Bank Turnover Dataset

Big Data & Predictive Analytics

MKTG 746

Section 002

Made by: Vivek Raj (301055278)

Thi Thuy Linh Le (301077496)

Gabriel Velasquez Fuentes (301084803)

For: Prof. David Parent

Date: March 15th, 2020

Table of Contents

Background/Objectives 3

Data Exploration and Partition 3

Step 1: Explore the Dataset 3

Step 2: Data Partition 5

Decision Tree Groups 6

Step 3: Maximal Tree 6

Step 4: Decision Tree Diagram 7

Step 5: Probability Tree Diagram 9

Cap and Floor / Stats Explore 14

Step 6: Cap and Floor 14

Step 7: StatsExplore 16

Regression Model Groups 18

Step 8: Full Regression Model 18

Step 9: Forward Regression Model 21

Step 10: Stepwise Regression Model 24

Step 11: Backward Regression Model 26

Neural Network Model Groups 29

Step 12: Neural Network Model 3H 29

Step 13: Neural Network Model 4 H 31

Step 14: Neural Network 5H 34

Step 15: Neural Network 4H Stepwise 37

Model Comparison 40

Step 16: Model Comparison 40

Conclusion 43

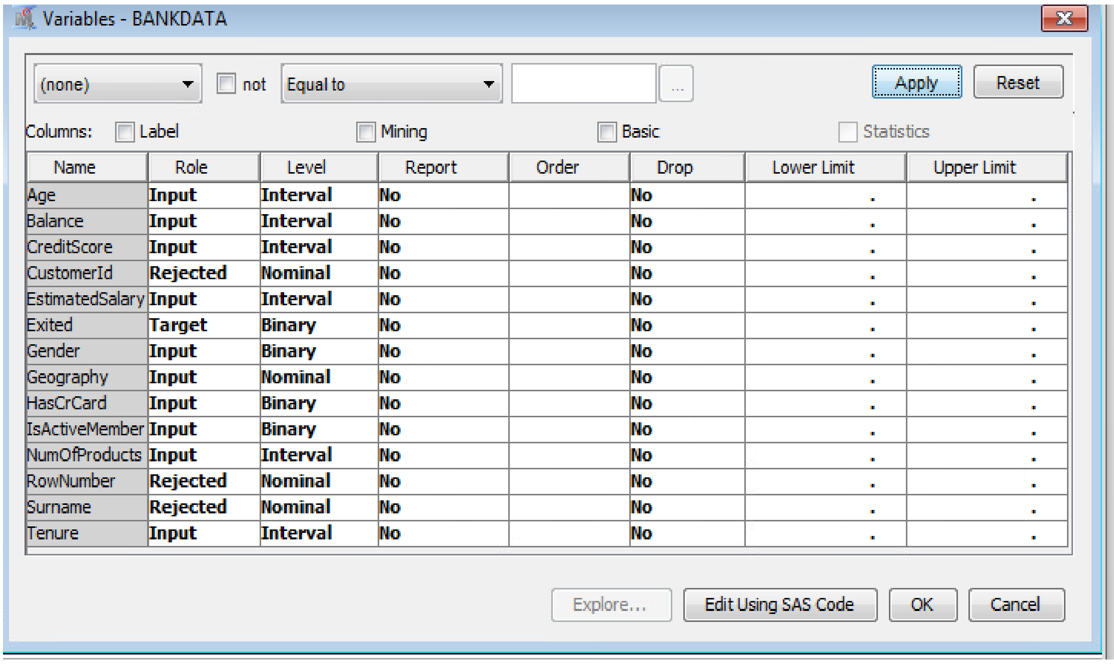
Data Source 43

Background/Objectives

The main objective of this project is to be able to determine among several projection models which is the most optimal to be able to predict the flight probability of clients within a bank. In order to develop this, we worked with a database where the most important variables to consider were: Gender, Geography, Age, Balance, Number of Products, Have a Credit Card or not, active member or not and still a customer of the Bank or not. To start this, the first thing was to determine the role and level of the variables.

Data Exploration and Partition

Step 1: Explore the Dataset



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Balance	Input	Interval	No		No	.	.
CreditScore	Input	Interval	No		No	.	.
CustomerId	Rejected	Nominal	No		No	.	.
EstimatedSalary	Input	Interval	No		No	.	.
Exited	Target	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
Geography	Input	Nominal	No		No	.	.
HasCrCard	Input	Binary	No		No	.	.
IsActiveMember	Input	Binary	No		No	.	.
NumOfProducts	Input	Interval	No		No	.	.
RowNumber	Rejected	Nominal	No		No	.	.
Surname	Rejected	Nominal	No		No	.	.
Tenure	Input	Interval	No		No	.	.

Fig 1: Dataset

After exploring the dataset, we found that we don't need the Customer ID, Row Number and Surname because these variables are too specific and not going to contribute in making the predictions. Our target variable is named as "Exited" (Binary target) where 1 is representing that customer has stopped using the bank and has left the bank.

After exploring the dataset in figure dataset, from Fig 2 we can see that we have no illogical missing values and unusual values so we don't need to make any replacement node before the data partition.

Sample Statistics

Obs #	Variable ...	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number of Levels	Mode Percentage	Mode
1	Gender		CLASS	0	.	.	.	2	54.57	MALE
2	Geography		CLASS	0	.	.	.	3	50.14	FRANCE
3	Surname		CLASS	0	.	.	.	128+	2.619693	MARTIN
4	Age		VAR	0	18	92	38.9218	.	.	
5	Balance		VAR	0	0	250898.1	76485.89	.	.	
6	CreditScore		VAR	0	350	850	650.5288	.	.	
7	CustomerId		VAR	0	15565701	15815690	15690941	.	.	
8	EstimatedS...		VAR	0	11.58	199992.5	100090.2	.	.	
9	Exited		VAR	0	0	1	0.2037	.	.	
10	HasCrCard		VAR	0	0	1	0.7055	.	.	
11	IsActiveMe...		VAR	0	0	1	0.5151	.	.	
12	NumOfProd...		VAR	0	1	4	1.5302	.	.	
13	RowNumber		VAR	0	1	10000	5000.5	.	.	
14	Tenure		VAR	0	0	10	5.0128	.	.	

Fig 2: Sample Statistics

We wanted to make our dataset more generalized and impartial, so for that we changed the sample method to “random” from “top” and changed the fetch size to max (fig 3).

Sample Properties

Property	Value
Rows	10000
Columns	14
Library	BANK
Member	BANKDATA
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	10000
Random Seed	12345

Apply Plot...

Fig. 3: Sample Properties

Step 2: Data Partition

In total we had 10,000 rows. We divided our dataset into two parts: 75% for training and 25% for validation. Since we had 10,000 observations so we thought to train our data on 7,498 observations and carry out the validations on 2,502 observations as we thought this would provide enough observations for carrying out both the functions i.e. train and validation (fig 4).

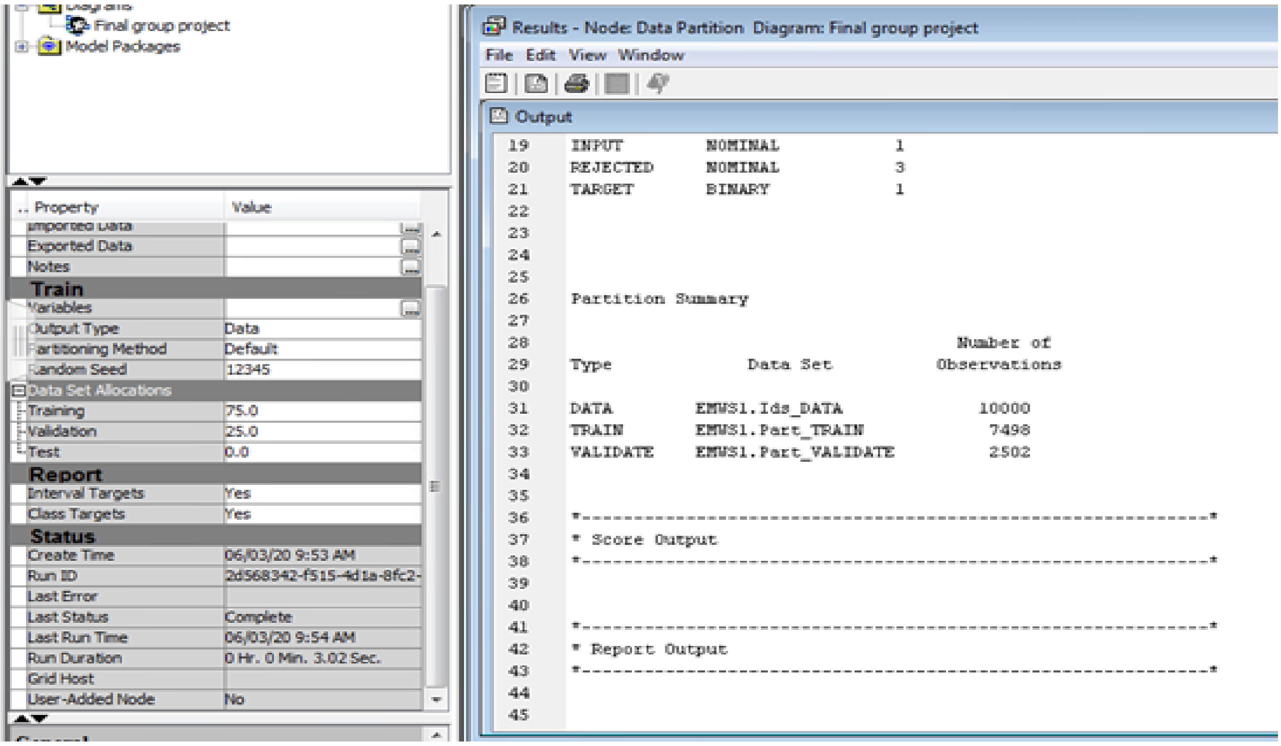


Fig 4: Results - Data Partition

Decision Tree Groups

Step 3: Maximal Tree

First we will start by creating a Maximal Tree. To create a Maximal Tree we launched interacting training and then freeze it. Following is the output of the Maximal Tree (fig 5).

Fig 5: Results - Maximal Tree

From the Fig 6 below, we can see that there are 36 leaves in total and the discrepancy between the training percent and validation is not much, which means the model is not over-fitted.

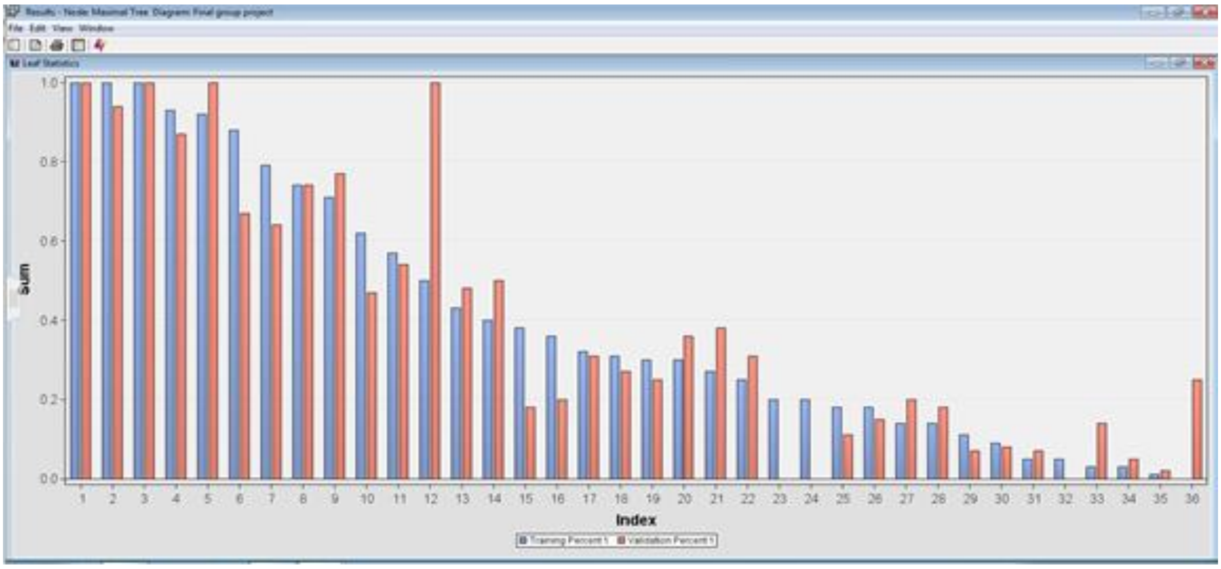


Fig 6: Maximal Tree - Leaf Statistics

Then from the fit statistics below, we checked the valid misclassification rate for maximal tree and it came out to be 0.143086 and average squared error came out to be 0.107738.

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Exited		_ASE_	Average Squared Error	0.102243	0.107738	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_MAX_	Maximum Absolute Error	0.992999	1	.
Exited		_MISC_	Misclassification Rate	0.135103	0.143086	.
Exited		_RASE_	Root Average Squared Error	0.319754	0.328235	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_SSE_	Sum of Squared Errors	1533.233	539.1217	.
Exited		_DFT_	Total Degrees of Freedom	7498	.	.

Fig 7 - Fit Statistics

Step 4: Decision Tree Diagram

We built one decision tree by taking an assessment measure as “decision”. The output can be seen in the fig 8.

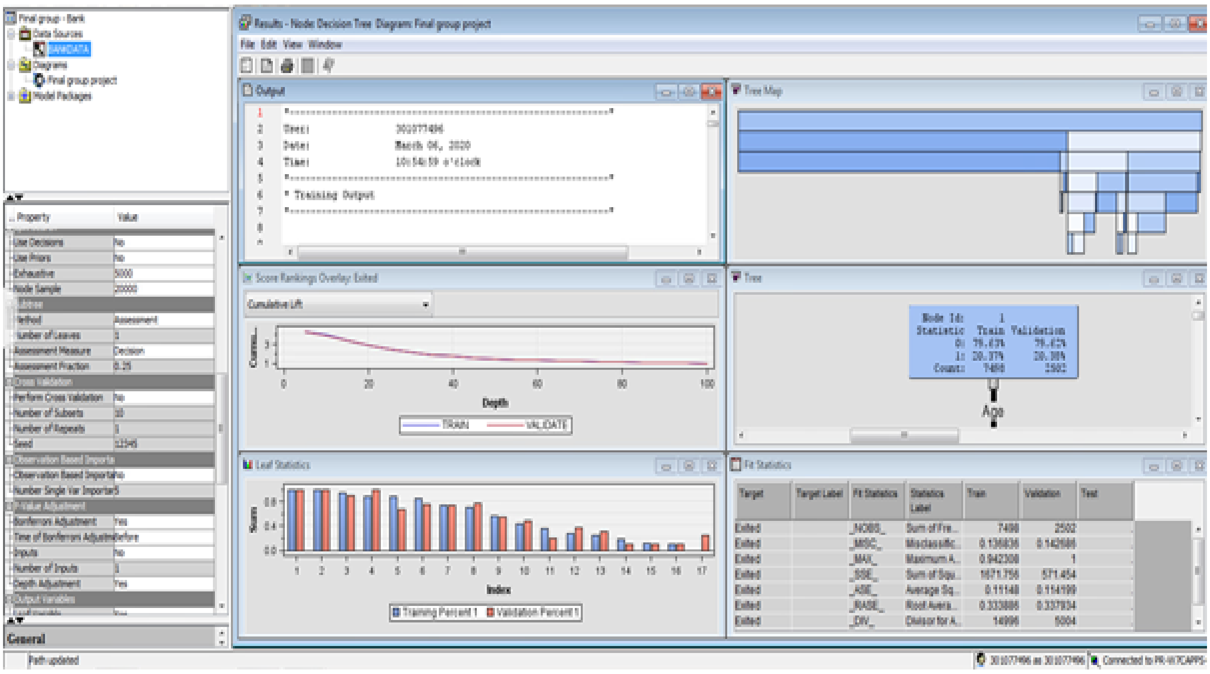


Fig 8: Results - Decision Tree

From the leaf statistics figure below, we can figure out the total number of leaves in the decision tree. We have only 17 leaves for this decision tree and the difference between training percent and validation percent is very small, which means the model is not over-fitted (Fig 9).

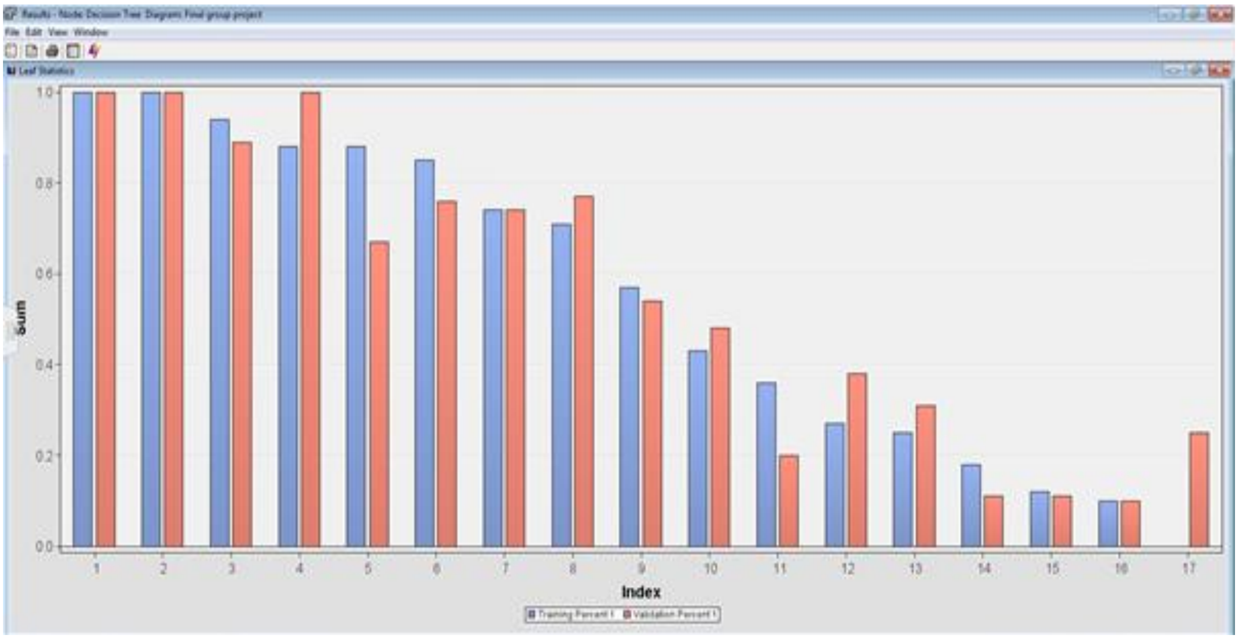
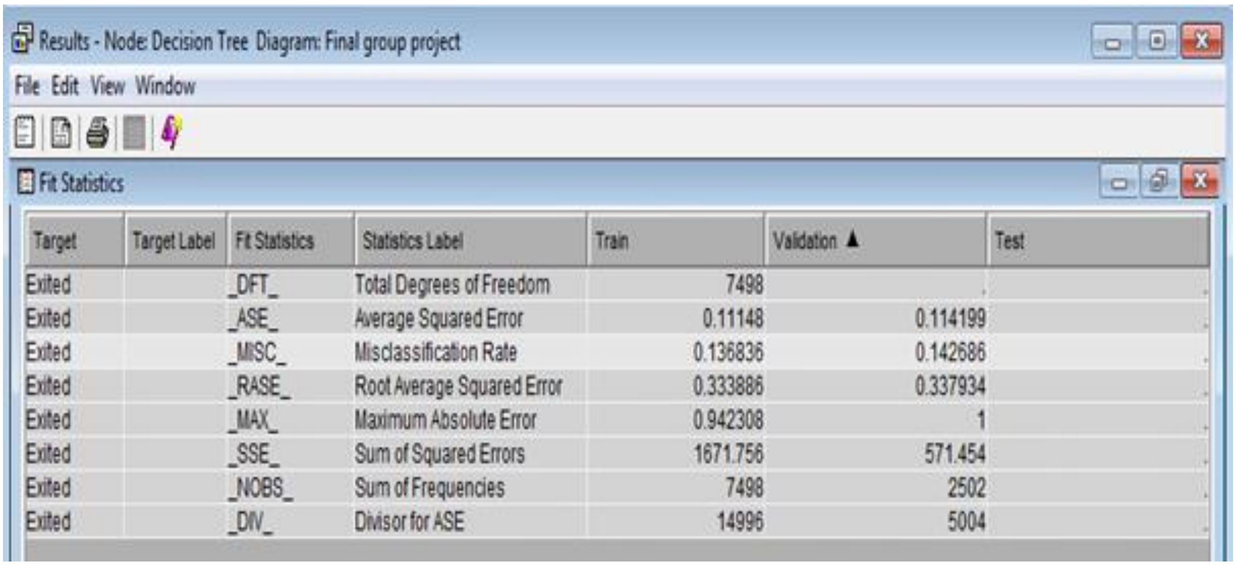


Fig 9: Decision Tree - Leaf statistics

After checking the fit statistics we found out that the Decision tree has smaller Validation Misclassification rate: 0.142686 (compared with Maximal tree with Validation misclassification of 0.143086) (Fig 10).



Results - Node: Decision Tree Diagram: Final group project

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation ▲	Test
Exited		_DFT_	Total Degrees of Freedom	7498	.	.
Exited		_ASE_	Average Squared Error	0.11148	0.114199	.
Exited		_MISC_	Misclassification Rate	0.136836	0.142686	.
Exited		_RASE_	Root Average Squared Error	0.333886	0.337934	.
Exited		_MAX_	Maximum Absolute Error	0.942308	1	.
Exited		_SSE_	Sum of Squared Errors	1671.756	571.454	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_DIV_	Divisor for ASE	14996	5004	.

Fig 10: Decision Tree - Fit Statistics

Step 5: Probability Tree Diagram

We built a Probability tree with assessment measure as “Average Square Error” to predict the probability of the target and the result window is as below:

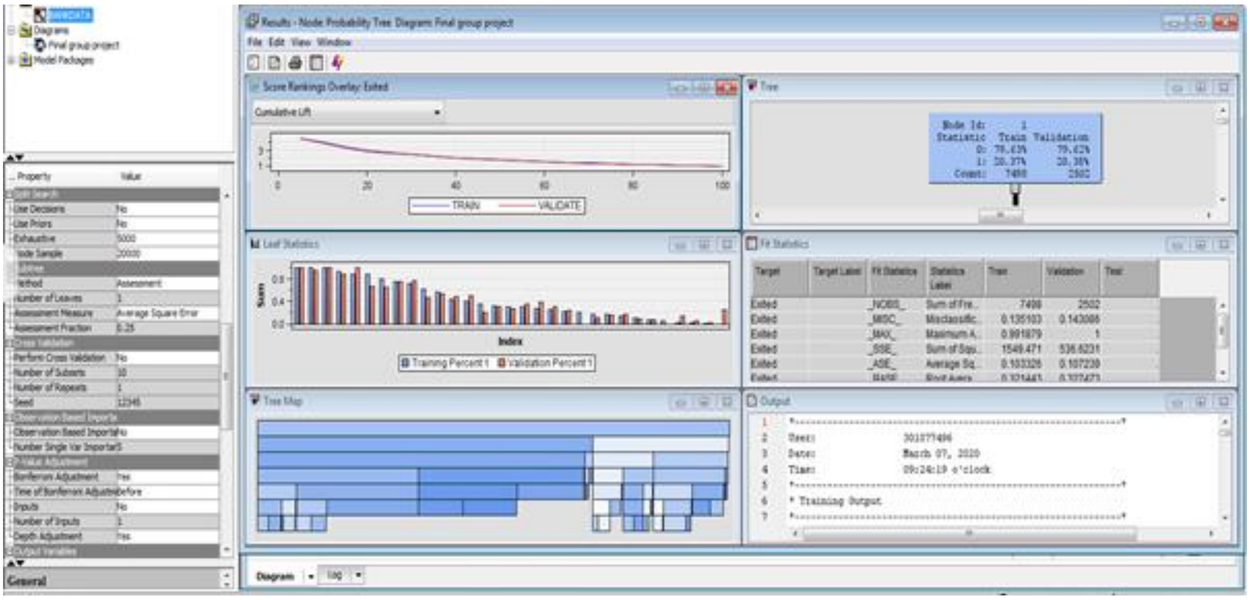


Fig 11: Probability Tree

From the Leaf Statistics (fig 12) we can see that we have 31 leaves for this model as shown below:

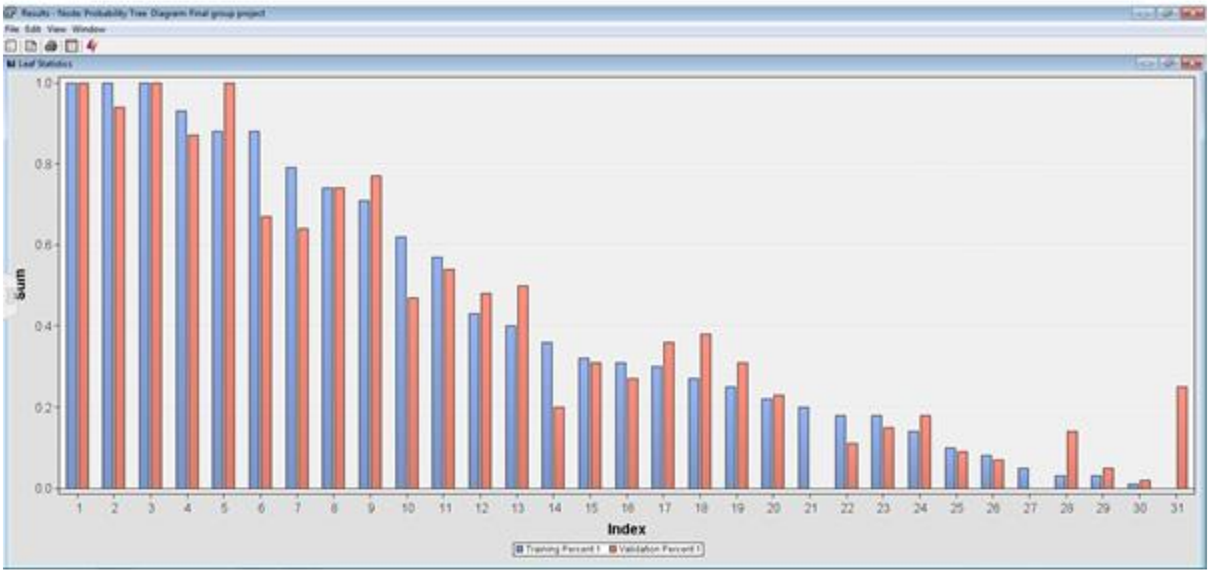


Fig 12: Probability Tree - Leaf Statistics

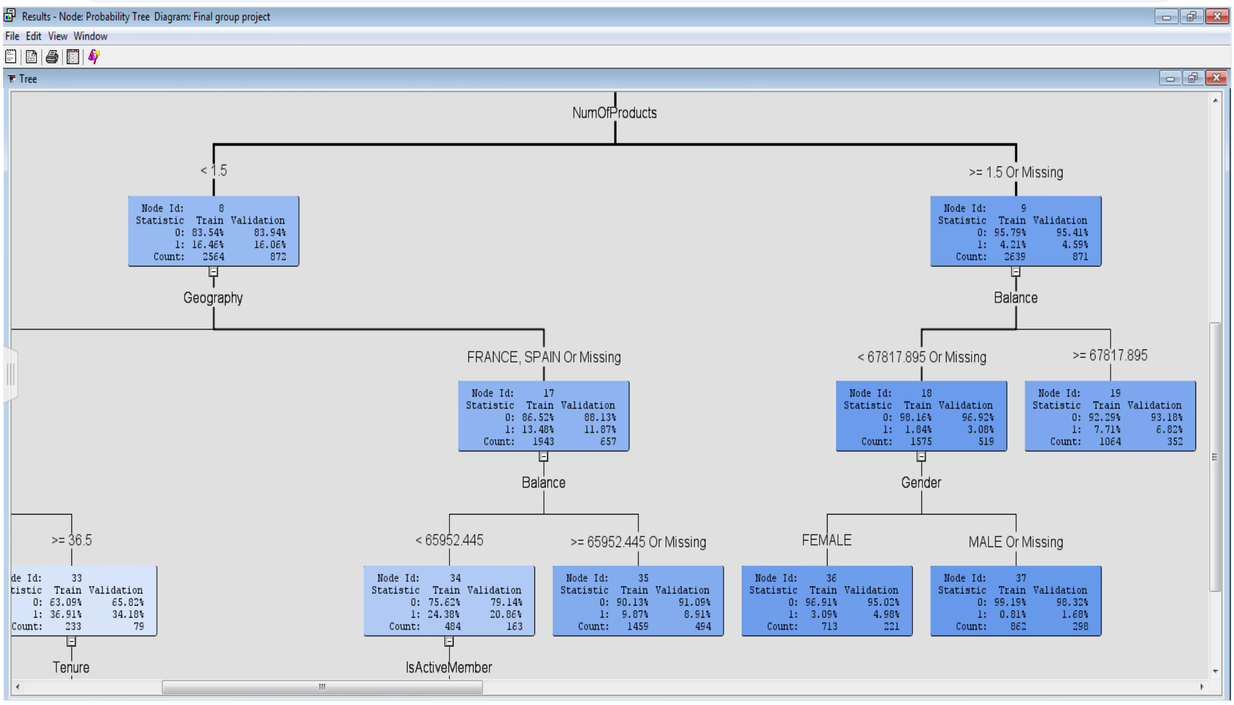
From the Fit Statistics we can see that the Probability Tree has lower validation average squared error: 0.107239 (compared with that of maximal tree: 0.107738).

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_ASE_	Average Squared Error	0.103326	0.107239	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_MAX_	Maximum Absolute Error	0.991879	1	.
Exited		_MISC_	Misclassification Rate	0.135103	0.143086	.
Exited		_RASE_	Root Average Squared Error	0.321443	0.327473	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_SSE_	Sum of Squared Errors	1549.471	536.6231	.
Exited		_DFT_	Total Degrees of Freedom	7498	.	.

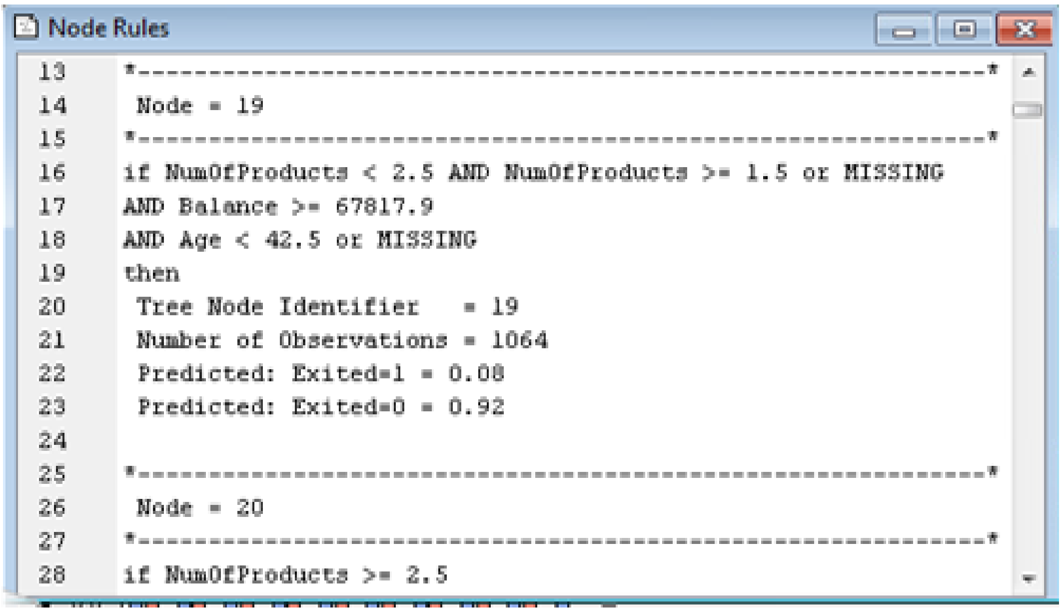
Fig 13: Probability Tree - Fit Statistics

Among three decision trees, the Probability tree seems to be the best model in predicting the leaving probability of clients because it has the lowest ASE. We can use this model to describe the type of people who will leave and the probability of leaving.

The finding is based on the highest number of observations in some nodes with major distinction between exit and non-exit rate as below:



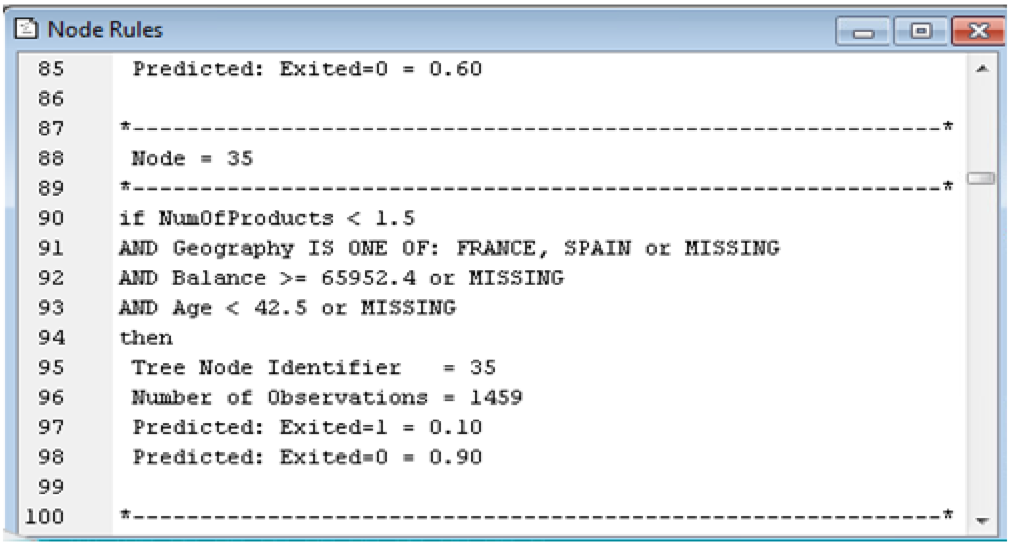
The people who will NOT exit are shown in fig. 13.1, 13.2, 13.3, 13.4 at Node 19, 35, 36, 37 as below:



```
Node Rules
*------------------------------------------------------------*
 Node = 19
*------------------------------------------------------------*
if NumOfProducts < 2.5 AND NumOfProducts >= 1.5 or MISSING
AND Balance >= 67817.9
AND Age < 42.5 or MISSING
then
 Tree Node Identifier   = 19
 Number of Observations = 1064
 Predicted: Exited=1 = 0.08
 Predicted: Exited=0 = 0.92

*------------------------------------------------------------*
 Node = 20
*------------------------------------------------------------*
if NumOfProducts >= 2.5
```

(Fig 13.1 - Node 19)

From the Node 19 (Fig 13.1 above), we can see that clients who have 2 products (NumOfProducts < 2.5 and >=1.5 or missing), having the balance more than $67,817 and age under 42.5 years old are 92% likely to stay with our bank. This prediction is based on 1,064 observations.



```
Node Rules
 Predicted: Exited=0 = 0.60

*------------------------------------------------------------*
 Node = 35
*------------------------------------------------------------*
if NumOfProducts < 1.5
AND Geography IS ONE OF: FRANCE, SPAIN or MISSING
AND Balance >= 65952.4 or MISSING
AND Age < 42.5 or MISSING
then
 Tree Node Identifier   = 35
 Number of Observations = 1459
 Predicted: Exited=1 = 0.10
 Predicted: Exited=0 = 0.90

*------------------------------------------------------------*
```

(Fig 13.2 - Node 35)

From the Node 35 (Fig 13.2 above), we can see that clients who have 1 product (NumOfProducts < 1.5), living in France or Spain, having balance more than $65,952 and age under 42.5 years old (or missing) are 90% likely to stay with our bank. This prediction is based on 1,459 observations.

Node Rules

```
*------------------------------------------------------------*
 Node = 36
*------------------------------------------------------------*
if NumOfProducts < 2.5 AND NumOfProducts >= 1.5 or MISSING
AND Gender IS ONE OF: FEMALE
AND Balance < 67817.9 or MISSING
AND Age < 42.5 or MISSING
then
 Tree Node Identifier   = 36
 Number of Observations = 713
 Predicted: Exited=1 = 0.03
 Predicted: Exited=0 = 0.97

*------------------------------------------------------------*
 Node = 37
*------------------------------------------------------------*
```

(Fig 13.3 - Node 36)

From the Node 36 (Fig 13.3 above), we can see that clients who have 2 products (NumOfProducts < 2.5 and >=1.5 or missing), are female, having balance less than $67,817 and age under 42.5 years old (or missing) are 97% likely to stay with our bank. This prediction is based on 713 observations.

Node Rules

```

*------------------------------------------------------------*
 Node = 37
*------------------------------------------------------------*
if NumOfProducts < 2.5 AND NumOfProducts >= 1.5 or MISSING
AND Gender IS ONE OF: MALE or MISSING
AND Balance < 67817.9 or MISSING
AND Age < 42.5 or MISSING
then
 Tree Node Identifier   = 37
 Number of Observations = 862
 Predicted: Exited=1 = 0.01
 Predicted: Exited=0 = 0.99

*------------------------------------------------------------*
 Node = 41
```

(Fig 13.4 - Node 37)

From the Node 37 (Fig 13.4 above), we can see that clients who have 2 products (NumOfProducts < 2.5 and >=1.5 or missing), are male, having balance less than $67,817 and age under 42.5 years old (or missing) are 99% likely to stay with our bank. This prediction is based on 862 observations.

- The people who will exit are shown in Fig. 13.5, 13.6 at Node 62 and Node 63 :

From the Node 62 (Fig 13.5 below), we can see that clients who have 1 product (NumOfProducts < 1.5 or missing), are not active members, and age from 47.5 years old (or missing) to 52.5 years old are 79% likely to leave our bank. This prediction is based on 173 observations.

Node Rules

```
 Node = 62
*------------------------------------------------------------*
if NumOfProducts < 1.5 or MISSING
AND IsActiveMember IS ONE OF: 0
AND Age < 52.5 AND Age >= 47.5 or MISSING
then
 Tree Node Identifier   = 62
 Number of Observations = 173
 Predicted: Exited=1 = 0.79
 Predicted: Exited=0 = 0.21

*------------------------------------------------------------*
 Node = 63
*------------------------------------------------------------*
if NumOfProducts < 1.5 or MISSING
AND IsActiveMember IS ONE OF: 0
AND Age < 73 AND Age >= 52.5
then
 Tree Node Identifier   = 63
 Number of Observations = 151
 Predicted: Exited=1 = 0.93
 Predicted: Exited=0 = 0.07

*------------------------------------------------------------*
```

(Fig 13.5 - Node 62 and Node 63)

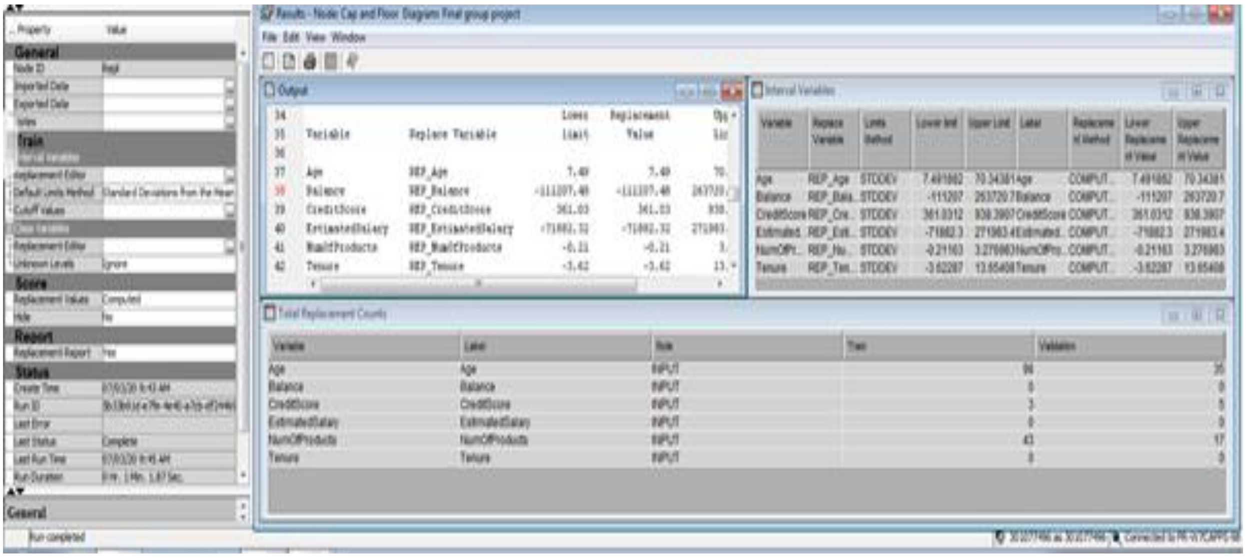
From the Node 63 (Fig 13.5 above), we can see that clients who have 1 product (NumOfProducts < 1.5 or missing), are not active members, and age from 52.5 years old to less than 73 years old are 93% likely to leave our bank. This prediction is based on 151 observations.

In short, people who have only 1 product, not active members, more than 47 years old have at least 79% chance to leave our bank. Meanwhile, at a higher level (node 9), people who are young (under 42.5 years old) and have 2 products have 95% chance to stay with our bank. In the case of geography, people who are young, have 1 product, are living in France or Spain, and have a high balance (over $65,000) have 90% chance to stay with us.

Cap and Floor / Stats Explore

Step 6: Cap and Floor

Now we are going to add the replacement node to cap and floor the outliers in the interval variables (the dataset within three standard deviations from the mean. We are going to remove the outliers in order to not bring the average up as they should and not bias the model. For consistency in data, we have set the limit to 3 standard deviations. Fig. 14 is showing the result for the output of the replacement node.



(Fig 14): Cap and Floor

Fig.15 shows that Age, Credit Score and Number of Products were capped and floored within three standard deviation from the mean with the number of replacement shown in the below Figure:

Total Replacement Counts

Variable	Label	Role	Train	Validation
Age	Age	INPUT	98	35
Balance	Balance	INPUT	0	0
CreditScore	CreditScore	INPUT	3	5
EstimatedSalary	EstimatedSalary	INPUT	0	0
NumOfProducts	NumOfProducts	INPUT	43	17
Tenure	Tenure	INPUT	0	0

Interval Variables

Variable	Replace Variable	Limits Method	Lower limit	Upper Limit	Label	Replacement Method	Lower Replacement Value	Upper Replacement Value
Age	REP_Age	STDDEV	7.491882	70.34381	Age	COMPUTED	7.491882	70.34381
Balance	REP_Balance	STDDEV	-111207	263720.7	Balance	COMPUTED	-111207	263720.7
CreditScore	REP_CreditS...	STDDEV	361.0312	938.3907	CreditScore	COMPUTED	361.0312	938.3907
EstimatedSal...	REP_Estimat...	STDDEV	-71882.3	271983.4	EstimatedSal...	COMPUTED	-71882.3	271983.4
NumOfProdu...	REP_NumOf...	STDDEV	-0.21163	3.276983	NumOfProdu...	COMPUTED	-0.21163	3.276983
Tenure	REP_Tenure	STDDEV	-3.62287	13.65408	Tenure	COMPUTED	-3.62287	13.65408

(Fig 15): Total Replacement Counts

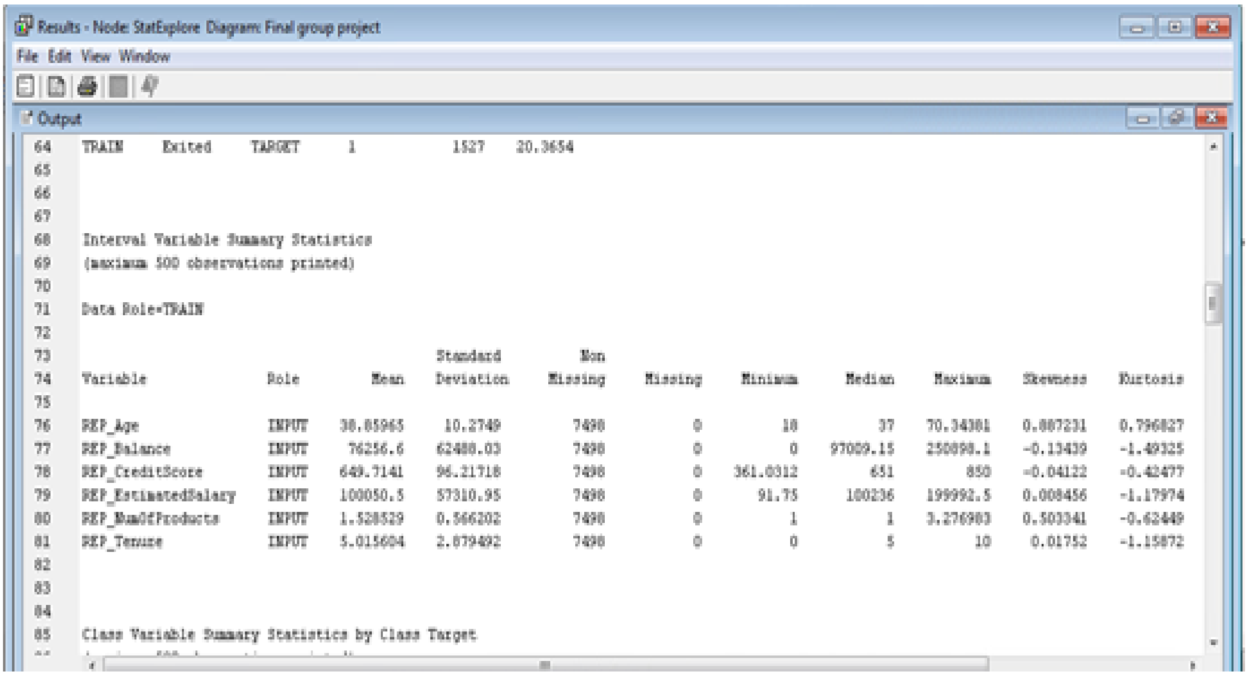
We have 98 outliers of *Age* in the training dataset and 35 in the validation dataset. We capped *Age* at 70 years old, so we have 98 people who are over 70 years old to be brought down to 70 years old.

We have 3 outliers of *CreditScore* in the training dataset and 5 in the validation dataset. So we capped *CreditScore* at 938, which means all the *CreditScore* records which are over 938 were brought down to 938.

We have 43 outliers of *NumOfProducts* in the training dataset and 17 in the validation dataset. So we capped all the records having over 3 *NumOfProducts* and brought them down to 3.

Step 7: StatsExplore

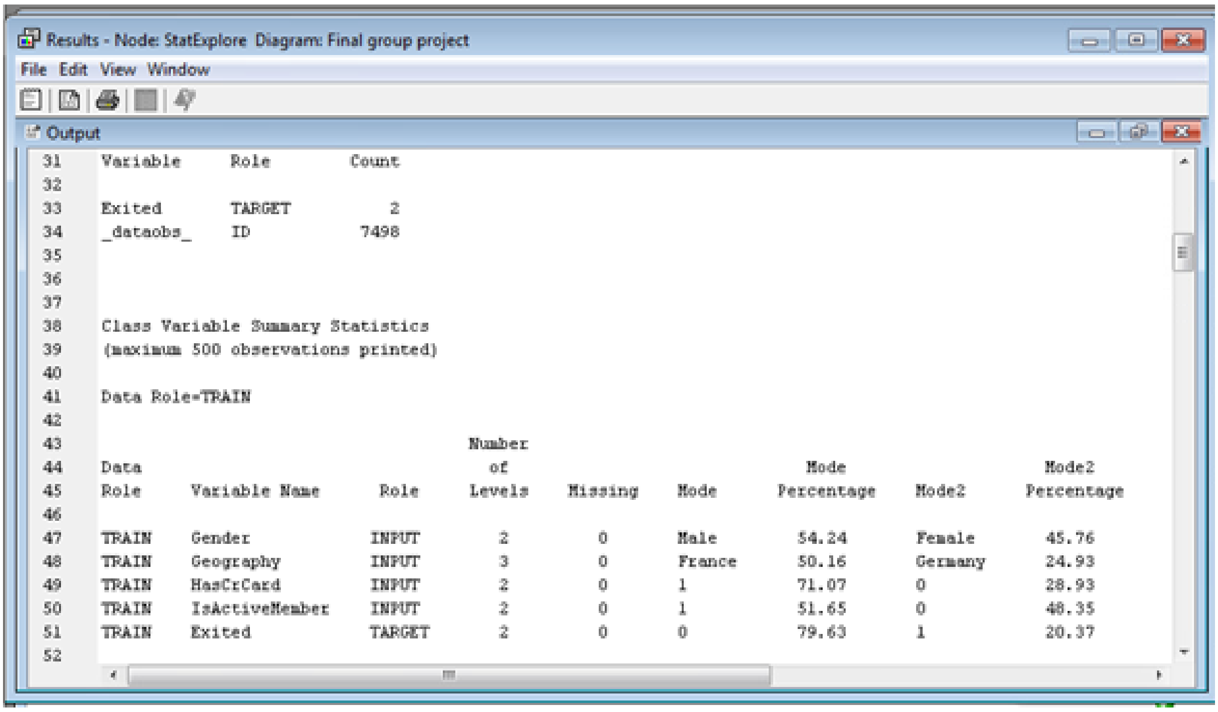
After cap and floor, we explore the dataset by using StatsExplore node to check the statistics i.e. if it has any missing value or if there is any Skewness or Kurtosis:



(Fig 16): StatExplore

From Fig 16 above, we can see that there is no missing values so we don't need to impute anything, Skewness of all variables are between -1 and 1 and Kurtosis are all under 3, which means we don't have any Skewness and Kurtosis so we don't need to do any transformation.

Also from Fig 17 below, we can see that all the class variables just have 2 or 3 levels, so they will not cause the curse of dimensionality. Therefore, we don't need to Recode Dummies before building Regression Models and Neural Network Models.



Results - Node: StatExplore Diagram: Final group project
File Edit View Window
Output

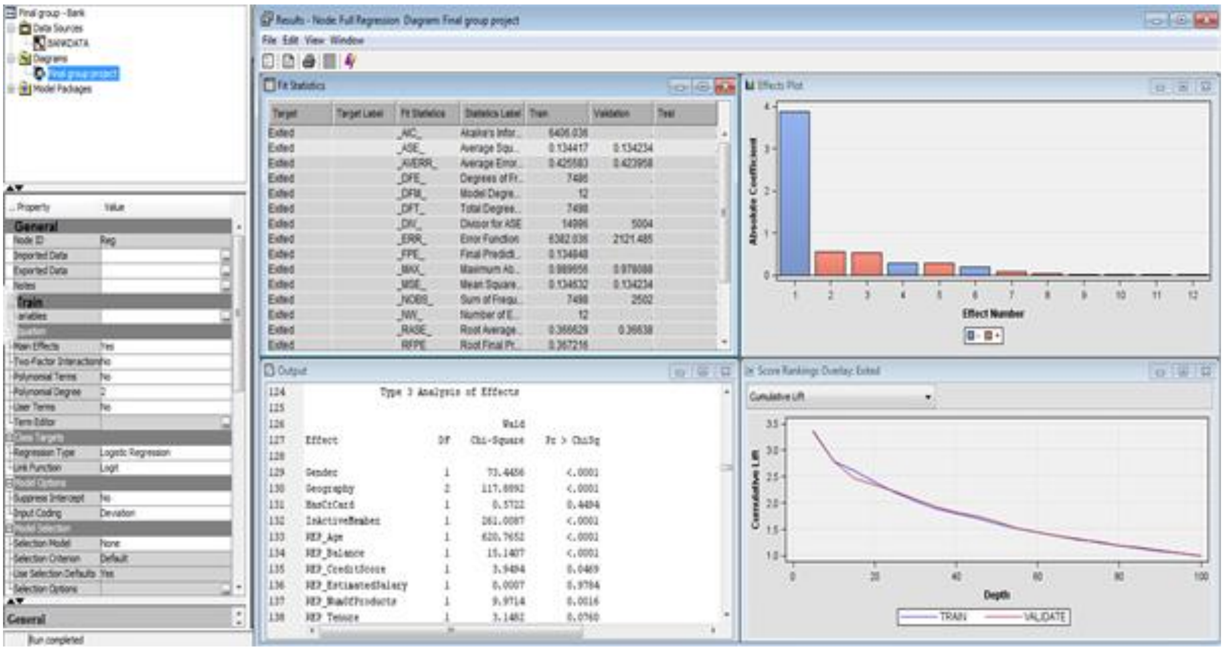
```
Variable      Role        Count

Exited        TARGET          2
_dataobs_     ID           7498



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                     Number
Data                                   of                              Mode                      Mode2
Role     Variable Name       Role    Levels    Missing    Mode      Percentage    Mode2       Percentage

TRAIN    Gender              INPUT      2         0       Male         54.24      Female         45.76
TRAIN    Geography           INPUT      3         0       France       50.16      Germany        24.93
TRAIN    HasCrCard           INPUT      2         0       1            71.07      0              28.93
TRAIN    IsActiveMember      INPUT      2         0       1            51.65      0              48.35
TRAIN    Exited              TARGET     2         0       0            79.63      1              20.37
```

(Fig 17): StatExplore Diagram

Regression Model Groups

Step 8: Full Regression Model

Now we are going to start with the Regression Models. Building Full Regression Model with Selection Model as "None" and Selection Criterion as "Default", our result window is shown as below:



(Fig 18): Full Regression Model

From Fig.19 below, we can see that there are three variables which are "*HasCrCard*"**,** "*REP_Estimated Salary*" and "*REP_Tenure*" are not statistically significant because they have Pr > ChiSq values larger than 0.05.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	73.4456	<.0001
Geography	2	117.8892	<.0001
HasCrCard	1	0.5722	0.4494
IsActiveMember	1	261.0087	<.0001
REP_Age	1	620.7652	<.0001
REP_Balance	1	15.1407	<.0001
REP_CreditScore	1	3.9494	0.0469
REP_EstimatedSalary	1	0.0007	0.9784
REP_NumOfProducts	1	9.9714	0.0016
REP_Tenure	1	3.1482	0.0760

Fig 19: Type 3 Analysis of Effects

From the Fig 20 below, we could see that In the case of the variable *Gender*, we can determine that Female are 71.9% more likely to leave the bank compared to Male. In case of *Geography*, customers in Germany are more than twice as likely to leave the bank as clients in Spain. In the case of *IsActiveMember*, members who are not active are almost 3 times more likely to leave the bank than active members. In the case of *NumOfProducts*, for each additional number of products, those customers are 16.4% less likely to leave.

Output

Odds Ratio Estimates

Effect		Point Estimate
Gender	Female vs Male	1.719
Geography	France vs Spain	0.980
Geography	Germany vs Spain	2.198
HasCrCard	0 vs 1	1.054
IsActiveMember	0 vs 1	2.942
REP_Age		1.079
REP_Balance		1.000
REP_CreditScore		0.999
REP_EstimatedSalary		1.000
REP_NumOfProducts		0.835
REP_Tenure		0.981

--

Fig 20 - Odds Ratio Estimates

From Fig 21, we found our Validation ASE is 0.134234, much higher that of any decision tree above.

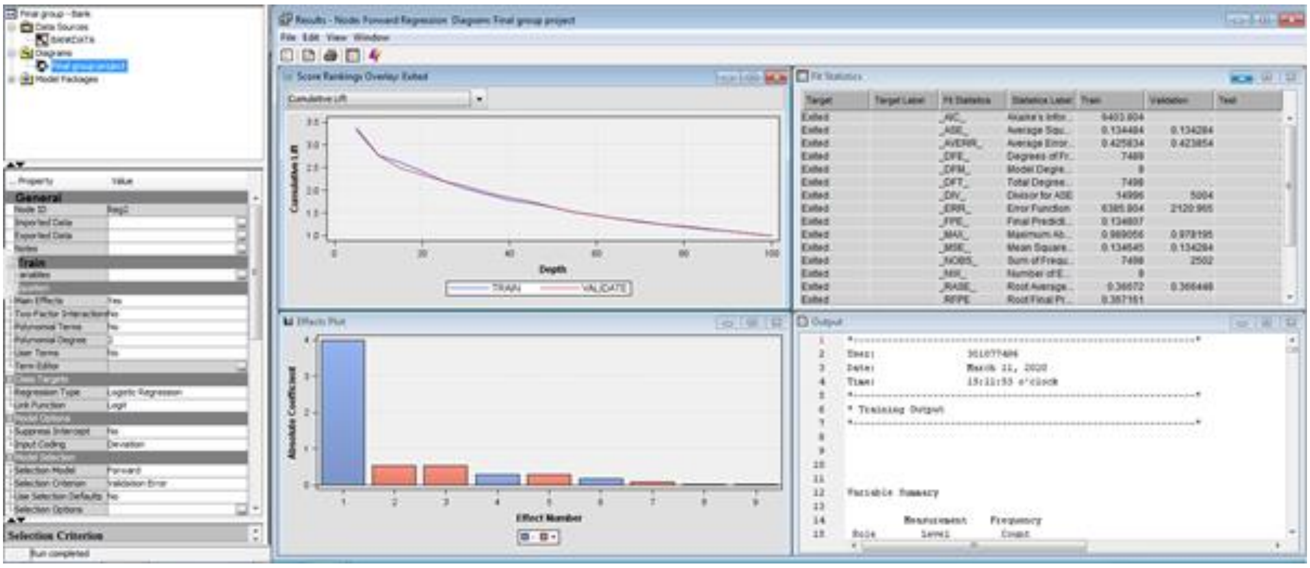
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation ▲	Test
Exited		_AIC_	Akaike's Inform...	6406.036	.	.
Exited		_DFE_	Degrees of Fre...	7486	.	.
Exited		_DFM_	Model Degrees...	12	.	.
Exited		_DFT_	Total Degrees ...	7498	.	.
Exited		_FPE_	Final Prediction...	0.134848	.	.
Exited		_NW_	Number of Esti...	12	.	.
Exited		_RFPE_	Root Final Pred...	0.367216	.	.
Exited		_SBC_	Schwarz's Baye...	6489.105	.	.
Exited		_ASE_	Average Squar...	0.134417	0.134234	.
Exited		_MSE_	Mean Square E...	0.134632	0.134234	.
Exited		_MISC_	Misclassificatio...	0.187517	0.182254	.
Exited		_RASE_	Root Average S...	0.366629	0.36638	.
Exited		_RMSE_	Root Mean Squ...	0.366923	0.36638	.
Exited		_AVERR_	Average Error F...	0.425583	0.423958	.
Exited		_MAX_	Maximum Abso...	0.989656	0.978088	.
Exited		_SSE_	Sum of Square...	2015.715	671.7067	.
Exited		_ERR_	Error Function	6382.036	2121.485	.
Exited		_NOBS_	Sum of Freque...	7498	2502	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_SUMW_	Sum of Case ...	14996	5004	.

(Fig 21): Fit Statistics

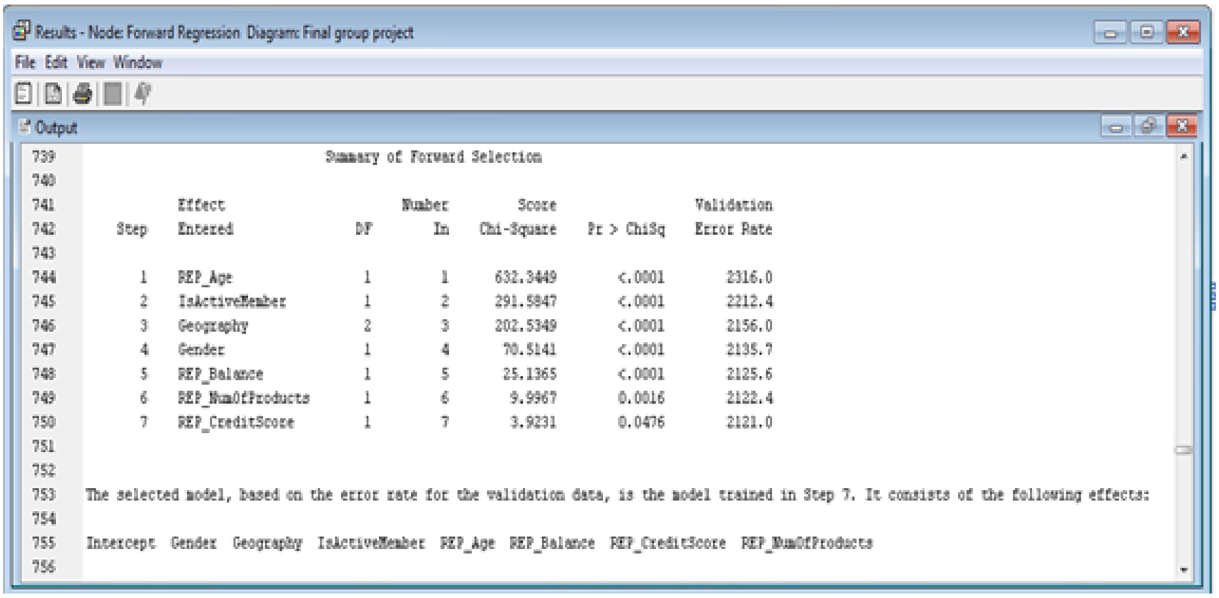
Step 9: Forward Regression Model

After building the Full Regression Model, we will build the Forward Regression Model with selection model as “Forward” and selection criterion as “Validation Error” and the result window is shown as below (Fig 22):



(Fig 22): Forward Regression Model

From Fig 23 below, the Forward Regression Model, based on the error rate for the validation data, is the model trained in Step 7 of forward selection. It consists of the following variables: I*ntercept, Gender, Geography, IsActiveMember, REP-Age, REP_Balance, REP_CreditScore* and *REP_NumOfProducts.*



Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	REP_Age	1	1	632.3449	<.0001	2316.0
2	IsActiveMember	1	2	291.5847	<.0001	2212.4
3	Geography	2	3	202.5349	<.0001	2156.0
4	Gender	1	4	70.5141	<.0001	2135.7
5	REP_Balance	1	5	25.1365	<.0001	2125.6
6	REP_NumOfProducts	1	6	9.9967	0.0016	2122.4
7	REP_CreditScore	1	7	3.9231	0.0476	2121.0

The selected model, based on the error rate for the validation data, is the model trained in Step 7. It consists of the following effects:

Intercept Gender Geography IsActiveMember REP_Age REP_Balance REP_CreditScore REP_NumOfProducts

Fig 23: Summary of Forward Selection

From Fig 24 below, in the variable *Gender*, females have 72.1% more chances of exiting the bank compared to males. In case of *Geography*, customers in Germany have over twice more chances of leaving the bank than the customers in Spain. For *IsActiveMember*, members who are not active have nearly 3 times more chances of leaving the bank than active members. By last NumOfProducts, for each additional number of product, those customers are 17.4% less likely to leave.

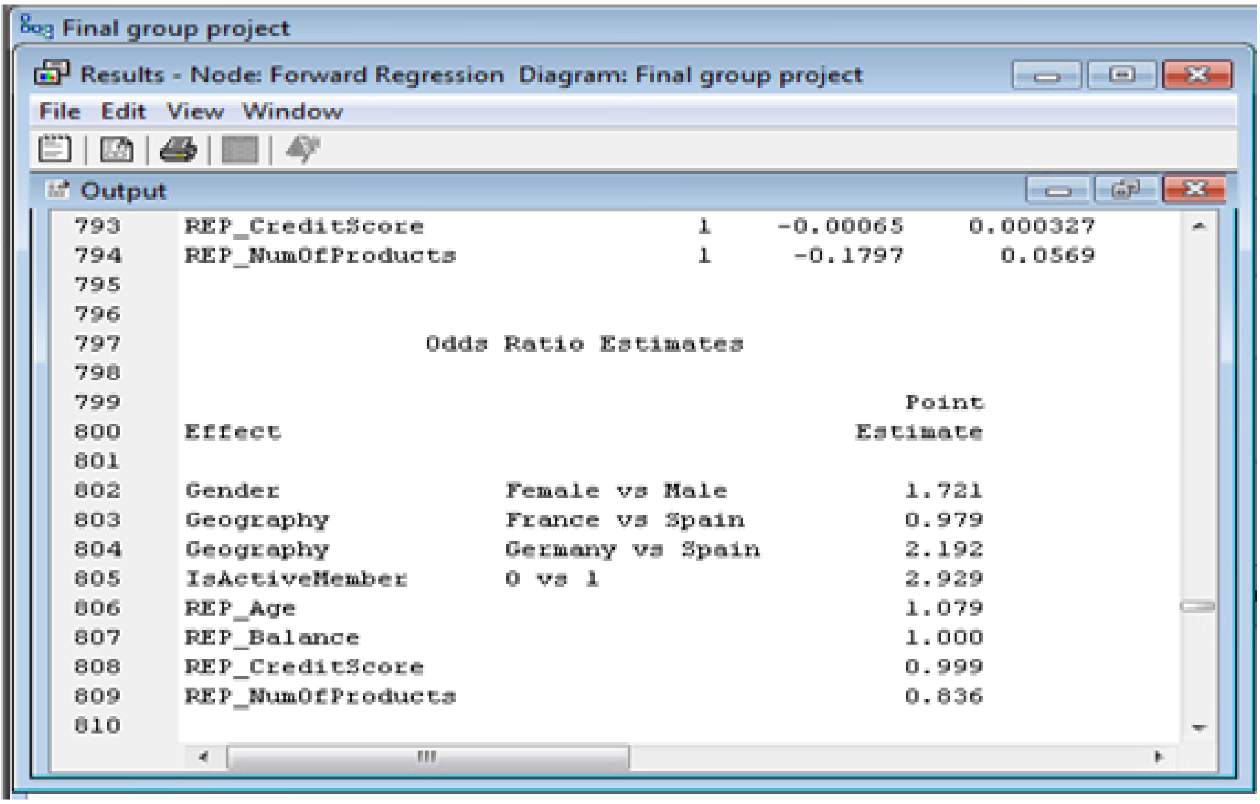


Fig 24: Odds Ratio Estimates

From the Fig 25 below, we have validation ASE which is 0.134284, slightly higher than that of Full Regression Model (0.134234). So we can see that this is also not a best model to find the probability.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_AIC_	Akaike's Information Criterion	6403.804	.	.
Exited		_ASE_	Average Squared Error	0.134484	0.134284	.
Exited		_AVERR_	Average Error Function	0.425834	0.423854	.
Exited		_DFE_	Degrees of Freedom for Error	7489	.	.
Exited		_DFM_	Model Degrees of Freedom	9	.	.
Exited		_DFT_	Total Degrees of Freedom	7498	.	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_ERR_	Error Function	6385.804	2120.965	.
Exited		_FPE_	Final Prediction Error	0.134807	.	.
Exited		_MAX_	Maximum Absolute Error	0.989056	0.978195	.
Exited		_MSE_	Mean Square Error	0.134645	0.134284	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_NW_	Number of Estimate Weights	9	.	.
Exited		_RASE_	Root Average Sum of Squares	0.36672	0.366448	.
Exited		_RFPE_	Root Final Prediction Error	0.367161	.	.
Exited		_RMSE_	Root Mean Squared Error	0.366941	0.366448	.
Exited		_SBC_	Schwarz's Bayesian Criterion	6466.106	.	.
Exited		_SSE_	Sum of Squared Errors	2016.72	671.9566	.
Exited		_SUMW_	Sum of Case Weights Times ...	14996	5004	.
Exited		_MISC_	Misclassification Rate	0.186583	0.185851	.

Fig 25: Fit Statistics

Step 10: Stepwise Regression Model

After building the Forward regression model, we will build the Stepwise Regression Model with Selection Model is "*Stepwise*" and Selection Criterion is "*Validation Error*" and the result window is shown as below:

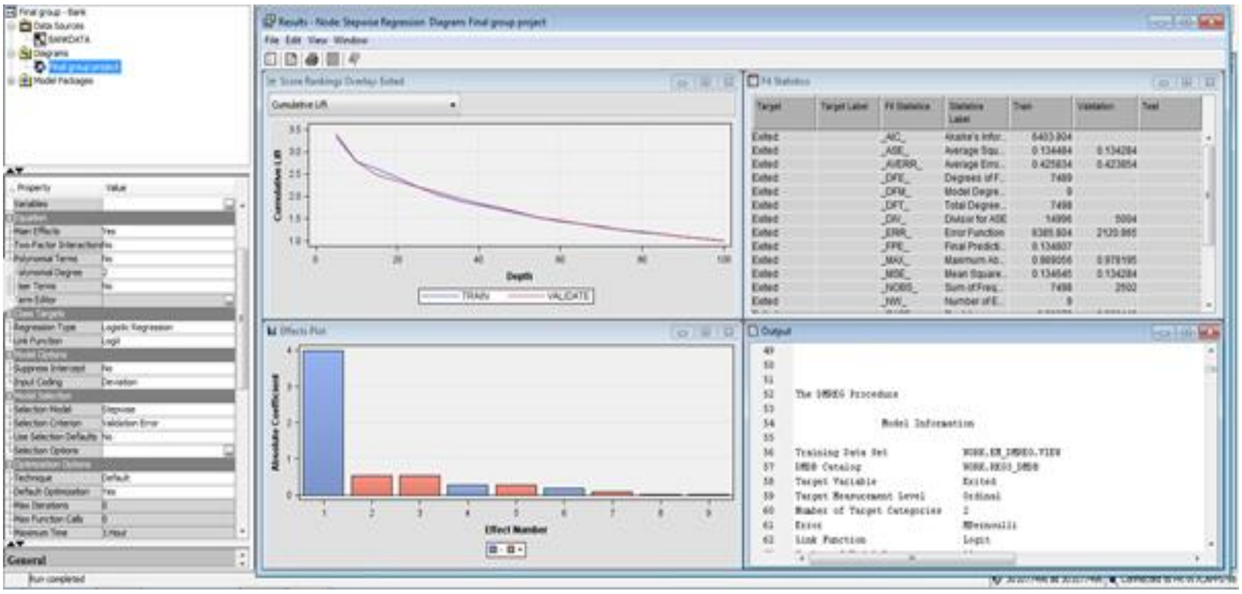
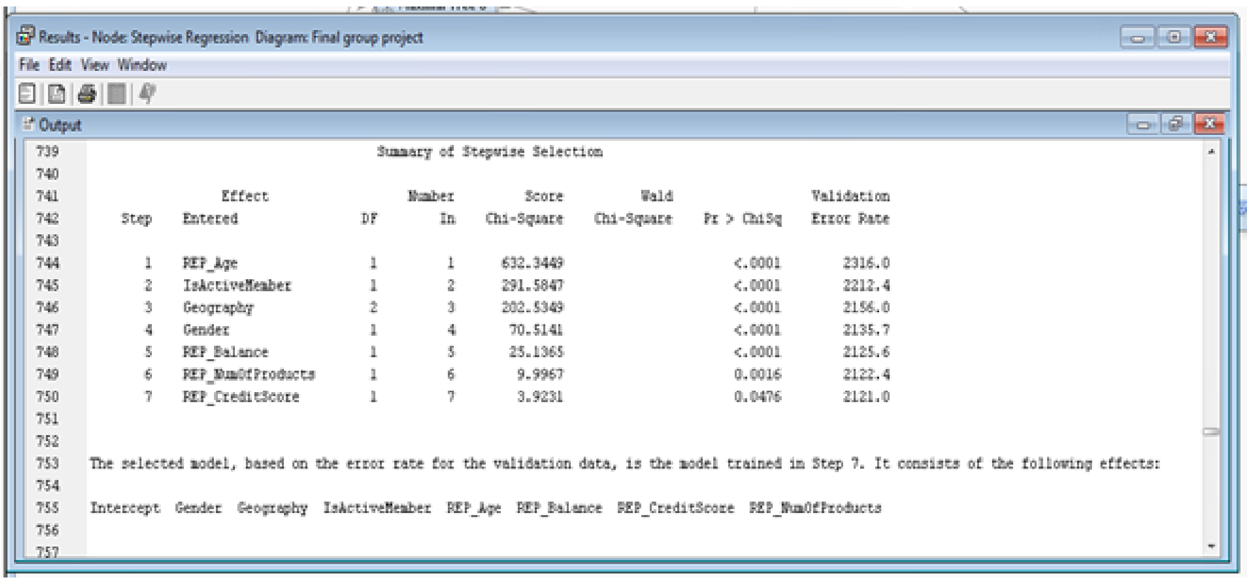


Fig 26: Stepwise Regression Model

From Fig 27, we can see that the Stepwise Regression Model, based on the error rate for the validation data, is also the model trained in Step 7 of Stepwise selectionIt also has the same variables like Forward Regression model, namely: *Intercept, Gender, Geography, IsActiveMember, REP-Age, REP_Balance, REP_CreditScore* and *REP_NumOfProducts.*



Results - Node: Stepwise Regression Diagram: Final group project

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	REP_Age	1	1	632.3449		<.0001	2316.0
2	IsActiveMember	1	2	291.5847		<.0001	2212.4
3	Geography	2	3	202.5349		<.0001	2156.0
4	Gender	1	4	70.5141		<.0001	2135.7
5	REP_Balance	1	5	25.1365		<.0001	2125.6
6	REP_NumOfProducts	1	6	9.9967		0.0016	2122.4
7	REP_CreditScore	1	7	3.9231		0.0476	2121.0

The selected model, based on the error rate for the validation data, is the model trained in Step 7. It consists of the following effects:

Intercept Gender Geography IsActiveMember REP_Age REP_Balance REP_CreditScore REP_NumOfProducts

(Fig 27): Summary of Stepwise Selection

From Fig 28 below, we can interpret the result of Stepwise Regression Model the same as the result of Forward Regression Model.

Results - Node: Stepwise Regression Diagram: Final group project

Output

Odds Ratio Estimates

Effect		Point Estimate
Gender	Female vs Male	1.721
Geography	France vs Spain	0.979
Geography	Germany vs Spain	2.192
IsActiveMember	0 vs 1	2.929
REP_Age		1.079
REP_Balance		1.000
REP_CreditScore		0.999
REP_NumOfProducts		0.836

Fig 28: Odds Ratio Estimates

From the Fig 29 below, we have validation ASE is 0.134284, same as that of Forward Regression Model.

Results - Node: Stepwise Regression Diagram: Final group project

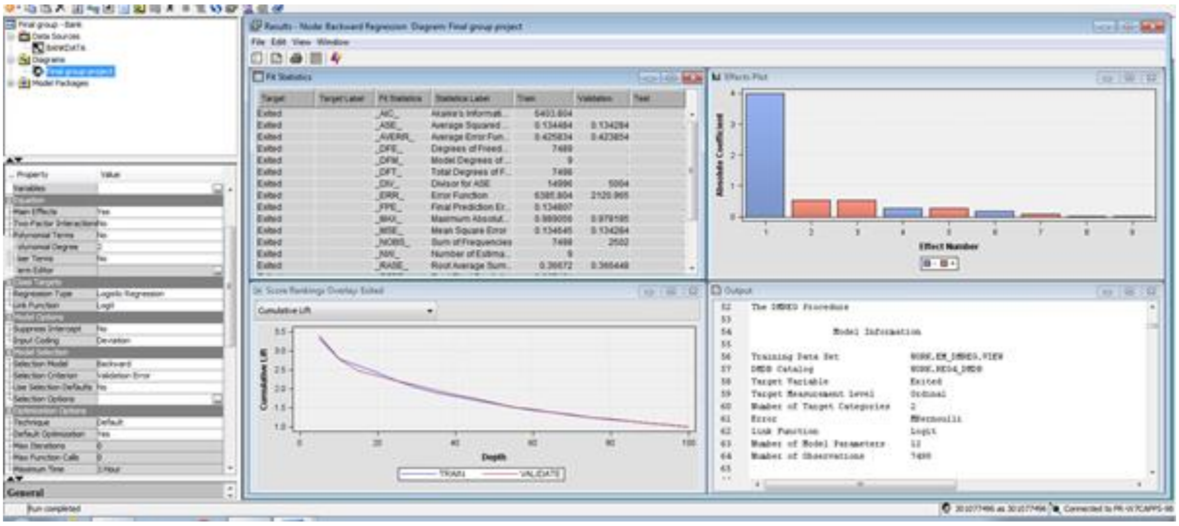
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_AIC_	Akaike's Information Criterion	6403.804	.	.
Exited		_ASE_	Average Squared Error	0.134484	0.134284	.
Exited		_AVERR_	Average Error Function	0.425834	0.423854	.
Exited		_DFE_	Degrees of Freedom for Error	7489	.	.
Exited		_DFM_	Model Degrees of Freedom	9	.	.
Exited		_DFT_	Total Degrees of Freedom	7498	.	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_ERR_	Error Function	6385.804	2120.965	.
Exited		_FPE_	Final Prediction Error	0.134807	.	.
Exited		_MAX_	Maximum Absolute Error	0.989056	0.978195	.
Exited		_MSE_	Mean Square Error	0.134645	0.134284	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_NW_	Number of Estimate Weights	9	.	.
Exited		_RASE_	Root Average Sum of Squares	0.36672	0.366448	.
Exited		_RFPE_	Root Final Prediction Error	0.367161	.	.
Exited		_RMSE_	Root Mean Squared Error	0.366941	0.366448	.

Fig 29: Fit Statistics

Step 11: Backward Regression Model

After building the Stepwise Regression Model, we will build the Backward Regression Model with Selection Model is "*Backward*" and Selection Criterion is "*Validation Error*" and the result window is shown as below:



(Fig 30): Backward Regression Model

From Fig 31, we can see that the Stepwise Regression Model removed three following variables: REP_EstimatedSalary, HasCrCard and REP_Tenure, which are not statistically significant as Pr> Chisq is more than 0.05. The selected model, based on the error rate for validation data, is the model trained in Step 3 of backward selection and consists of the following effects: *Intercept, Gender, Geography, IsActiveMember, REP-Age, REP_Balance, REP_CreditScore* and *REP_NumOfProducts.*

Results - Node: Backward Regression Diagram: Final group project

File Edit View Window

Output

```
487
488                     Summary of Backward Elimination
489
490         Effect                      Number      Wald                  Validation
491  Step   Removed                DF     In    Chi-Square  Pr > ChiSq   Error Rate
492
493     1   REP_EstimatedSalary     1      9       0.0007      0.9784      2121.5
494     2   HasCrCard               1      8       0.5724      0.4493      2121.5
495     3   REP_Tenure              1      7       3.1935      0.0739      2121.0
496
497
498  The selected model, based on the error rate for the validation data, is the model trained in Step 3. It consists of the following effects:
499
500  Intercept  Gender  Geography  IsActiveMember  REP_Age  REP_Balance  REP_CreditScore  REP_NumOfProducts
501
```

Fig 31: Summary of Backward Elimination

From the Fig 32 below, we have validation ASE is 0.134284, same as that of Forward, and Stepwise Regression Model.

Results - Node: Backward Regression Diagram: Final group project

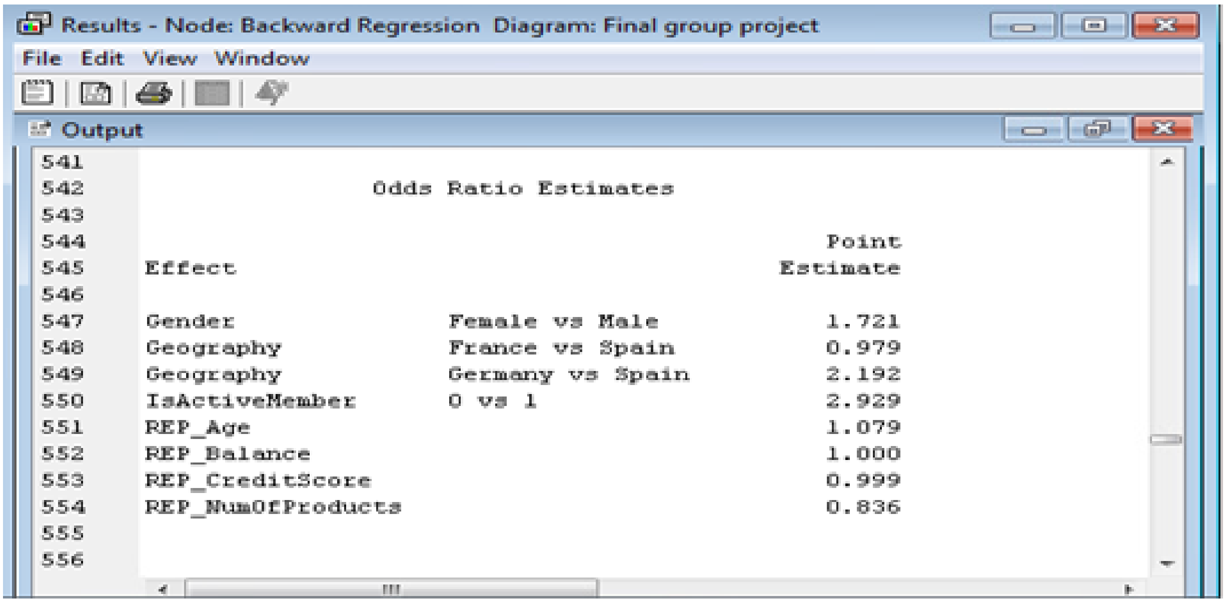
File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_AIC_	Akaike's Information Criterion	6403.804	.	.
Exited		_ASE_	Average Squared Error	0.134484	0.134284	.
Exited		_AVERR_	Average Error Function	0.425834	0.423854	.
Exited		_DFE_	Degrees of Freedom for Error	7489	.	.
Exited		_DFM_	Model Degrees of Freedom	9	.	.
Exited		_DFT_	Total Degrees of Freedom	7498	.	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_ERR_	Error Function	6385.804	2120.965	.
Exited		_FPE_	Final Prediction Error	0.134807	.	.
Exited		_MAX_	Maximum Absolute Error	0.989056	0.978195	.
Exited		_MSE_	Mean Square Error	0.134645	0.134284	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_NW_	Number of Estimate Weights	9	.	.
Exited		_RASE_	Root Average Sum of Squares	0.36672	0.366448	.
Exited		_RFPE_	Root Final Prediction Error	0.367161	.	.
Exited		_RMSE_	Root Mean Squared Error	0.366941	0.366448	.
Exited		_SBC_	Schwarz's Bayesian Criterion	6466.106	.	.
Exited		_SSE_	Sum of Squared Errors	2016.72	671.9566	.
Exited		_SUMW_	Sum of Case Weights Time...	14996	5004	.
Exited		_MISC_	Misclassification Rate	0.186583	0.185851	.

Fig 32: Fit Statistics

From Fig 33, we can interpret the result of Backward Regression Model as the same as the result of Forward and Stepwise Regression Model.



Odds Ratio Estimates

Effect		Point Estimate
Gender	Female vs Male	1.721
Geography	France vs Spain	0.979
Geography	Germany vs Spain	2.192
IsActiveMember	0 vs 1	2.929
REP_Age		1.079
REP_Balance		1.000
REP_CreditScore		0.999
REP_NumOfProducts		0.836

Fig 33: Odds Ratio Estimates

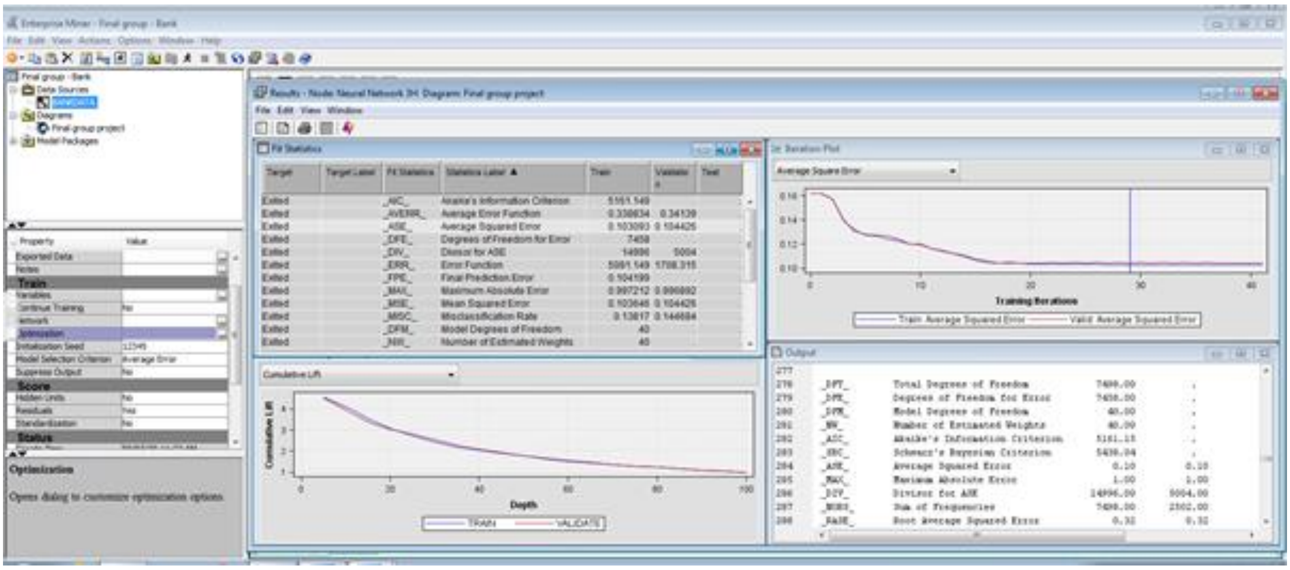
In summary, Full Regression model has the lowest validation Average squared (0.134234) in comparison with those of Forward, Backward and Stepwise Regression Models which all have the same validation Average Squared Error (0.134284). Therefore, Full Regression model is the best model among all Regression Models and its result can be interpreted as below: In the case of the variable *Gender*, we can determine that Female are 71.9% more likely to leave the bank compared to Male. In case of *Geography*, customers in Germany are more than twice as likely to leave the bank as clients in Spain. In the case of *IsActiveMember*, members who are not active are almost 3 times more likely to leave the bank than active members. In the case of *NumOfProducts*, for each additional number of products, those customers are 16.4% less likely to leave. (Fig 20)

Now we will work on neural network models to see if we can find a model with the least Average Squared Error.

Neural Network Model Groups

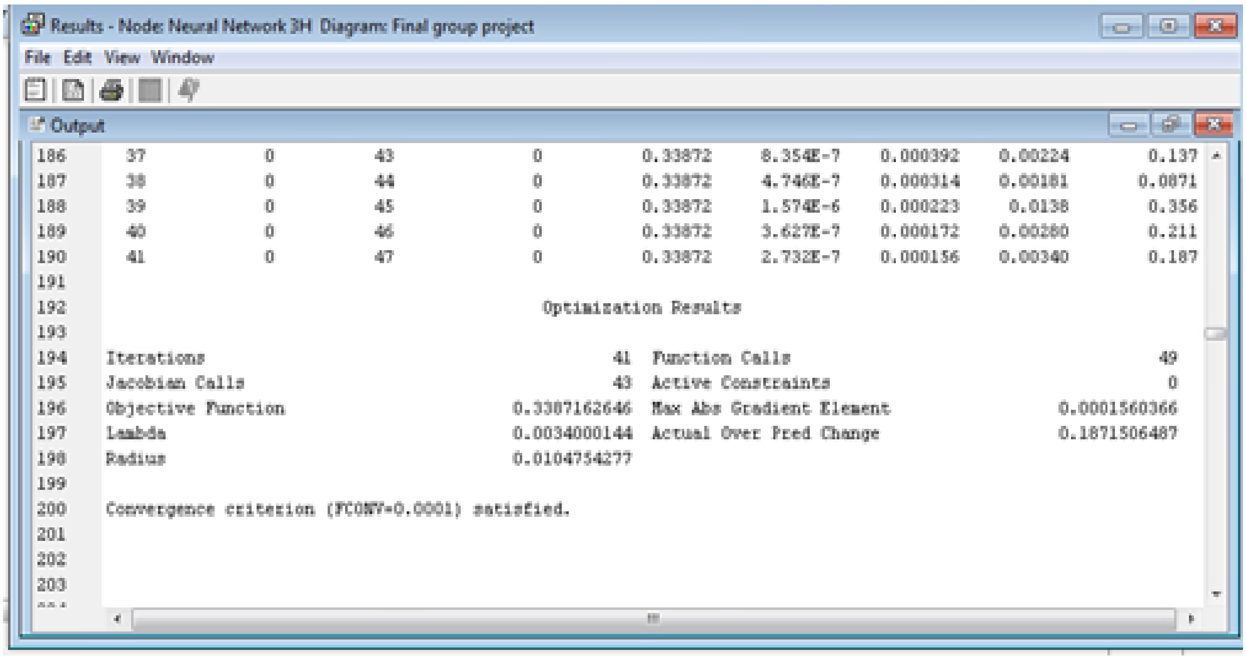
Step 12: Neural Network Model 3H

We start to build Neural Network Model 3 Hidden Units as default, with Model Selection Criterion as "*Average Error*", and "Enable preliminary training: *No*" and the result window is shown as below:

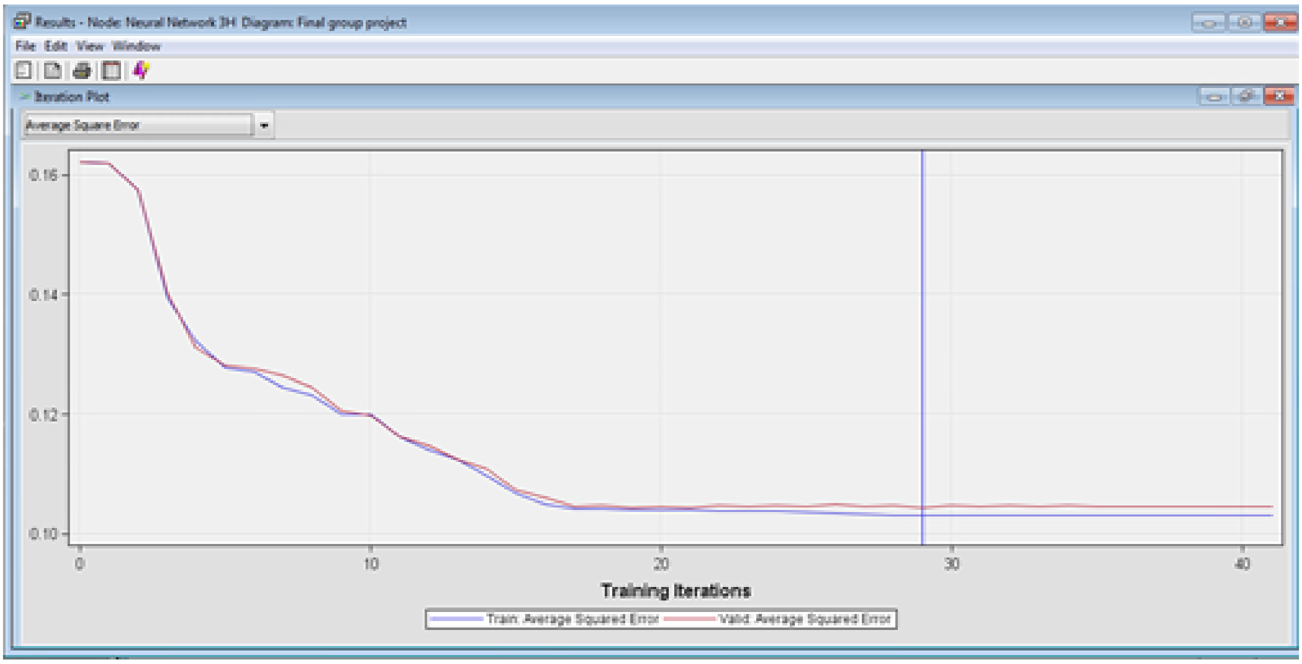


(Fig 34): Neural Network Model 3H

From Fig 35, Fig 36 and Fig 37 below, we can see that training stopped at iteration 41 and convergence criterion satisfied (Fig 35). The iteration plot shows the optimal validation Average Squared Error occurring at iteration 29 (Fig 36), with the validation ASE of 0.104426 (Fig 37). We can see that the Average Squared Error through this model is the least as compared to any other model which we tested earlier. So at this point this is the best model which can be used if we just want to know the probability of each individual of leaving the bank or not.



(Fig 35): Optimization Results



(Fig 36): Iteration Plot

Results - Node: Neural Network 3H Diagram: Final group project

File Edit View Window

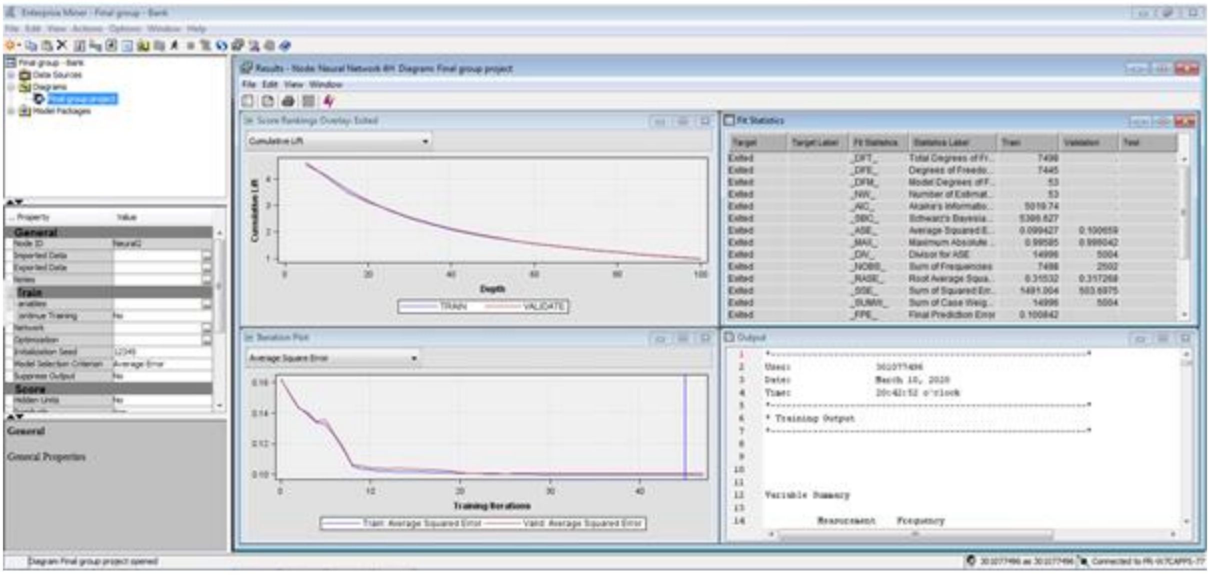
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Exited		_AIC_	Akaike's Information Criterion	5161.149	.	.
Exited		_AVERR_	Average Error Function	0.338834	0.34139	.
Exited		_ASE_	Average Squared Error	0.103093	0.104426	.
Exited		_DFE_	Degrees of Freedom for Error	7458	.	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_ERR_	Error Function	5081.149	1708.315	.
Exited		_FPE_	Final Prediction Error	0.104199	.	.
Exited		_MAX_	Maximum Absolute Error	0.997212	0.996892	.
Exited		_MSE_	Mean Squared Error	0.103646	0.104426	.
Exited		_MISC_	Misclassification Rate	0.13817	0.144684	.
Exited		_DFM_	Model Degrees of Freedom	40	.	.
Exited		_NW_	Number of Estimated Weights	40	.	.

(Fig 37): Fit Statistics

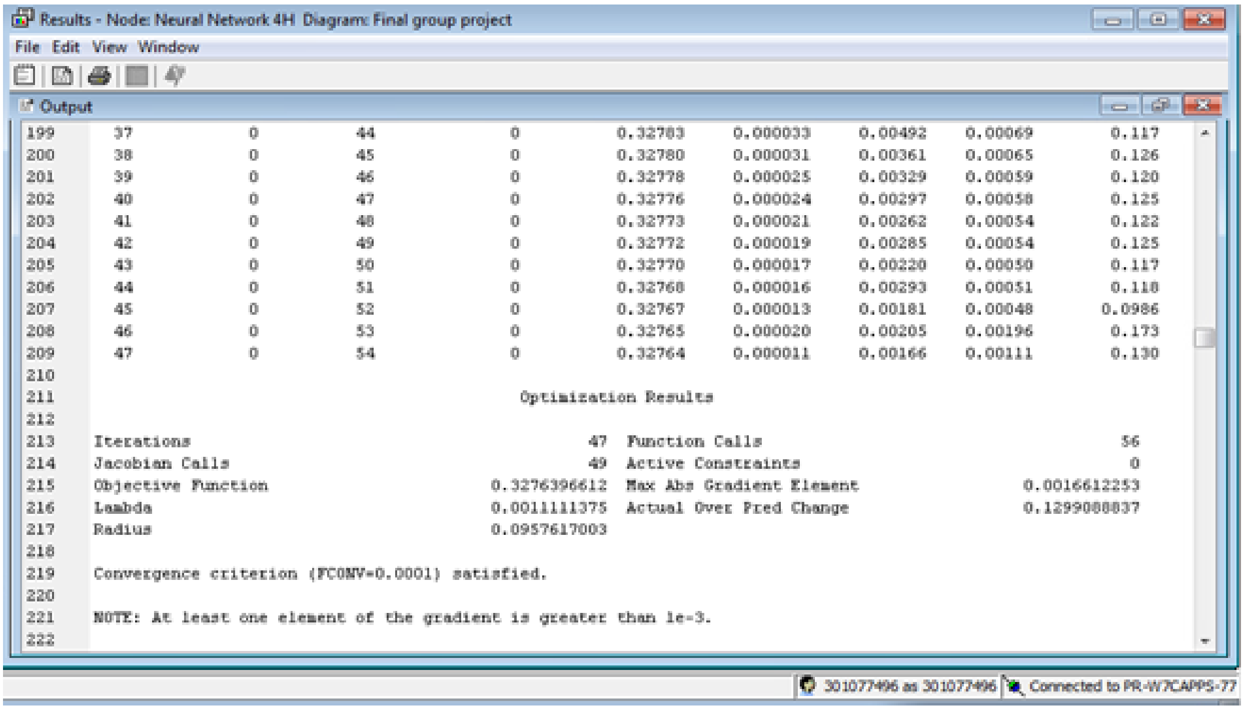
Step 13: Neural Network Model 4 H

Since we got our best model by Neural Network so we want to test the best neural network by trying with different hidden units. So In order to explore the best neural network, we continue to build Neural Network Model with 4 Hidden Units, with Model Selection Criterion as "*Average Error*", and "Enable preliminary training: *No*" and the result window is shown as below:

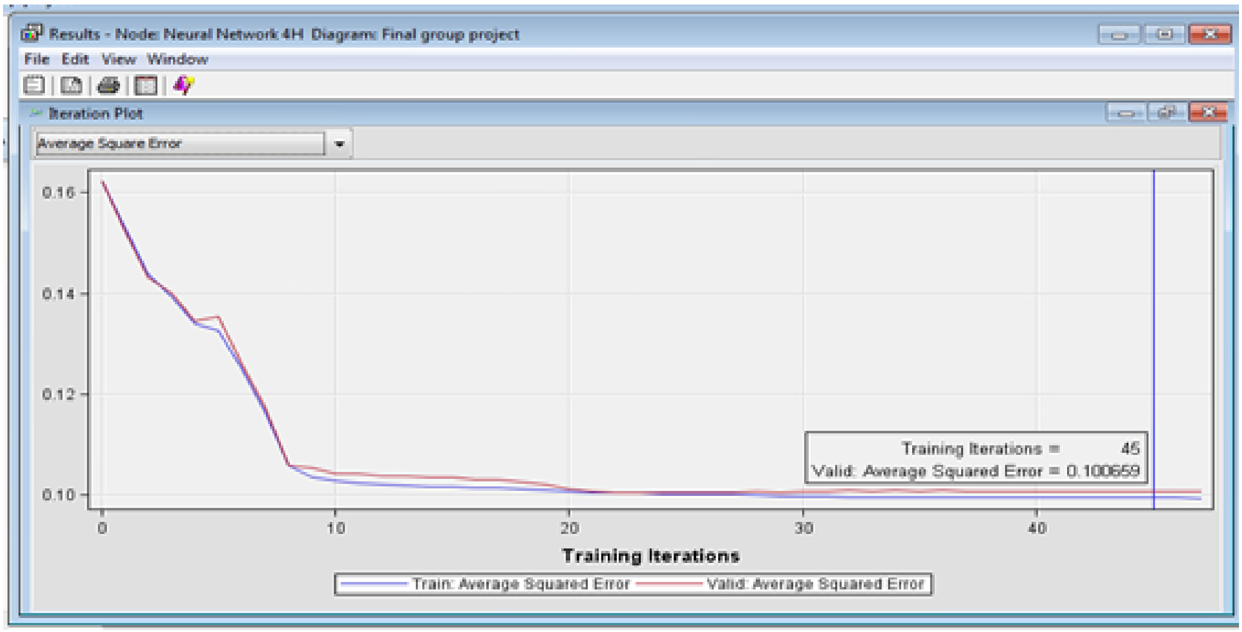


(Fig 38): Neural Network Model 4 H

From Fig 39, Fig 40 and Fig 41 below, we can see that training stopped at iteration 47 and convergence criterion satisfied (Fig 39). The iteration plot shows the optimal validation average squared error occuring at iteration 45 (Fig 40), with the validation ASE of 0.100659 (Fig 41), which is significantly lower than validation ASE of Neural network 3H (0.104426)



(Fig 39)- Optimization results



(Fig 40): Iteration Plot

Results - Node: Neural Network 4H Diagram: Final group project

File Edit View Window

Fit Statistics

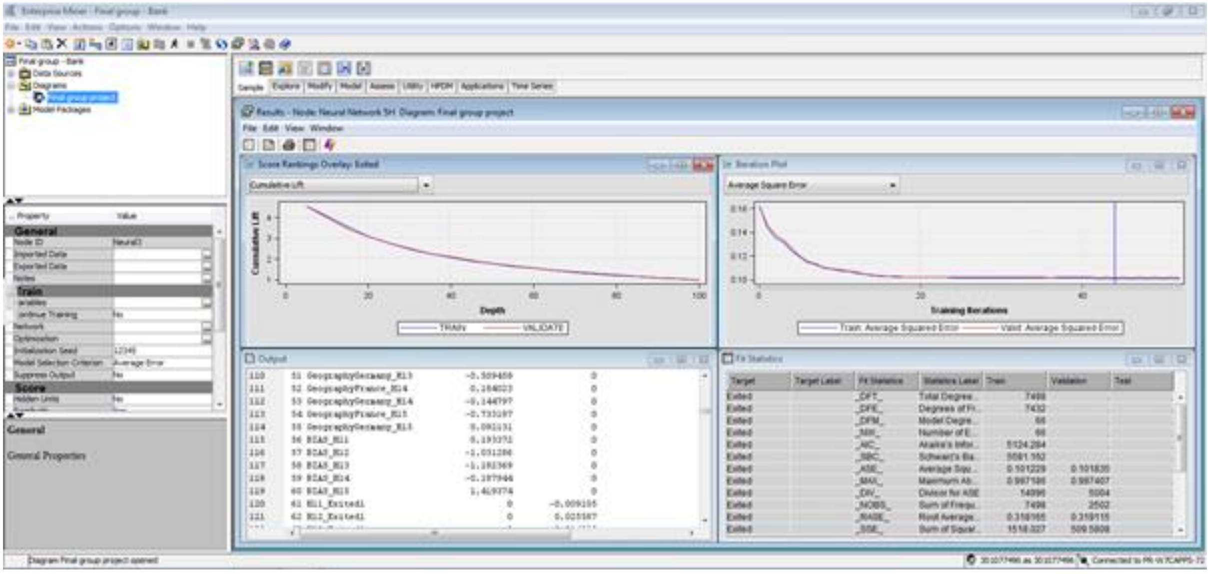
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_DFT_	Total Degrees of Freed...	7498	.	.
Exited		_DFE_	Degrees of Freedom for...	7445	.	.
Exited		_DFM_	Model Degrees of Free...	53	.	.
Exited		_NW_	Number of Estimated W...	53	.	.
Exited		_AIC_	Akaike's Information Cri...	5019.74	.	.
Exited		_SBC_	Schwarz's Bayesian Crit...	5386.627	.	.
Exited		_ASE_	Average Squared Error	0.099427	0.100659	.
Exited		_MAX_	Maximum Absolute Error	0.99585	0.996042	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_RASE_	Root Average Squared ...	0.31532	0.317268	.
Exited		_SSE_	Sum of Squared Errors	1491.004	503.6975	.
Exited		_SUMW_	Sum of Case Weights T...	14996	5004	.
Exited		_FPE_	Final Prediction Error	0.100842	.	.
Exited		_MSE_	Mean Squared Error	0.100135	0.100659	.
Exited		_RFPE_	Root Final Prediction Er...	0.317557	.	.
Exited		_RMSE_	Root Mean Squared Error	0.316441	0.317268	.

(Fig 41): Fit Statistics

So up to this point, our best model is the Neural Network with 4 Hidden units.

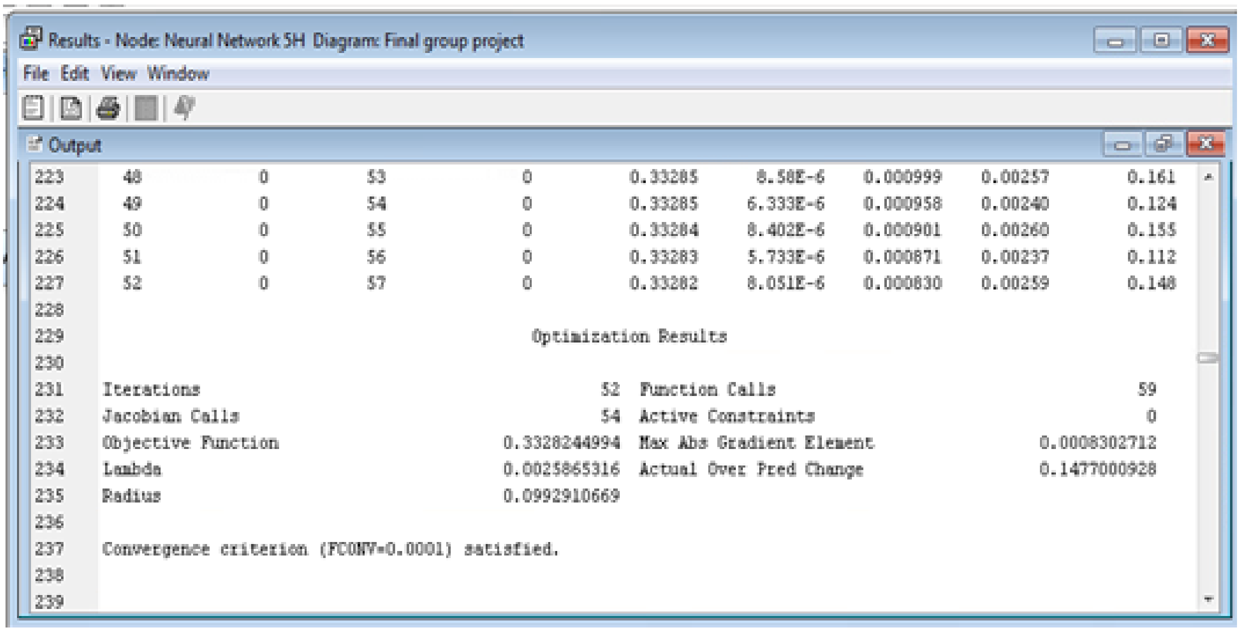
Step 14: Neural Network 5H

Neural Network 4H is better model than Neural Network 3H with lower validation ASE so we continue to build Neural Network Model 5 Hidden Units to find whether Neural Network 5H units is better than 4H or not, with Model Selection Criterion as "*Average Error*", and "Enable preliminary training: *No"* and the result window is shown as below:

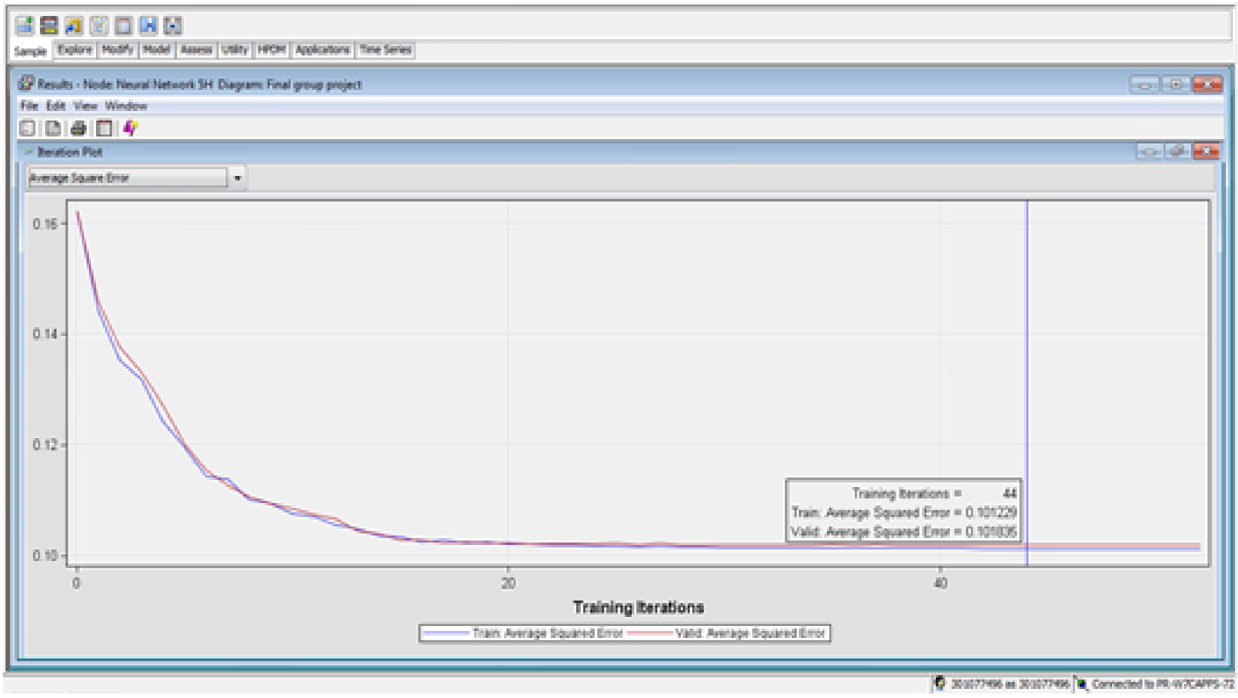


(Fig 42): Neural Network Model 5H

From Fig 43, Fig 44 and Fig 45 below, we can see that the training stopped at iteration 52 and convergence criterion satisfied (Fig 43). The iteration plot shows the optimal validation Average Squared Error occurring at iteration 44 (Fig 44), with the validation ASE of 0.101835 (Fig 41), which is lower than validation ASE of Neural network 3H of 0.10442 but higher than validation ASE of Neural network 4H of 0.100659. Therefore, among three neural network models, the neural network model with 4 hidden units is the best model.



(Fig 43): Optimization Results



(Fig 44): Iteration Plot

Fit Statistics

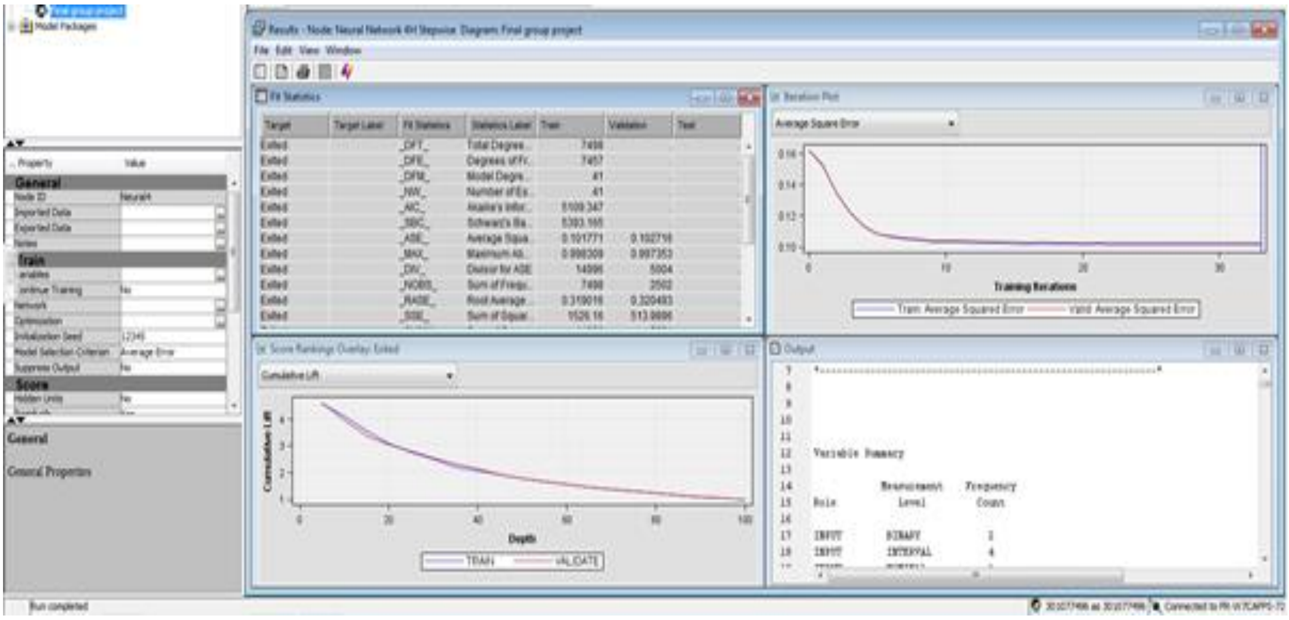
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_DFT_	Total Degrees of Fre...	7498	.	.
Exited		_DFE_	Degrees of Freedo...	7432	.	.
Exited		_DFM_	Model Degrees of Fr...	66	.	.
Exited		_NW_	Number of Estimate...	66	.	.
Exited		_AIC_	Akaike's Information...	5124.284	.	.
Exited		_SBC_	Schwarz's Bayesian ...	5581.162	.	.
Exited		_ASE_	Average Squared Er...	0.101229	0.101835	.
Exited		_MAX_	Maximum Absolute ...	0.997186	0.997407	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_RASE_	Root Average Squar...	0.318165	0.319115	.
Exited		_SSE_	Sum of Squared Err...	1518.027	509.5808	.
Exited		_SUMW_	Sum of Case Weigh...	14996	5004	.
Exited		_FPE_	Final Prediction Error	0.103027	.	.
Exited		_MSE_	Mean Squared Error	0.102128	0.101835	.
Exited		_RFPE_	Root Final Predictio...	0.320978	.	.
Exited		_RMSE_	Root Mean Squared ...	0.319574	0.319115	.
Exited		_AVERR_	Average Error Functi...	0.332908	0.336079	.
Exited		_ERR_	Error Function	4992.284	1681.74	.
Exited		_MISC_	Misclassification Rate	0.134036	0.140288	.
Exited		_WRONG_	Number of Wrong Cl...	1005	351	.

(Fig 45): Fit Statistics

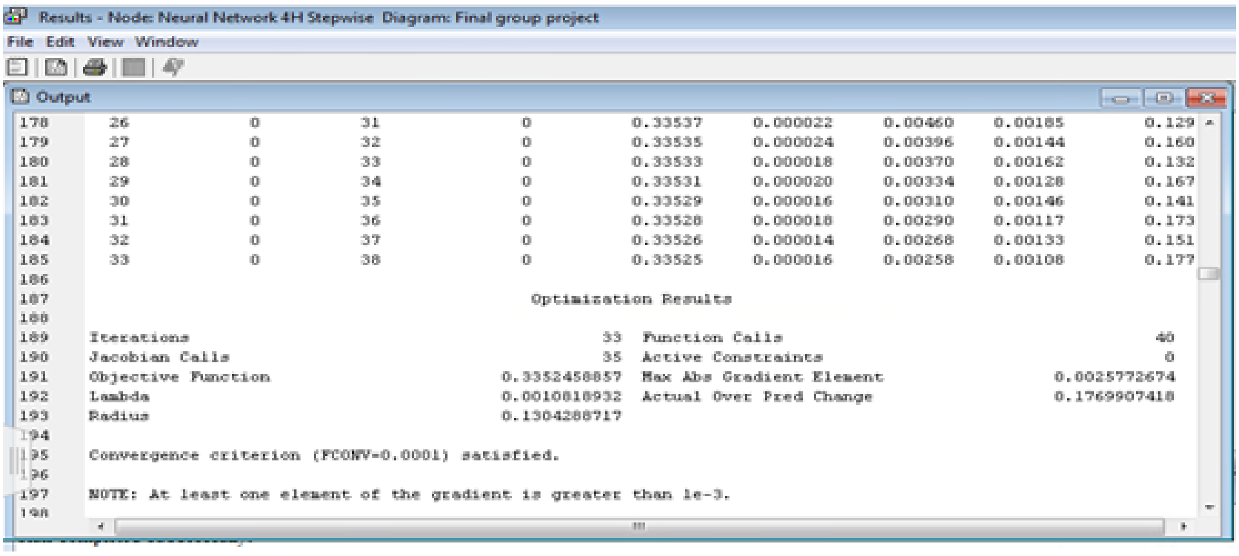
For the Neural Network Model Group, we started with 3 hidden units, then 4 hidden units and saw ASE coming down, so we continued with 5 hidden units but its validation ASE was going up so we stopped trying other neural network models with different hidden units.

Step 15: Neural Network 4H Stepwise model

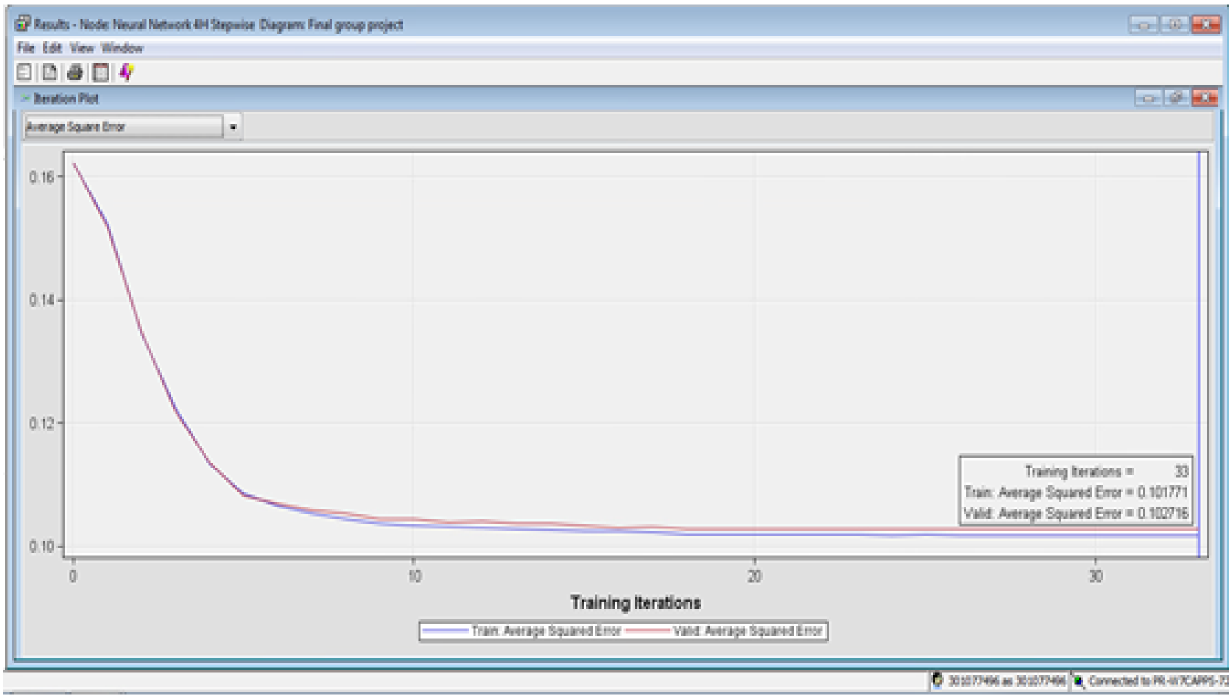
In order to take advantage of the statistically significant variables selected by Stepwise Regression Model, we build another Neural Network with 4 hidden units (because Neural network 4H is the best one among 3 neural network models built before) based on Stepwise Regression Model with model selection criterion is "*Average Error*" and the result is shown as below:



(Fig 46): Neural Network 4H Stepwise



(Fig 47): Optimization Results



(Fig 48): Iteration Plot

Results - Node: Neural Network 4H Stepwise Diagram: Final group project

File Edit View Window

Fit Statistics

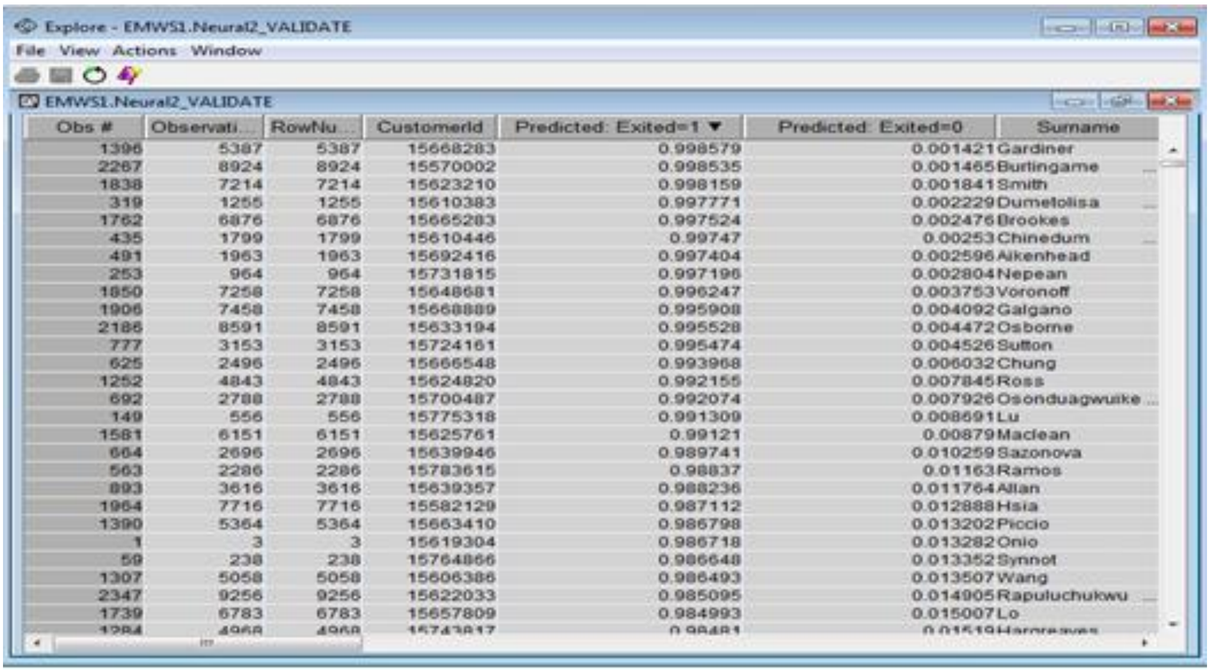
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_DFT_	Total Degrees of Freedom	7498	.	.
Exited		_DFE_	Degrees of Freedom for Error	7457	.	.
Exited		_DFM_	Model Degrees of Freedom	41	.	.
Exited		_NW_	Number of Estimated Weights	41	.	.
Exited		_AIC_	Akaike's Information Criterion	5109.347	.	.
Exited		_SBC_	Schwarz's Bayesian Criterion	5393.165	.	.
Exited		_ASE_	Average Squared Error	0.101771	0.102716	.
Exited		_MAX_	Maximum Absolute Error	0.998309	0.997353	.
Exited		_DIV_	Divisor for ASE	14996	5004	.
Exited		_NOBS_	Sum of Frequencies	7498	2502	.
Exited		_RASE_	Root Average Squared Error	0.319016	0.320493	.
Exited		_SSE_	Sum of Squared Errors	1526.16	513.9896	.
Exited		_SUMW_	Sum of Case Weights Times Freq	14996	5004	.
Exited		_FPE_	Final Prediction Error	0.10289	.	.
Exited		_MSE_	Mean Squared Error	0.102331	0.102716	.
Exited		_RFPE_	Root Final Prediction Error	0.320765	.	.
Exited		_RMSE_	Root Mean Squared Error	0.319892	0.320493	.
Exited		_AVERR_	Average Error Function	0.335246	0.33484	.
Exited		_ERR_	Error Function	5027.347	1675.539	.

(Fig 49): Fit Statistics

From Fig 47, Fig 48 and Fig 49 below, we can see that training stopped at iteration 33 and convergence criterion satisfied (Fig 47). The iteration plot shows the optimal validation average squared error occurring at iteration 33 (Fig 48), with the validation ASE of 0.102716 (Fig 49), which is higher than the validation ASE of Neural network 3H model (0.10442) but smaller than those of Neural Network 4H (0.100659) and Neural Network 5H Models (0.101835).

In summary, among all four neural network models, the neural network model with 4 hidden units is still the best model because it has the lowest validation average squared error (0.100659), in comparison with Neural network 3H model (0.10442) and Neural network 5H model (0.101835) and Neural Network 4H Stepwise model (0.102716)

From the Fig 50 below, we explore the validation dataset to see the result of prediction of probability of exiting for each specific customer. Predicted probability of exiting is listed in descending order. For example, Customer with Surname "Gardiner" and CustomerId "15668283" has 99.8% of leaving our bank.



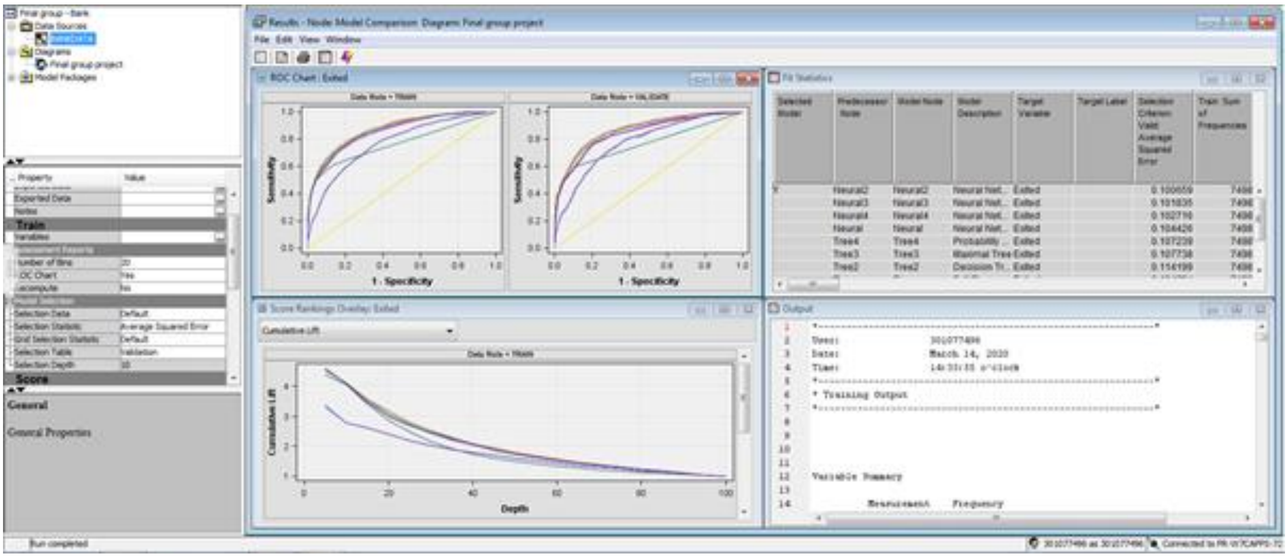
Obs #	Observati...	RowNu...	CustomerId	Predicted: Exited=1 ▼	Predicted: Exited=0	Surname
1396	5387	5387	15668283	0.998579	0.001421	Gardiner
2267	8924	8924	15570002	0.998535	0.001465	Burlingame
1838	7214	7214	15623210	0.998159	0.001841	Smith
319	1255	1255	15610383	0.997771	0.002229	Dumetolisa
1762	6876	6876	15665283	0.997524	0.002476	Brookes
435	1799	1799	15610446	0.99747	0.00253	Chinedum
491	1963	1963	15692416	0.997404	0.002596	Aikenhead
253	964	964	15731815	0.997196	0.002804	Nepean
1850	7258	7258	15648681	0.996247	0.003753	Voronoff
1906	7458	7458	15668889	0.995908	0.004092	Galgano
2186	8591	8591	15633194	0.995528	0.004472	Osborne
777	3153	3153	15724161	0.995474	0.004526	Sutton
625	2496	2496	15666548	0.993968	0.006032	Chung
1252	4843	4843	15624820	0.992155	0.007845	Ross
692	2788	2788	15700487	0.992074	0.007926	Osonduagwuike ...
149	556	556	15775318	0.991309	0.008691	Lu
1581	6151	6151	15625761	0.99121	0.00879	Maclean
664	2696	2696	15639946	0.989741	0.010259	Sazonova
563	2286	2286	15783615	0.98837	0.01163	Ramos
893	3616	3616	15639357	0.988236	0.011764	Allan
1964	7716	7716	15582129	0.987112	0.012888	Hsia
1390	5364	5364	15663410	0.986798	0.013202	Piccio
1	3	3	15619304	0.986718	0.013282	Onio
59	238	238	15764866	0.986648	0.013352	Synnot
1307	5058	5058	15606386	0.986493	0.013507	Wang
2347	9256	9256	15622033	0.985095	0.014905	Rapuluchukwu ...
1739	6783	6783	15657809	0.984993	0.015007	Lo
1284	4968	4968	15743817	0.98481	0.01519	Hargreaves

(Fig 50): Validate

Model Comparison

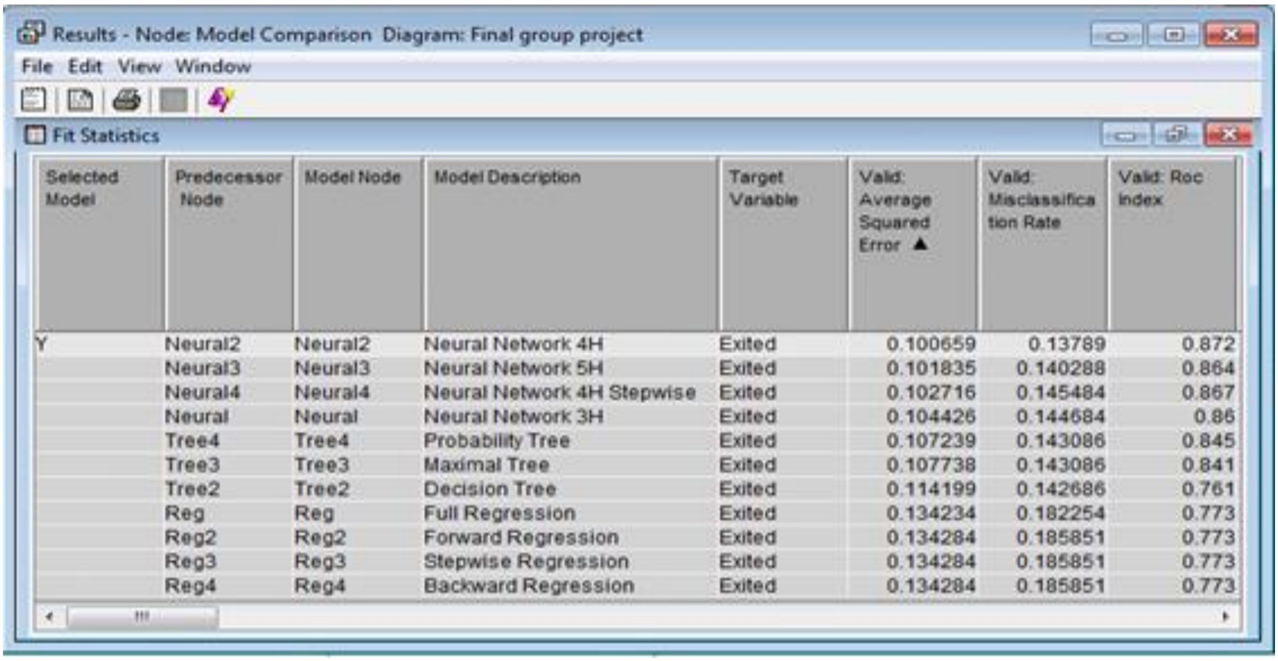
Step 16: Model Comparison

In order to figure out which model is the best one to predict our target, we use model comparison node with: Selection statistic is "*Validation squared error*" and Selection table is "*Validation*" and result is shown as below:



(Fig 51): Model Comparison

From the Fig 52 below, we can compare the validation average squared error of each model.



Results - Node: Model Comparison Diagram: Final group project

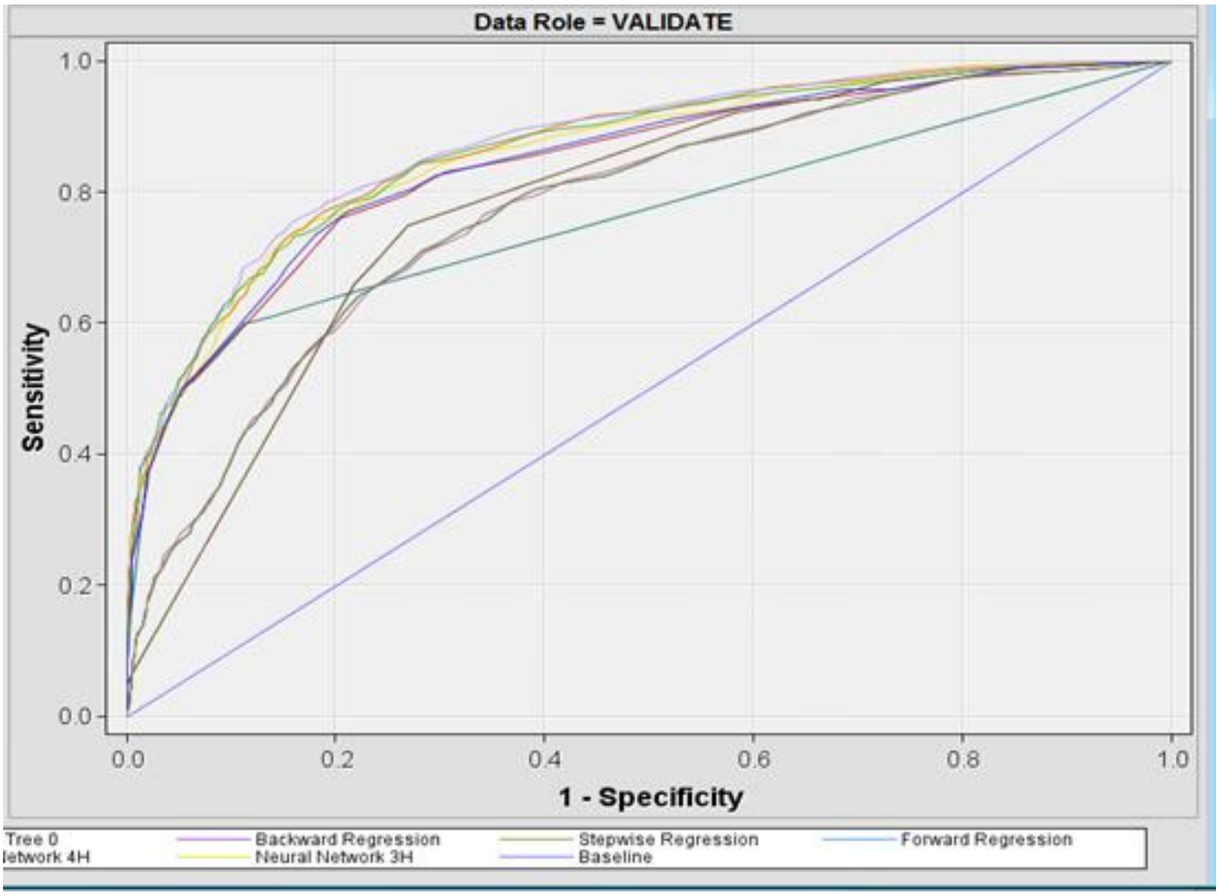
File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Average Squared Error ▲	Valid: Misclassification Rate	Valid: Roc Index
Y	Neural2	Neural2	Neural Network 4H	Exited	0.100659	0.13789	0.872
	Neural3	Neural3	Neural Network 5H	Exited	0.101835	0.140288	0.864
	Neural4	Neural4	Neural Network 4H Stepwise	Exited	0.102716	0.145484	0.867
	Neural	Neural	Neural Network 3H	Exited	0.104426	0.144684	0.86
	Tree4	Tree4	Probability Tree	Exited	0.107239	0.143086	0.845
	Tree3	Tree3	Maximal Tree	Exited	0.107738	0.143086	0.841
	Tree2	Tree2	Decision Tree	Exited	0.114199	0.142686	0.761
	Reg	Reg	Full Regression	Exited	0.134234	0.182254	0.773
	Reg2	Reg2	Forward Regression	Exited	0.134284	0.185851	0.773
	Reg3	Reg3	Stepwise Regression	Exited	0.134284	0.185851	0.773
	Reg4	Reg4	Backward Regression	Exited	0.134284	0.185851	0.773

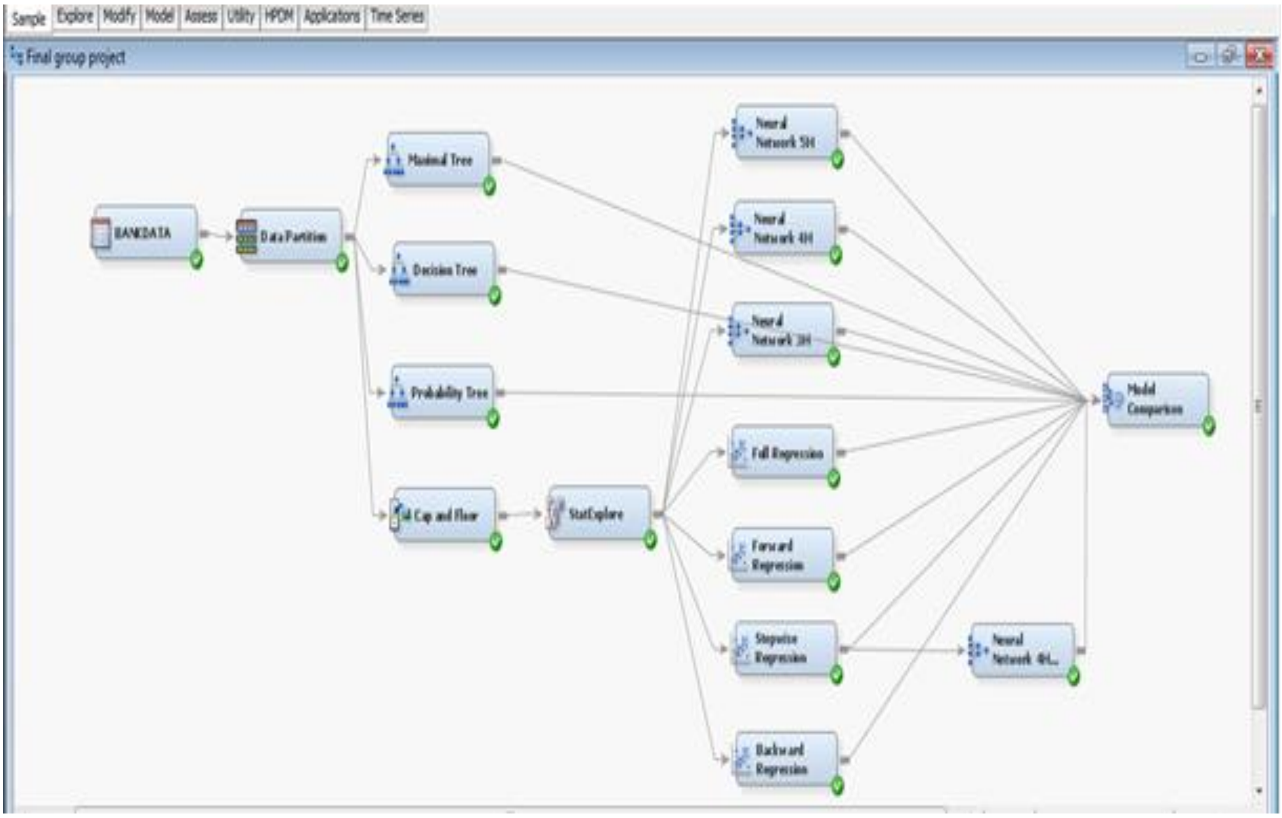
(Fig 52): Fit Statistics

For this project, we make a decision based on average square error because we are using probability as our goal. Based on Fig 52, Neural Network 4H (4 hidden units) is selected as the best model because it has the lowest validation average squared error (0.100659), and it's also concordant with its validation misclassification rate and ROC index because it also has the lowest validation Misclassification rate (0.13789) and highest validation ROC index (0.872). We can also see the ROC index in the Fig. 53 below.



(Fig 53): Validation ROC Index

Fig. 54 is showing our complete diagram with all the nodes. From this figure, we can see each step taken by us from adding the data to the final Model Comparison Node.



(Fig 54): Final Diagram

Conclusion

Based on the results of the different models and keeping in mind our goal, we have reached to the conclusion that if our goal is to predict who is going to leave, then the Neural Network model with 4 hidden units is the best model to find the probability of leaving for each individual. From Neural Network we can find out the ID of each individual and then can focus on those customers by running some campaigns for these customers.

However, if we want to know the particular segments of the people so that the bank can focus on those particular segments then we can use the probability tree since it has the lowest ASE of 0.107239 among the various models that can be used to explain the results and describe the characteristics of targeted segments. We have the following findings: people who have only 1 product, not active members, more than 47 years old have at least 79% chance to leave the bank. Meanwhile, people who are young (under 42.5 years old) and have 2 products have 95% chance to stay with the bank. In the case of geography, people who are young, have 1 product, are living in France or Spain, and have a high balance (over $65,000) have 90% chance to stay with the bank. So based on these segments, the bank can take certain initiatives based on its priorities and needs.

Data Source

https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling