_	Degrees of Freedom for Error	18389	.	.
y	y	_DFM_	Model Degrees of Freedom	26	.	.
y	y	_DFT_	Total Degrees of Freedom	18415	.	.
y	y	_DIV_	Divisor for ASE	36830	36836	.
y	y	_ERR_	Error Function	6083.526	6153.828	.
y	y	_FPE_	Final Prediction Error	0.050657	.	.
y	y	_MAX_	Maximum Absolute Error	0.991973	0.996631	.
y	y	_MSE_	Mean Square Error	0.050585	0.051874	.
y	y	_NOBS_	Sum of Frequencies	18415	18418	.
y	y	_NW_	Number of Estimate Weights	26	.	.
y	y	_RASE_	Root Average Sum of Squares	0.224753	0.227759	.
y	y	_RFPE_	Root Final Prediction Error	0.225071	.	.
y	y	_RMSE_	Root Mean Squared Error	0.224912	0.227759	.
y	y	_SBC_	Schwarz's Bayesian Criterion	6338.87	.	.
y	y	_SSE_	Sum of Squared Errors	1860.429	1910.83	.
y	y	_SUMW_	Sum of Case Weights Times Freq	36830	36836	.
y	y	_MISC_	Misclassification Rate	0.071301	0.074221	.

Figure 5.5.2. Fit Statistics of Stepwise Selection

```
                                  Odds Ratio Estimates

                                                                        Point
Effect                                                        y      Estimate

REP_campaign                                                  yes       0.935
REP_duration                                                  yes       1.005
REP_pdays                                                     yes       0.997
contact       cellular vs telephone                           yes       3.344
default       no vs yes                                       yes     161.730
default       unknown vs yes                                  yes      85.123
job           admin. vs unknown                               yes       0.608
job           blue-collar vs unknown                          yes       0.388
job           entrepreneur vs unknown                         yes       0.391
job           housemaid vs unknown                            yes       0.576
job           management vs unknown                           yes       0.467
job           retired vs unknown                              yes       1.047
job           self-employed vs unknown                        yes       0.614
job           services vs unknown                             yes       0.431
job           student vs unknown                              yes       1.241
job           technician vs unknown                           yes       0.557
job           unemployed vs unknown                           yes       0.552
month         apr vs sep                                      yes       0.600
month         aug vs sep                                      yes       0.264
month         dec vs sep                                      yes       1.664
month         jul vs sep                                      yes       0.196
month         jun vs sep                                      yes       0.670
month         mar vs sep                                      yes       3.376
month         may vs sep                                      yes       0.232
month         nov vs sep                                      yes       0.247
month         oct vs sep                                      yes       1.676
poutcome      failure vs success                              yes       0.439
poutcome      nonexistent vs success                          yes       0.448
```

Figure 5.5.3. Output – Odds Ratio Estimates

5.6. Regression Summary

Following the replacement node, four different forms of regression (forward, backward, stepwise, and complete) were performed, and the "best" regression model was found by assessing the ASE. The regression model with the lowest ASE would be the most successful.

Although the logistic regression type is a common component of all regression models, they differ in their selection models and criteria. This is so because the forward, backward, and stepwise regression models only use the validation error criterion. In summary, it is not unexpected that either might be regarded as the "best" model as both forward regression and stepwise regression yielded the same ASE of 0.051774.

Regression Model	Selection Criterion	ASE (Validation)
The Full Regression	None	0.051774
The Forward Regression	Validation Error	0.051874
The Backward Regression	Validation Error	0.051774
The Stepwise Regression	Validation Error	0.051874

Figure 5.6.1. Summary of Logistic Regression

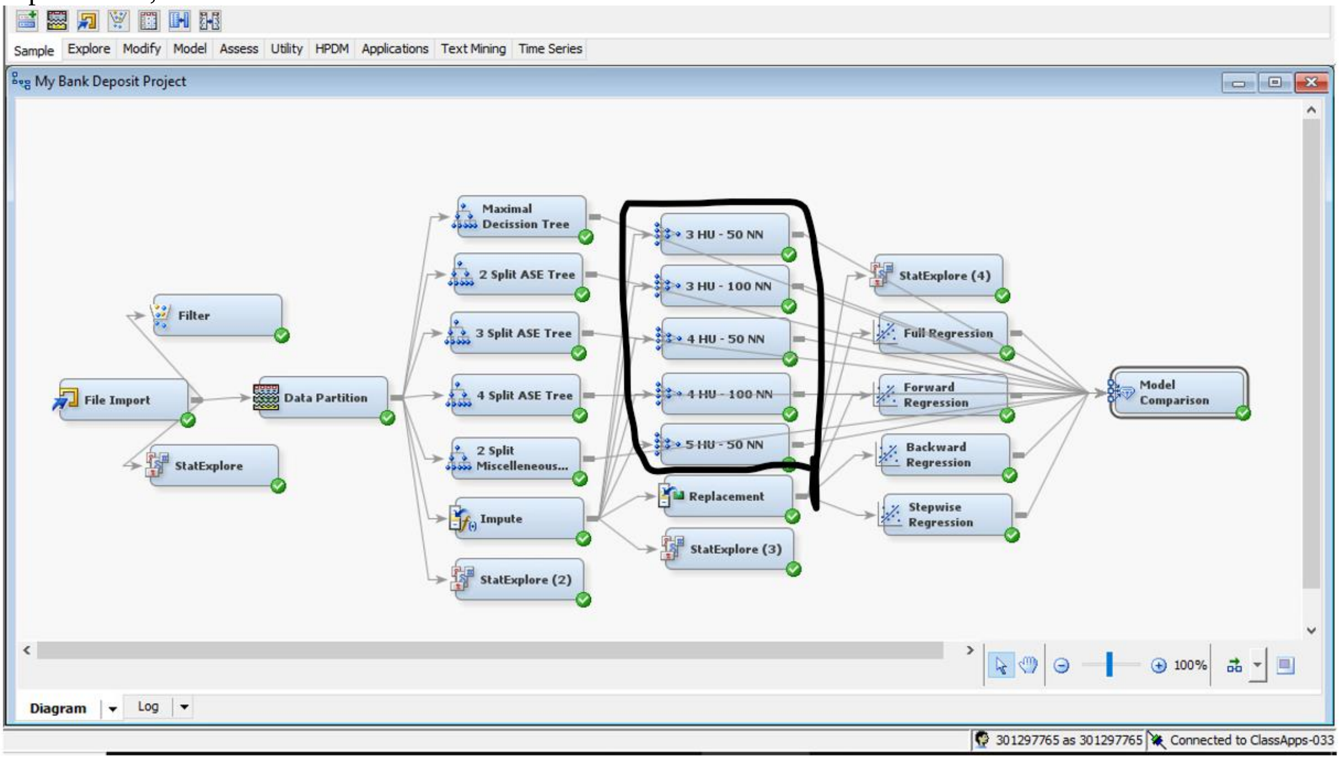
6. Neural Network

A model that can handle a wide range of nonlinear data is used to manage the unevenness of the variables. This goes beyond conventional analysis techniques and permits the adjusting of inputs to better match the findings. These models are better at performing predictive analysis since they have hidden layers

Nodes from multiple Neural networks with hidden layers and iterations will be attached to different stages of the process in order to determine which model will function best under certain circumstances. The process prevents random weight formation during initial data training by turning off specific components. The selection criteria for these models will be based on the average errors of the validation data, much like for conventional predictive models.

6.1. Neural Networks connected from Impute Node

To get the best neural network model, a total of 5 different neural network configurations were connected to the imputed data, as shown below:



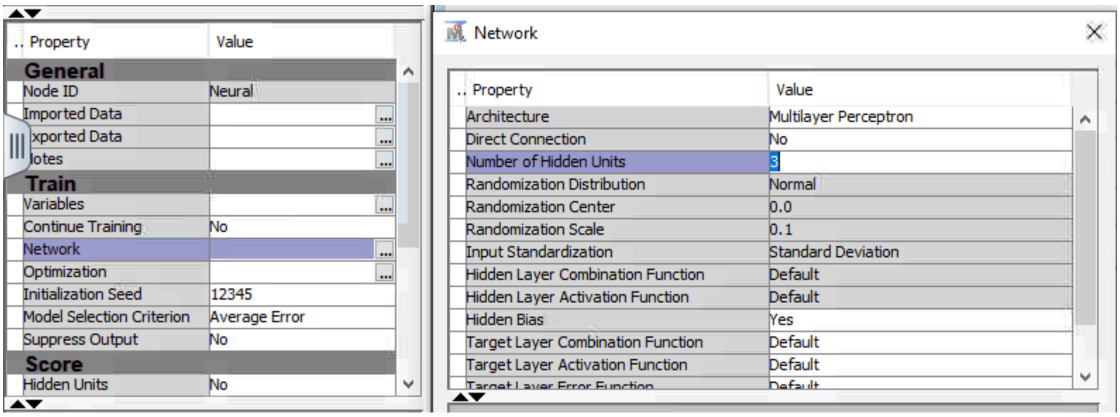


Figure 6.2.0 Setting of Neural Network with 3 Hidden Units and 50 Iterations

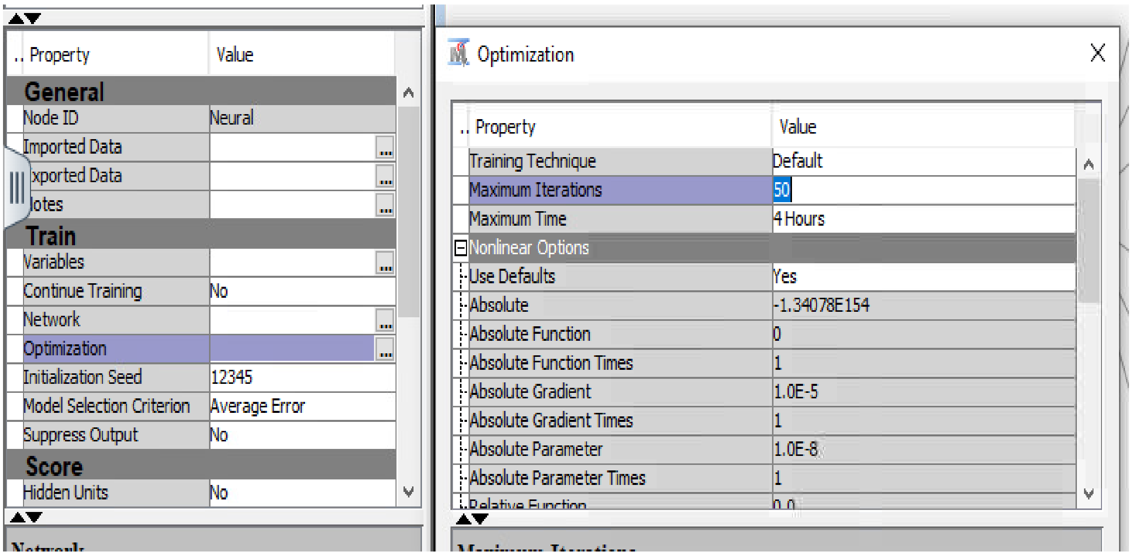


Figure 6.2.0.a Setting of Neural Network with 3 Hidden Units and 50 Iterations

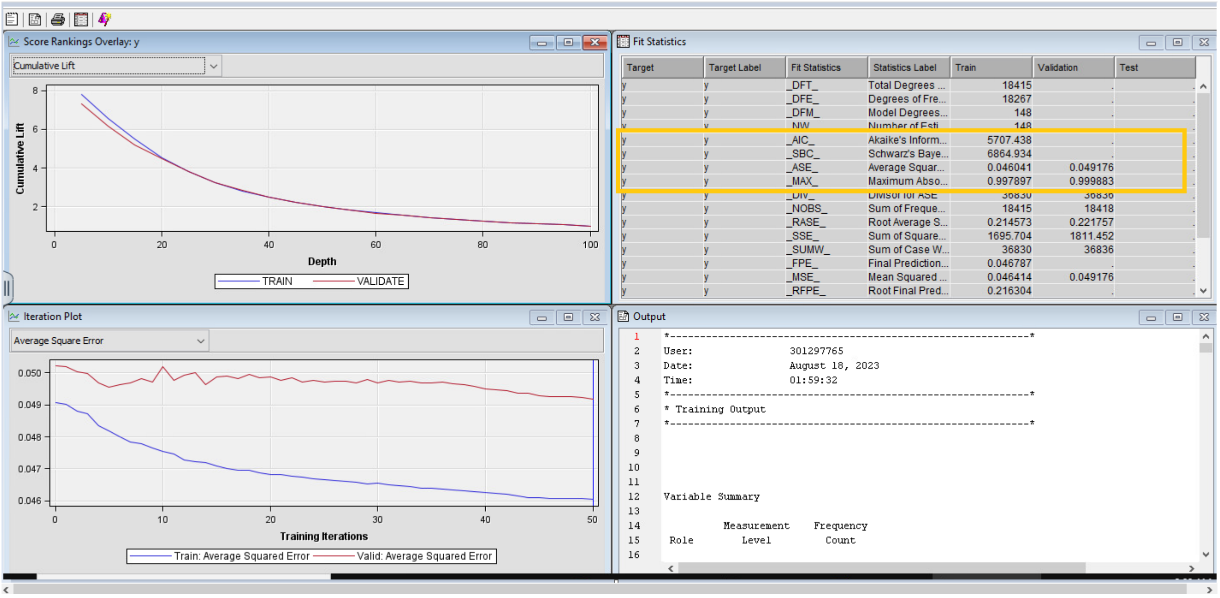


Figure 6.2.0.b Result of Neural Network with 3 Hidden Units and 50 Iterations

The first neural network with a configuration of 3 hidden units at 50 iterations had an ASE of 0.049176.

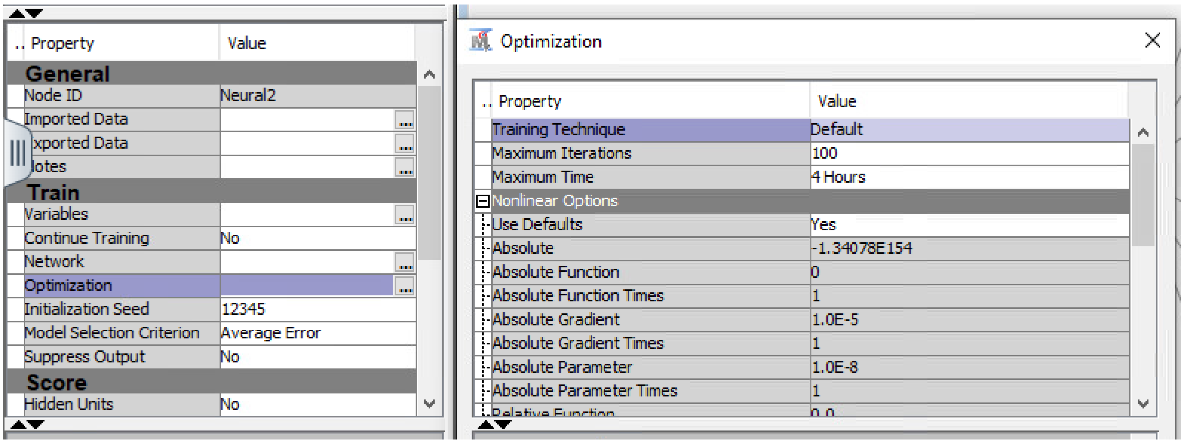


Figure 6.2.1a Setting of Neural Network with 3 Hidden Units and 100 Iterations

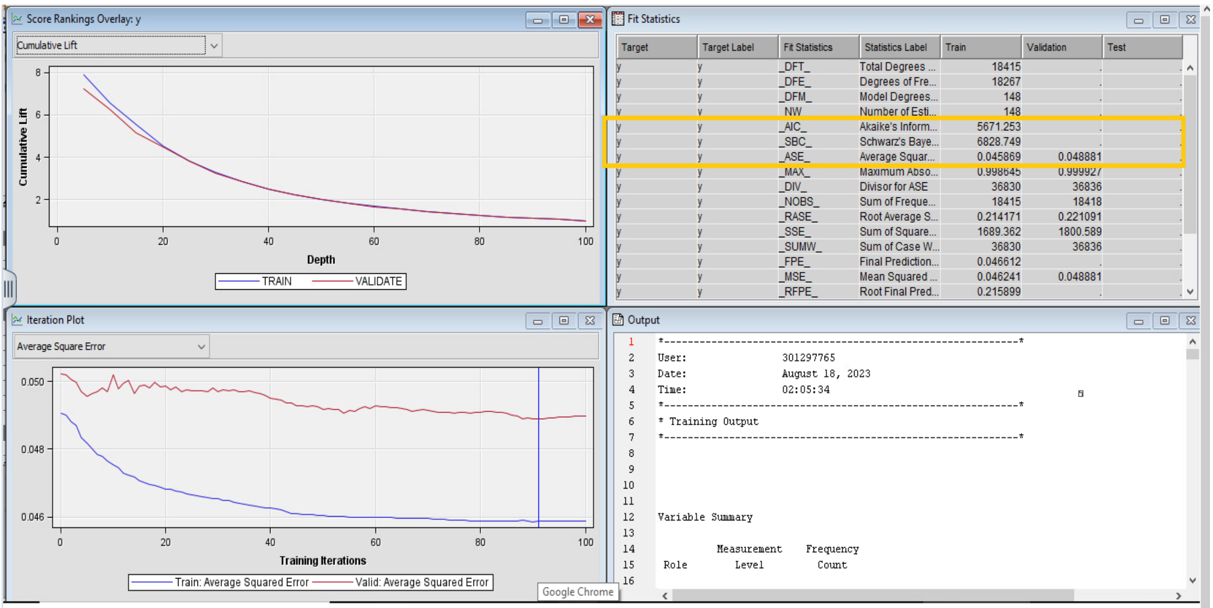


Figure 6.2.1b Result of Neural Network with 3 Hidden Units and 100 Iterations

To increase the accuracy of the model, another 50 iterations were added. Hence, a neural network with a configuration of 3 hidden units at 100 iterations was then implemented. The second neural network result was the same as the first node and had an ASE of 0.048881.

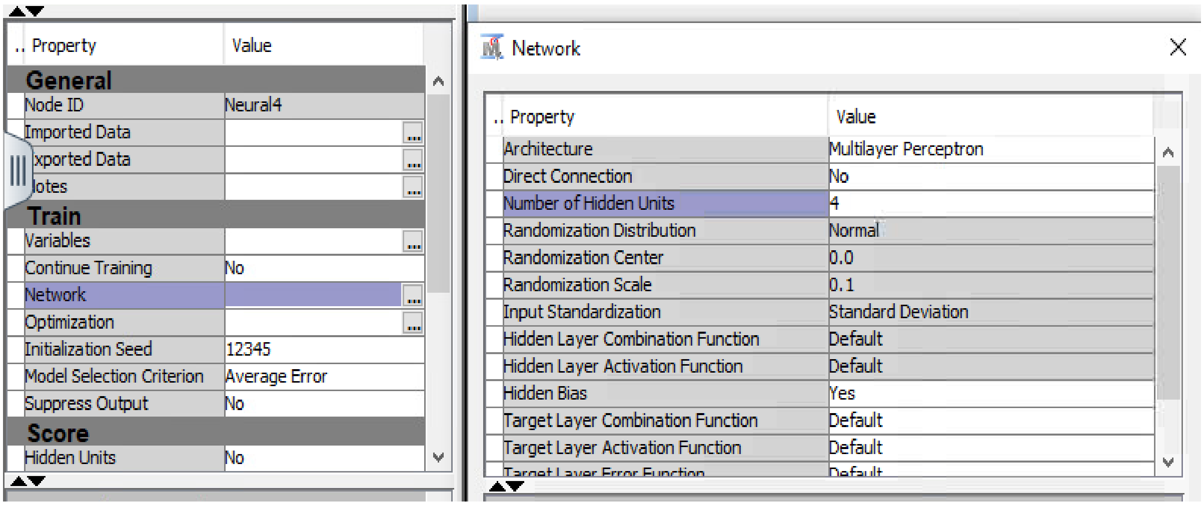


Figure 6.2.2.a Setting of Neural Network with 4 Hidden Units and 50 Iterations

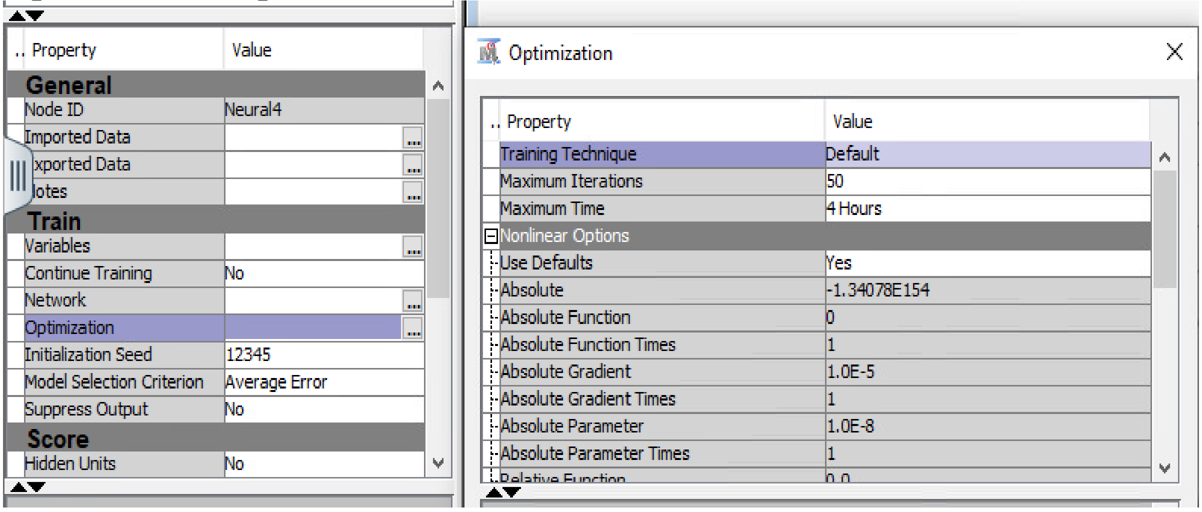


Figure 6.2.2.b Setting of Neural Network with 4 Hidden Units and 50 Iterations

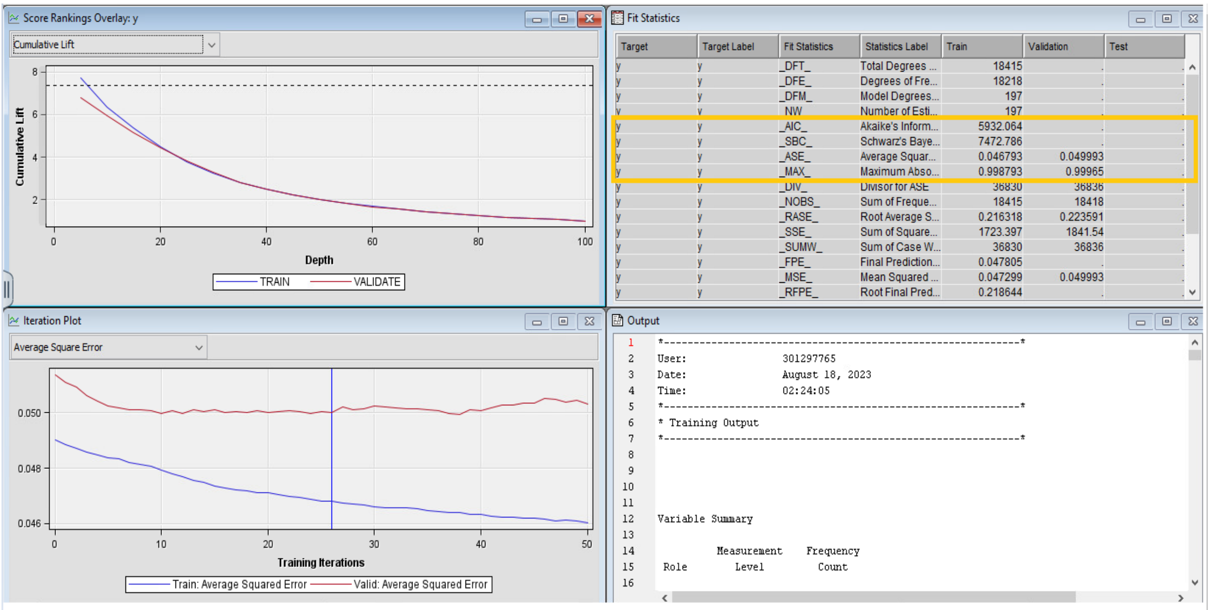


Figure 6.2.2.c Output of Neural Network with 4 Hidden Units & 50 iterations.

Results of a Neural Network with 4 Hidden Units and 50 Iterations are shown in Figure 6.1.1.c. To achieve a reduced average squared error, a third neural network is added with a configuration of four hidden units at 50 iterations. It came out to 0.049993.

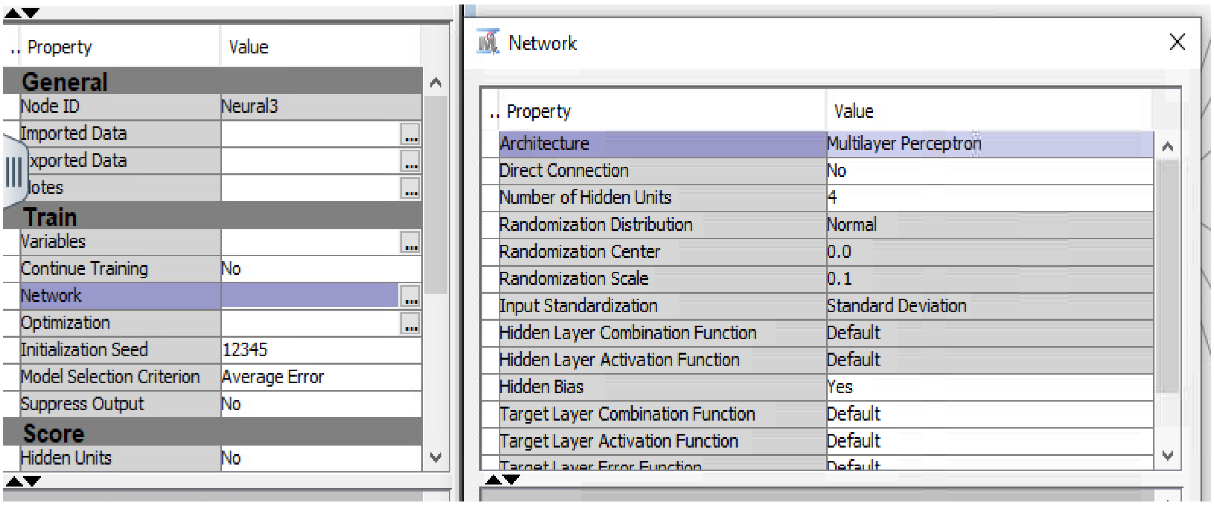


Figure 6.2.3.a Setting of Neural Network with 4 Hidden Units and 100 Iterations

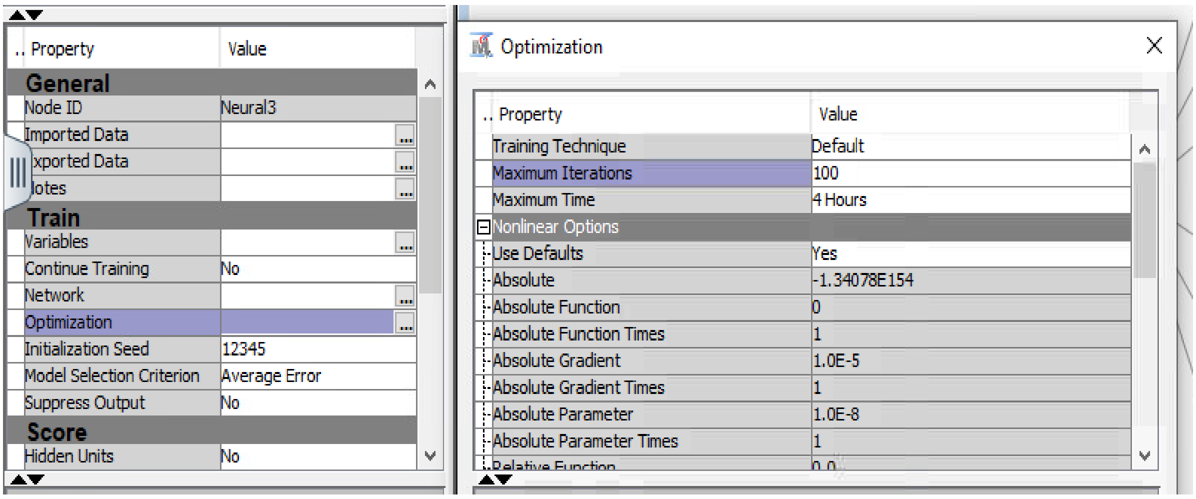


Figure 6.2.3.b Setting of Neural Network with 4 Hidden Units and 100 Iterations

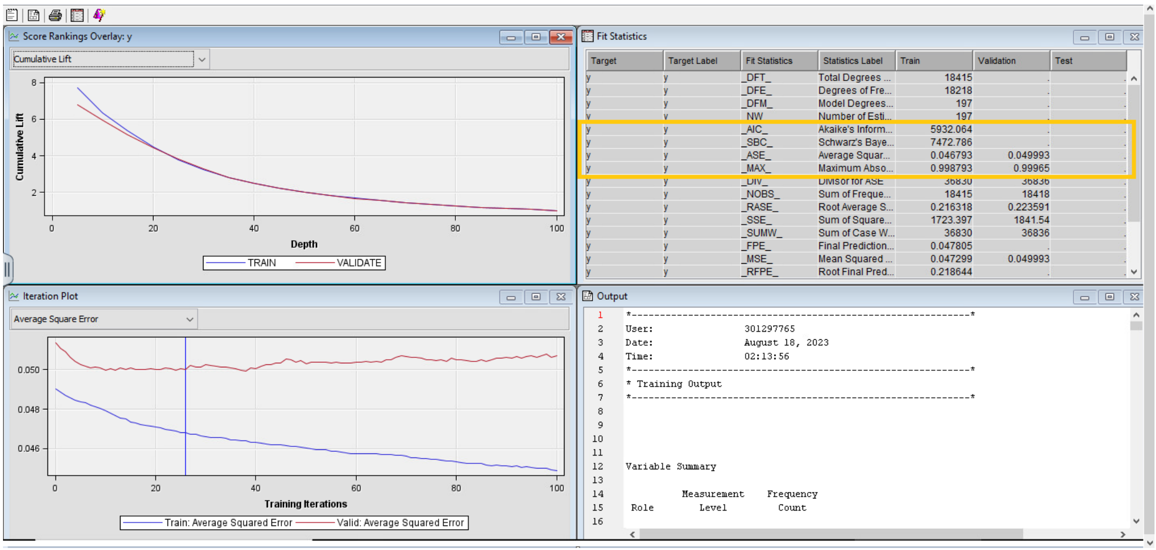


Figure 6.2.3.c. Result of Neural Network with 100 Iterations and 4 Hidden Units

The results indicated that the model required an additional 50 iterations, thus a fourth version with four hidden units was developed and run through 100 iterations. The outcome was identical to the third iteration, with a squared average error of 0.049993.

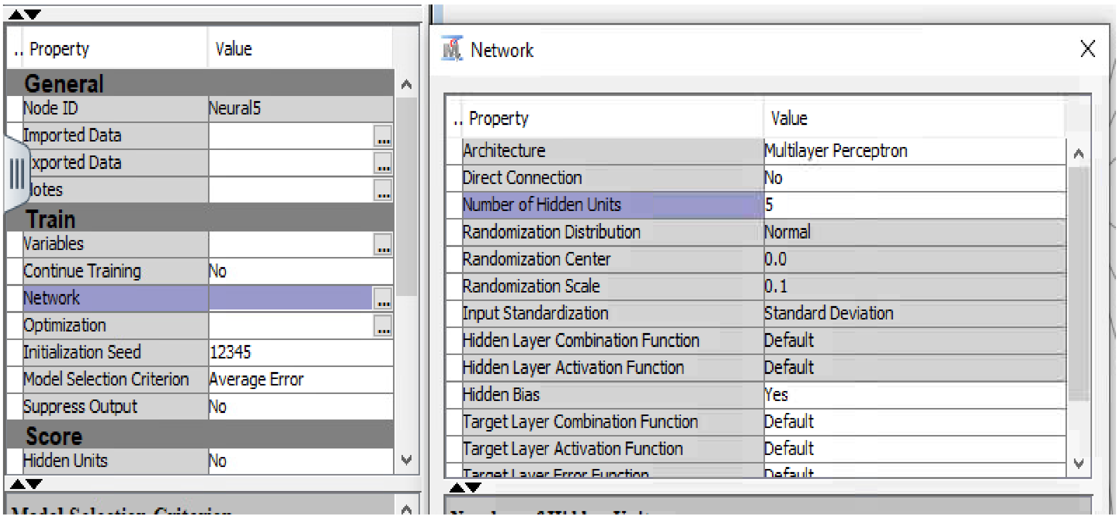


Figure 6.2.4.a Setting of Neural Network with 5 Hidden Units and 50 Iterations

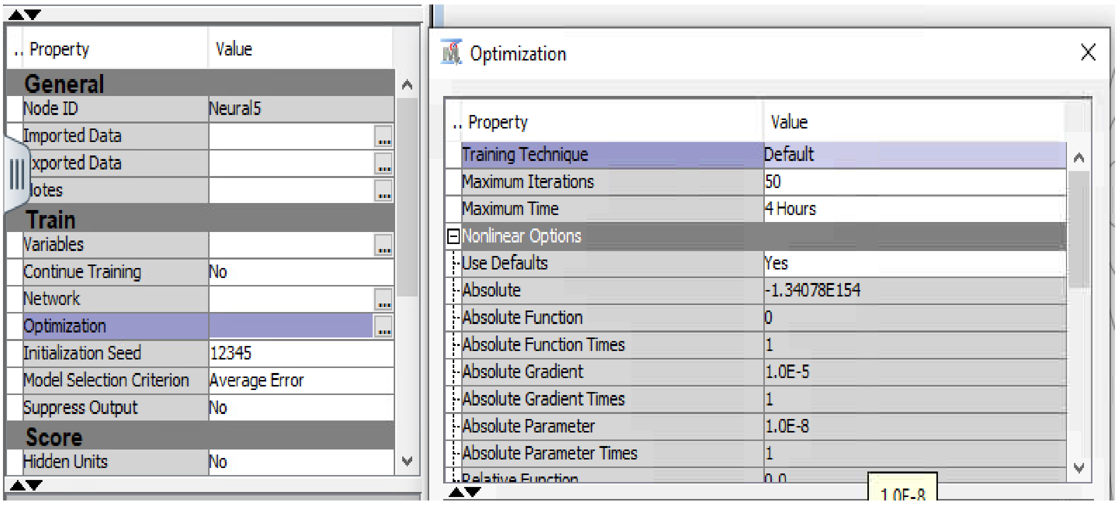


Figure 6.2.4.b Setting of Neural Network with 5 Hidden Units and 50 Iterations

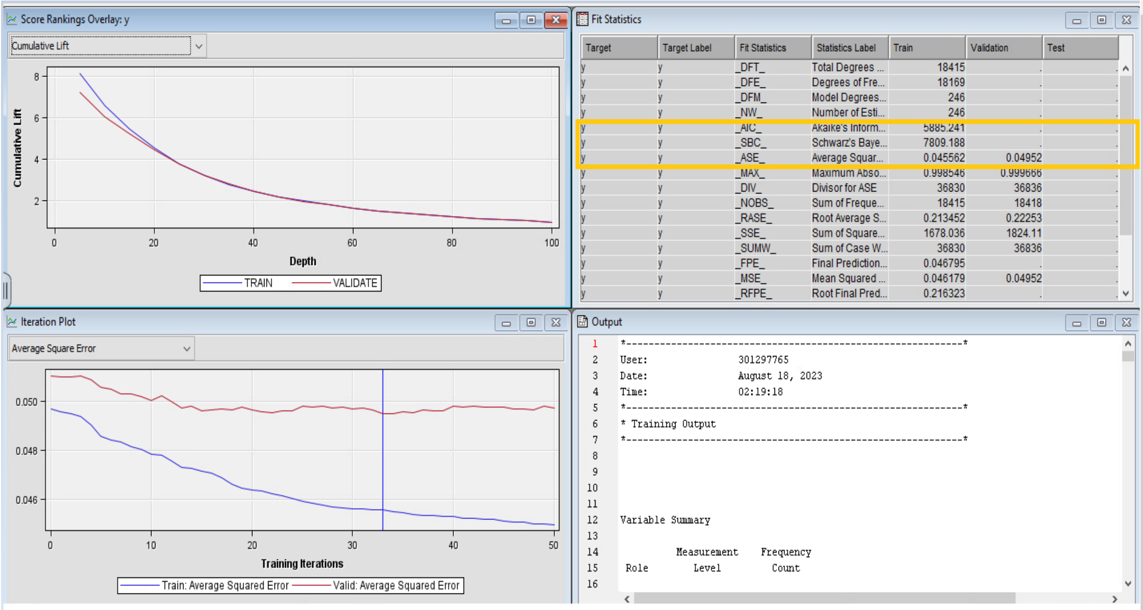


Figure 6.2.4.c Result of Neural Network with 5 Hidden Units and 50 Iterations

To try and lower the average squared error, a fifth Neural Network was developed with 5 hidden units and 50 iterations. The outcome, which was 0.04952, was higher than the previous result.

No further adjustments were required because the configuration of 5 hidden components at 50 iterations resulted in an increase in the average squared errors. With 50 iterations, the best version using the altered data had four hidden units.

6.3 Neural Network Summary.

The summary of the Neural Networks for the imputed data are indicated in the table below:

Neural Network	Hidden Units	Iterations	ASE
3 Hidden Units-50 Neural Network	3	50	0.049176
3 Hidden Units 100 Neural Network	3	100	0.048881
4 Hidden Units -50 Neural Network	4	50	0.049993
4 Hidden Units -100 Neural Network	4	100	0.049993
5 Hidden Units -50 Neural Network	5	50	0.04952

7. Model Comparison

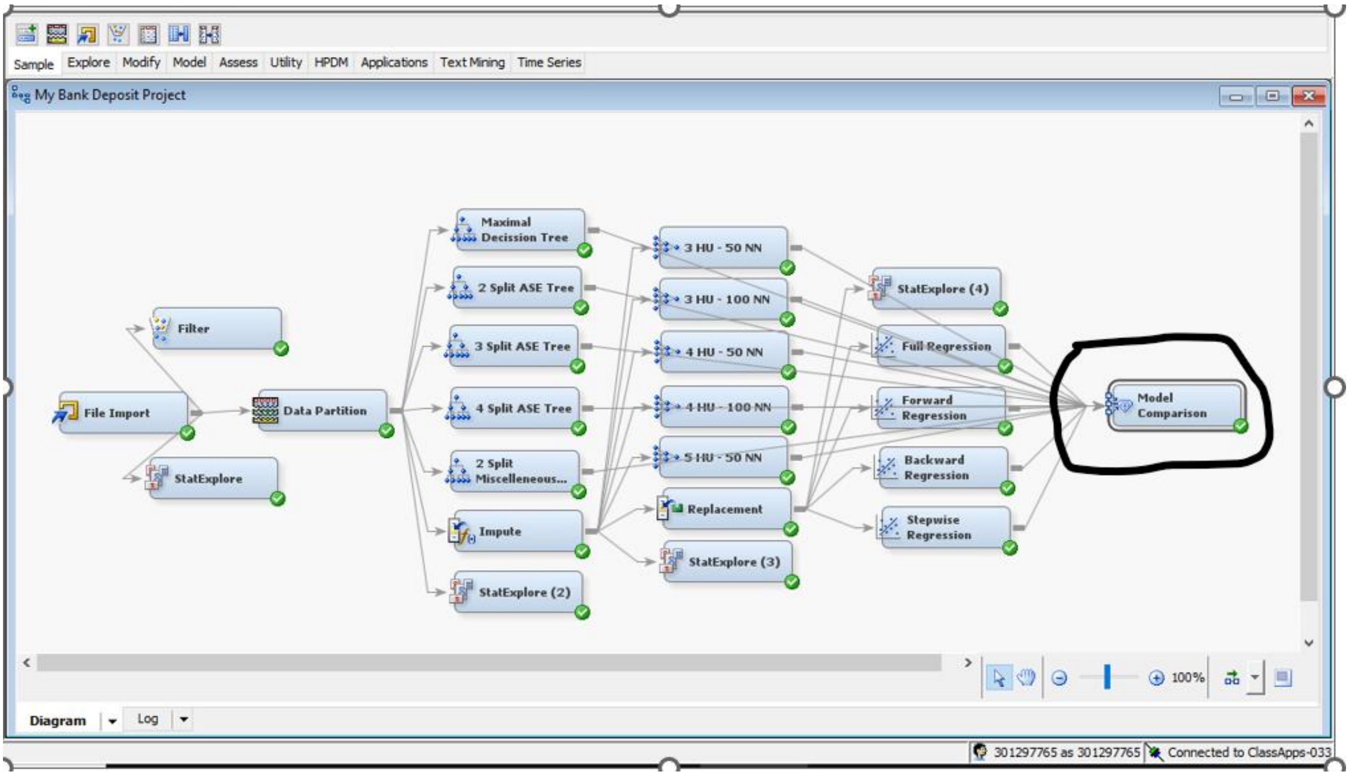


Figure 7.1. Model Comparison Node Connecting All Processed Models

All the alternative models, including decision trees, regression models, and Neural Networks connected to a Model Comparison node to determine which model was the best

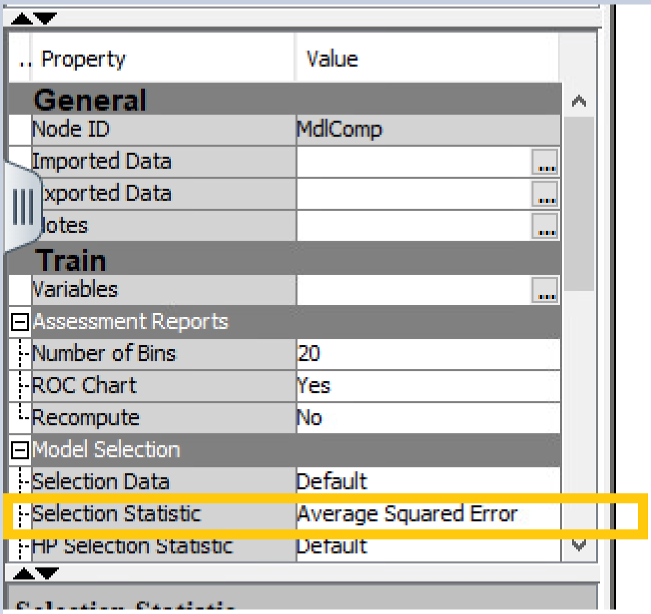
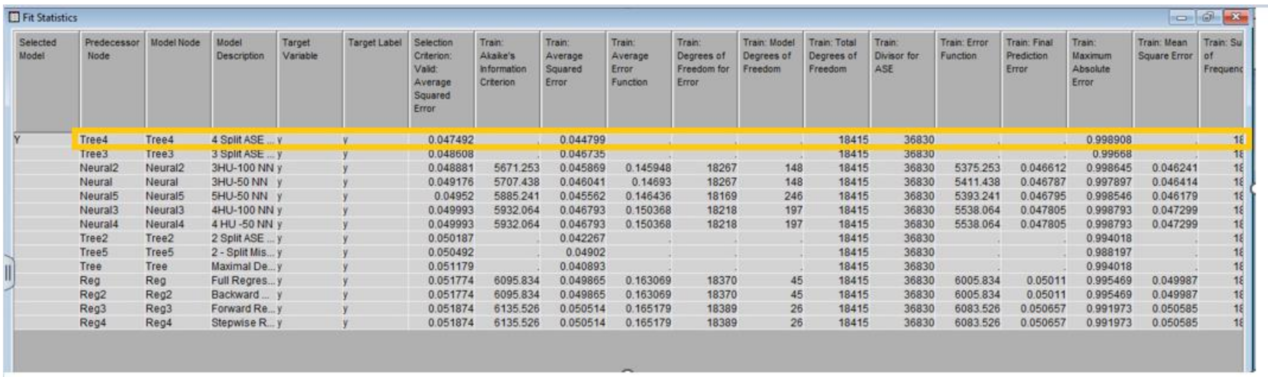


Figure 7.2. The setting of Model Comparison Node



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Su of Frequenc
Y	Tree4	Tree4	4 Split ASE ...	y	y	0.047492	.	0.044799	.	.	.	18415	36830	.	.	0.998908	.	18
	Tree3	Tree3	3 Split ASE ...	y	y	0.048608	.	0.046735	.	.	.	18415	36830	.	.	0.99668	.	18
	Neural2	Neural2	3HU-100 NN	y	y	0.048881	5671.253	0.045869	0.145948	18267	148	18415	36830	5375.253	0.046612	0.998645	0.046241	18
	Neural	Neural	3HU-50 NN	y	y	0.049176	5707.438	0.046041	0.14693	18267	148	18415	36830	5411.438	0.046787	0.997897	0.046414	18
	Neural5	Neural5	5HU-50 NN	y	y	0.04952	5885.241	0.045562	0.146436	18169	246	18415	36830	5393.241	0.046795	0.998546	0.046179	18
	Neural3	Neural3	4HU-100 NN	y	y	0.049993	5932.064	0.046793	0.150368	18218	197	18415	36830	5538.064	0.047805	0.998793	0.047299	18
	Neural4	Neural4	4 HU -50 NN	y	y	0.049993	5932.064	0.046793	0.150368	18218	197	18415	36830	5538.064	0.047805	0.998793	0.047299	18
	Tree2	Tree2	2 Split ASE ...	y	y	0.050187	.	0.042267	.	.	.	18415	36830	.	.	0.994018	.	18
	Tree5	Tree5	2 - Split Mis...	y	y	0.050492	.	0.04902	.	.	.	18415	36830	.	.	0.988197	.	18
	Tree	Tree	Maximal De...	y	y	0.051179	.	0.040893	.	.	.	18415	36830	.	.	0.994018	.	18
	Reg	Reg	Full Regres...	y	y	0.051774	6095.834	0.049865	0.163069	18370	45	18415	36830	6005.834	0.05011	0.995469	0.049987	18
	Reg2	Reg2	Backward ...	y	y	0.051774	6095.834	0.049865	0.163069	18370	45	18415	36830	6005.834	0.05011	0.995469	0.049987	18
	Reg3	Reg3	Forward Re...	y	y	0.051874	6135.526	0.050514	0.165179	18389	26	18415	36830	6083.526	0.050657	0.991973	0.050585	18
	Reg4	Reg4	Stepwise R...	y	y	0.051874	6135.526	0.050514	0.165179	18389	26	18415	36830	6083.526	0.050657	0.991973	0.050585	18

Figure 7.3 – Fit Statistics of Model Comparison Node

8. **Conclusion**

In this project, we utilized SAS Enterprise Miner 15.2 to predict potential clients for bank direct marketing campaigns utilizing several models such as Decision Trees, Regression, and Neural Networks. We used real-world and recent data from a bank, as well as several iterations, to fine-tune the prediction model findings. In practice, each iteration has proven to be quite valuable, as the resulting prediction performances have improved. The best model, as evidenced by ASE [Average Square error], demonstrated great predictive performance. We estimated the input importance in each model by analyzing the outcomes from all models, and this knowledge may be utilized by managers to improve campaigns (for example, by requesting agents to extend the length of their phone conversations or timing campaigns to certain months).

8.1 Summary:

- We have found that the most important feature in building these models is duration.
- For the duration feature, it's shown that the longer the bank contacts a customer, the more likely the customer is predicted to open a deposit account.
- The next most important feature is month, indicating that customers contacted in the later and earlier months of the year tend to influence them to be predicted as opening a deposit account. However, most customers, based on the month feature, are inclined to be predicted as not opening a deposit account.
- In the contact feature, it's apparent that customers who were not contacted (with 'unknown' values) tend to have a greater influence on being predicted to open a deposit account.
- The poutcome feature suggests that customers who were successfully acquired in the previous campaign by the bank are more likely to be predicted to open a deposit account.
- Regarding the housing feature, individuals who do not have a housing loan are more likely to be predicted to open a deposit account.

8.2 Recommendation for the business

Here are some strategies to increase the potential for customers to engage in deposit offerings:

- Offer campaigns with longer durations to customers, as evidenced by the impact of longer durations on customer engagement with deposits.
- Maximize deposit campaign offerings during Quarter 1 (January, February, March), Quarter 3 (July, August, September), and Quarter 4 (October, November, December).
- Evaluate the success of each campaign; the analysis indicates that customers who were successfully engaged in a previous deposit campaign are more likely to engage in subsequent campaigns.
- Focus deposit campaigns on customers who do not have home loan installments, as the analysis suggests that customers without home loan installments are more likely to engage with deposit offerings.

8.3 Recommendations for the model

- Here are some strategies to further develop the model for better performance:
- Introduce new features to the data, such as the offered deposit interest rate, which could potentially influence a person's decision to engage in a deposit.
- Add a new feature indicating the monthly income of customers, as this could provide deeper insights into customer characteristics and influence the prediction of whether a customer will engage in a deposit.
- Quantify predictions that result in errors, particularly focusing on false positives, given their larger numbers, to better understand and mitigate prediction inaccuracies.

9. References

Aslan Ahmedov. (2021). Predict Term Deposit, Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2865805

Predict term deposit. (2021, November 29). Kaggle. https://www.kaggle.com/datasets/aslanahmedov/predict-term-deposit?resource=download