GROUP PROJECT

BA 706

PROFESSOR: DAVID PARENT

PREPARED BY:

SAMSON OGWUCHE: 301196569

CHARLES NTAMACK: 301209795

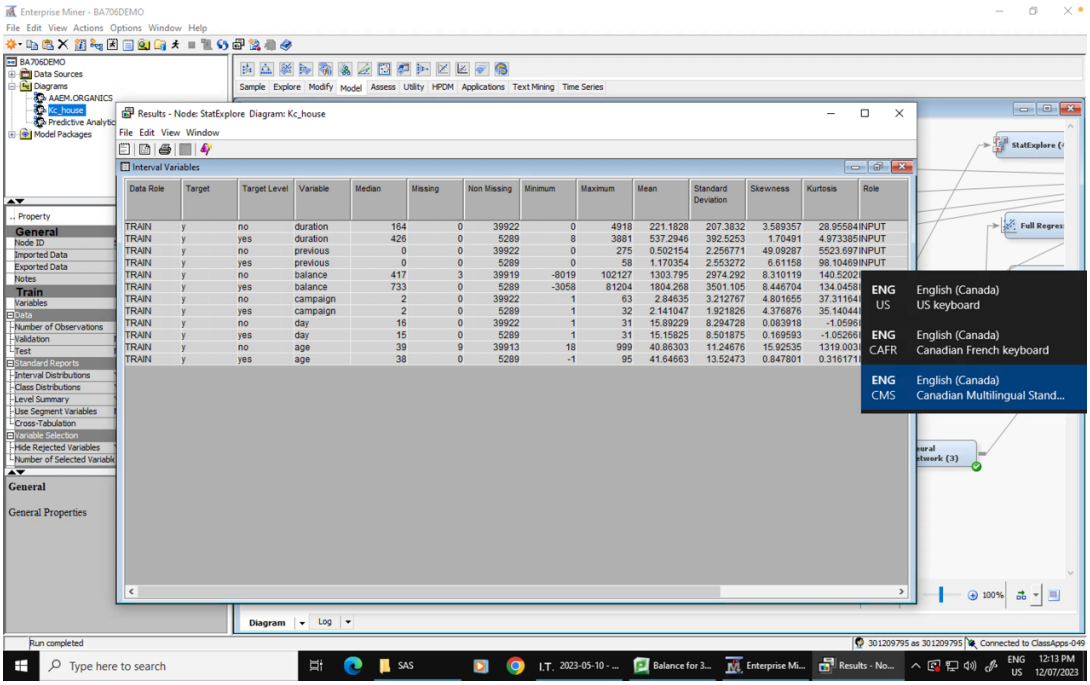
Introduction

The aim of the model is to provide a comprehensive analysis of the bank loan campaign and highlight the key variables that determined customers' acceptance or rejection of the loan campaign. The dataset is about the campaign carried out by a bank that was intended to get its customers to subscribe to bank loans. The dataset has 17 variables and 45,211 observations. There are several variables of interest in the dataset. Below is the dictionary to define each of the 17 variables in the dataset.

1. Age: Age in the dataset is a numeric variable that describes the ages of the customers in the loan campaign.
2. Job: This refers to the profession of the bank customers targeted for the loan campaign.
3. Marital status: This refers to the marital status of the customers contacted for the bank loan campaign.
4. Education: Education describes the level of education of the customers contacted.
5. Default: Default describes if a customer has had a credit default previously.
6. Balance: Balance relates to the average yearly euro balances in the customers' accounts.
7. Housing: This refers to whether a customer has a housing loan.
8. Loan: Loan refers if a customer has an existing personal loan.
9. Contact: This refers to the medium of communication with the customers.
10. Day: Day refers to the last contact day of the month.
11. Month: Month refers to the last contact month of the year.
12. Duration: This refers to the last contact duration in seconds.
13. Campaign: Campaign refers to the number of contacts performed during the campaign and with each customer.
14. Pdays: This refers to the number of days that passed after the customer was last contacted.
15. Previous: Previous refers to the number of campaigns performed before the current campaign and with each customer.
16. Poutcome: This refers to the outcome of the previous marketing campaigns performed.
17. Y: This is the binary variable in the dataset that refers to customers' responses which is either a yes or a no. This is our target variable in the model.

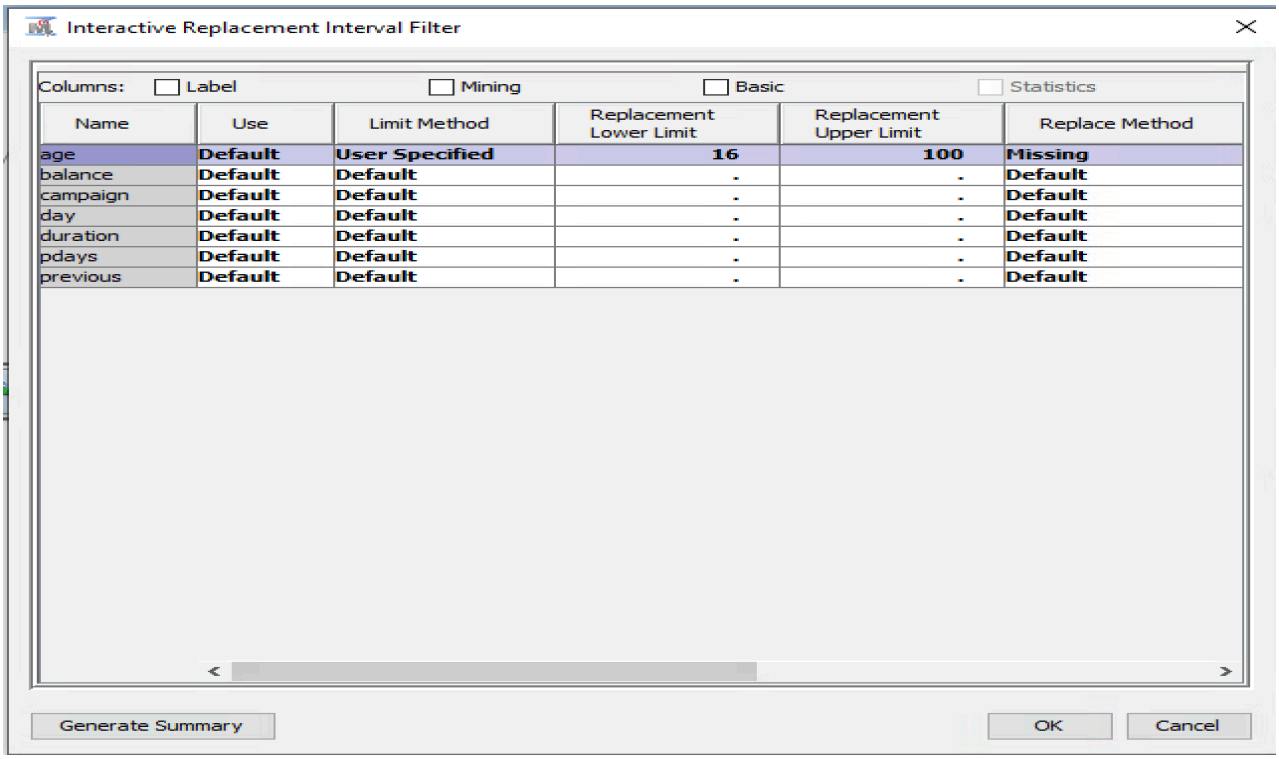
Exploring the data source:

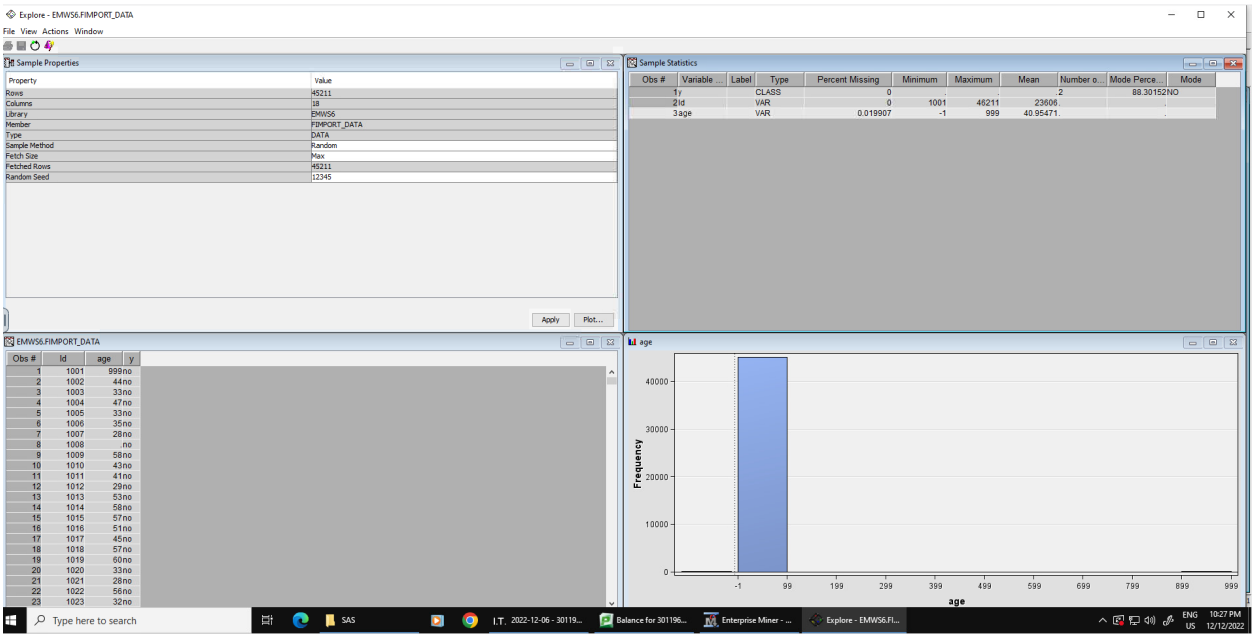
We explored the data source to see the properties of the sample. We changed the setting of the sample method from top to random and fetch size from default to maximum. From the exploration, we observed the mean age of the customers who participated in the campaign is 40.95471 years. This implies that the loan campaign was targeted at customers who are in their youthful age.

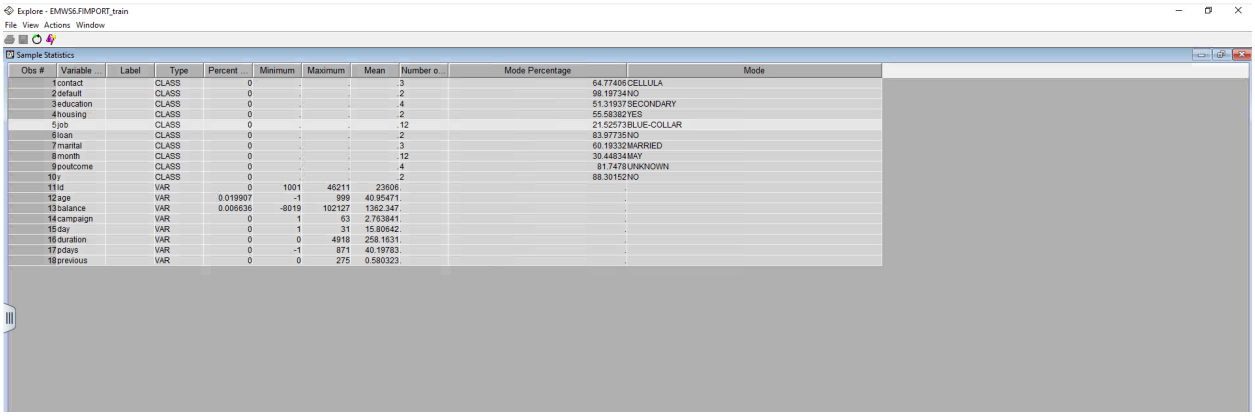


We have a normal range in our dataset except for age that has -1 which has been corrected by setting the age range between 16 and 100.

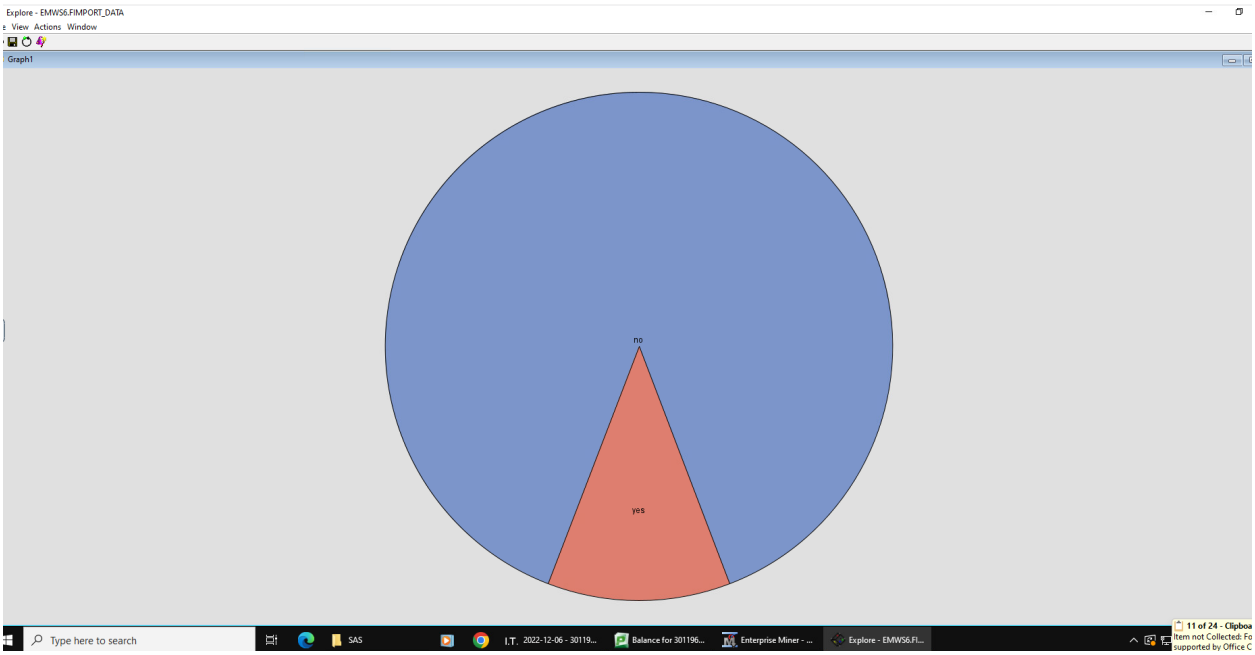
See the screenshot below for the normal age range of 16 to 100.


Interactive Replacement Interval Filter
Columns: Label Mining Basic Statistics
Name | Use | Limit Method | Replacement Lower Limit | Replacement Upper Limit | Replace Method
age | Default | User Specified | 16 | 100 | Missing
balance | Default | Default | . | . | Default
campaign | Default | Default | . | . | Default
day | Default | Default | . | . | Default
duration | Default | Default | . | . | Default
pdays | Default | Default | . | . | Default
previous | Default | Default | . | . | Default
Generate Summary
OK
Cancel


Explore - EMWS6.FIMPORT_DATA
File View Actions Window
Sample Properties
Sample Statistics
EMWS6.FIMPORT_DATA
age
Frequency
age



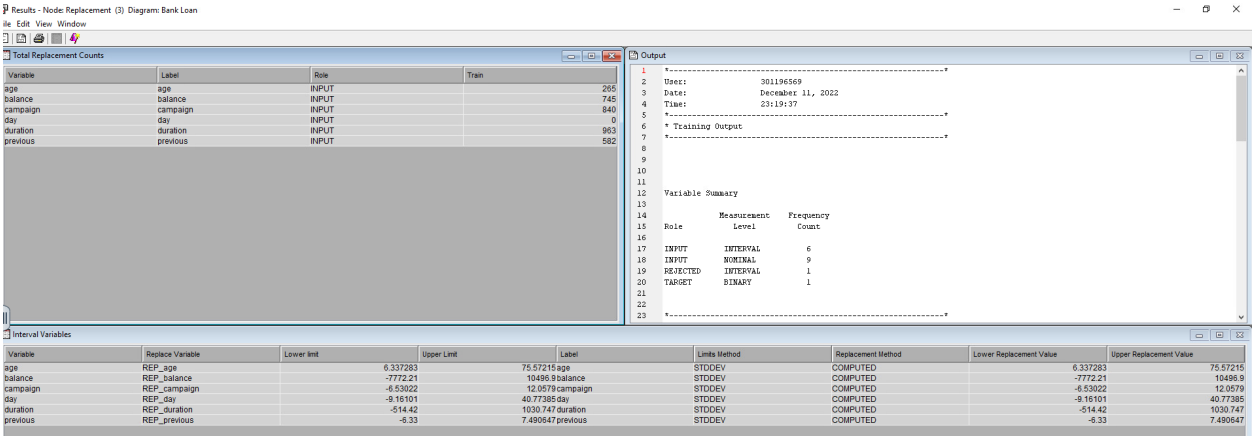
Obs #	Variable	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Percentage	Mode
1	contact		CLASS	0				3	64.77406	CELLULA
2	default		CLASS	0				2	98.19734	NO
3	education		CLASS	0				4	51.31937	SECONDARY
4	housing		CLASS	0				2	55.58382	YES
5	job		CLASS	0				12	21.52573	BLUE-COLLAR
6	loan		CLASS	0				2	83.97735	NO
7	marital		CLASS	0				3	60.19332	MARRIED
8	month		CLASS	0				12	30.44834	MAY
9	poutcome		CLASS	0				4	81.7478	UNKNOWN
10	y		CLASS	0				2	88.30152	NO
11	id		VAR	0	1001	46211	23606.			
12	age		VAR	0.019907	-1	999	40.95471.			
13	balance		VAR	0.006636	-8019	102127	1362.347.			
14	campaign		VAR	0	1	63	2.763841.			
15	day		VAR	0	1	31	15.80642.			
16	duration		VAR	0	0	4918	258.1631.			
17	pdays		VAR	0	-1	871	40.19783.			
18	previous		VAR	0	0	275	0.580323.			



From the pie chart above, the percentage of customers who said yes to the loan campaign is 11.6984% while that of no is 88.3015%.

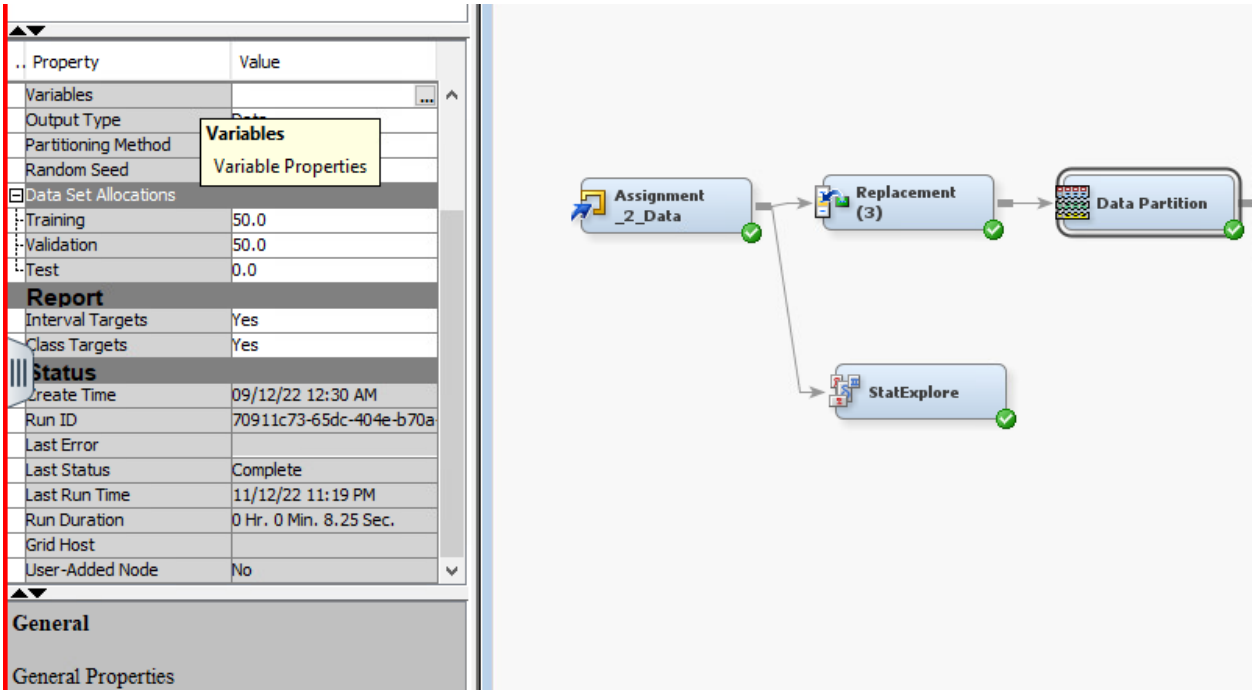
Replacement:

The dataset has some missing values and to replace them, we did imputation for the missing values. The replacement node was used to perform the task of replacing the missing values by replacing all the lower limit's standard deviations with the upper limit standard deviation values.


Results - Node: Replacement (3) Diagram: Bank Loan
File Edit View Window
Total Replacement Counts
Variable Label Role Train
age age INPUT 265
balance balance INPUT 745
campaign campaign INPUT 840
day day INPUT 0
duration duration INPUT 963
previous previous INPUT 582
Output
Training Output
Variable Summary
Role Measurement Level Frequency Count
INPUT INTERVAL 6
INPUT NOMINAL 9
REJECTED INTERVAL 1
TARGET BINARY 1
Interval Variables
Variable Replace Variable Lower limit Upper Limit Label Limits Method Replacement Method Lower Replacement Value Upper Replacement Value
age REP_age 6.337283 75.57215 age STDDEV COMPUTED 6.337283 75.57215
balance REP_balance -7772.21 10496.9 balance STDDEV COMPUTED -7772.21 10496.9
campaign REP_campaign -6.53022 12.0579 campaign STDDEV COMPUTED -6.53022 12.0579
day REP_day -9.16101 40.77385 day STDDEV COMPUTED -9.16101 40.77385
duration REP_duration -514.42 1030.747 duration STDDEV COMPUTED -514.42 1030.747
previous REP_previous -6.33 7.490647 previous STDDEV COMPUTED -6.33 7.490647

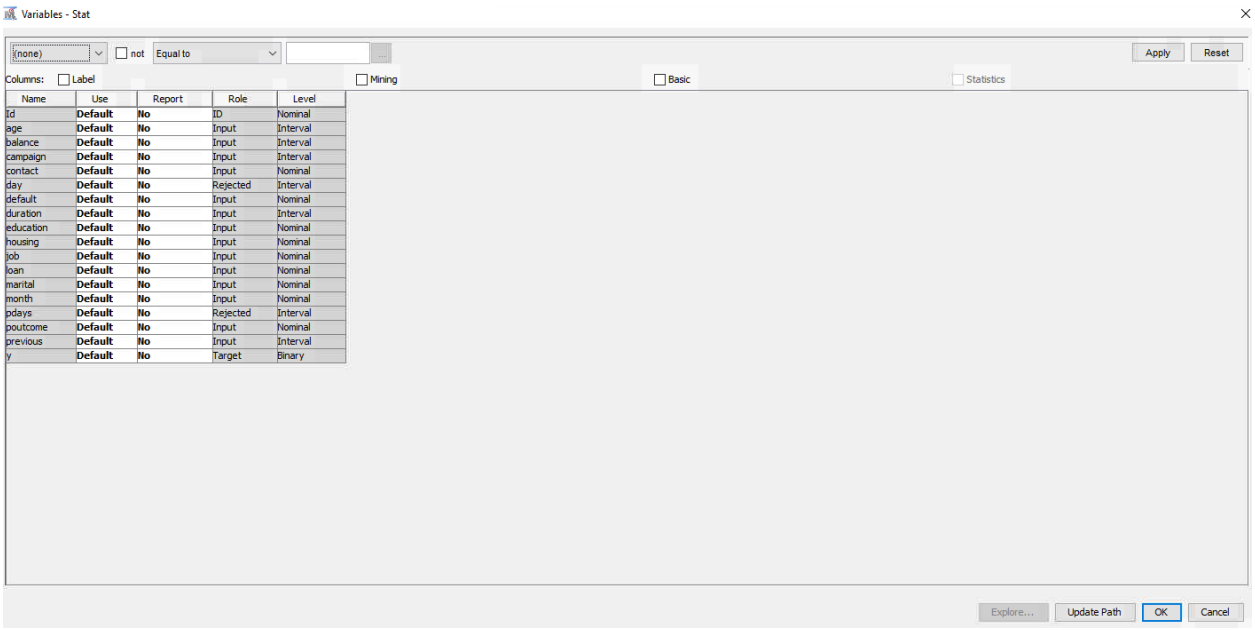
Data Participation:

We partitioned the data into two sets – training and validation. The training set is 50% of the dataset while validation is also 50% of the dataset.



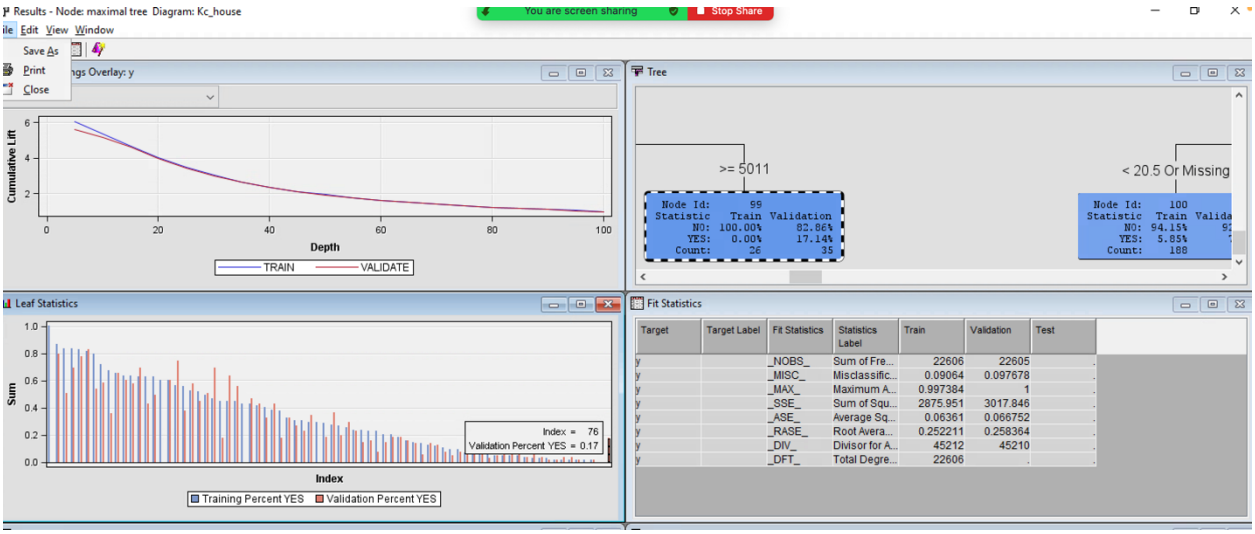
StatExplore.

We used the StatExplore update to reject two of the variables that are not relevant to the model and also used it to explore the skewness of the variables. The variables we rejected are day and pdays. The variable ‘day’ in the dataset represents the last contact day of the month while the variable ‘plays’ represents the last day that passed after the customer was contacted from the previous campaign. The two variables were dropped because the dataset was skewed with the two variables, day and pday. The result was not good because of the two variables in the dataset; hence we dropped them.



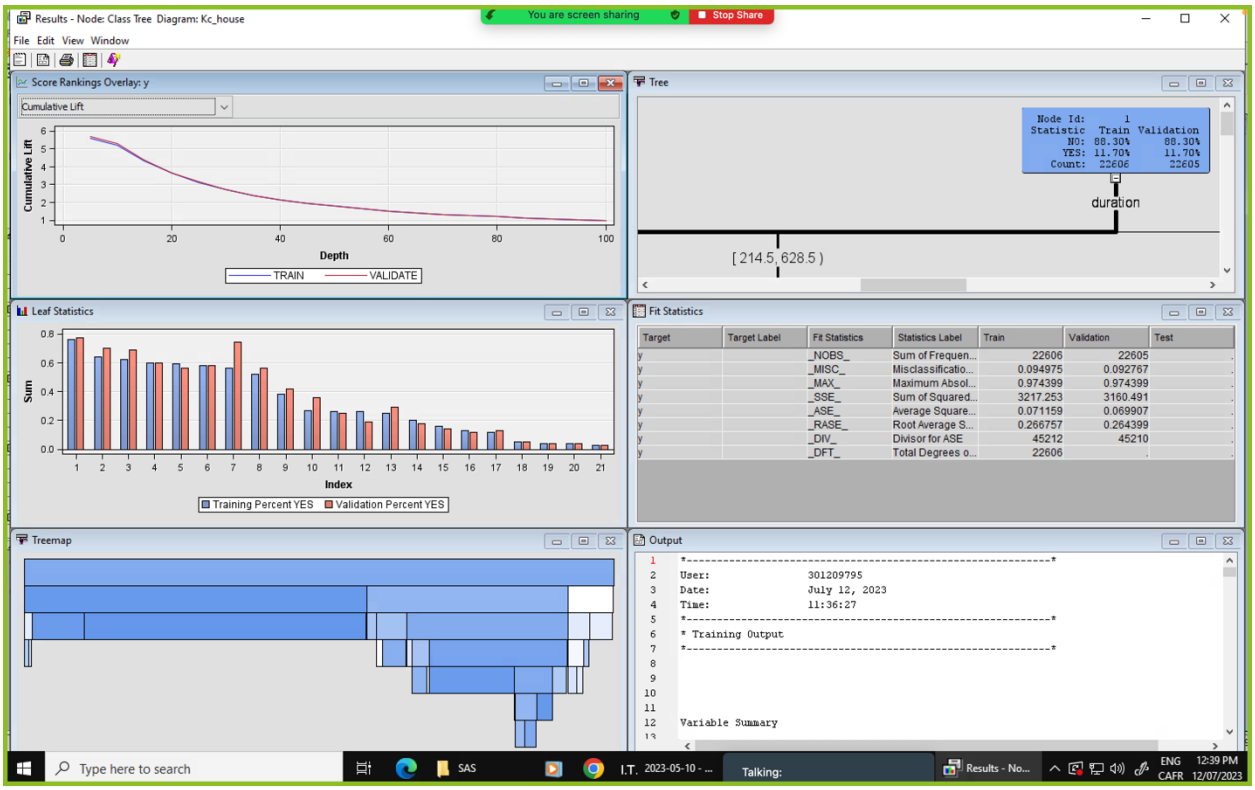
Decision Tree statistics

We built three decision trees using the average squared error (ASE) and misclassification as the assessment measures. Our first decision tree is the Maximum decision tree and we used 'largest' as the subtree method with maximum branch of 3. The Maximum decision tree has 76 leaf statistics and an ASE of **0.066752**. However, our target was to get an optimal ASE and so we have to run the optimal decision tree.



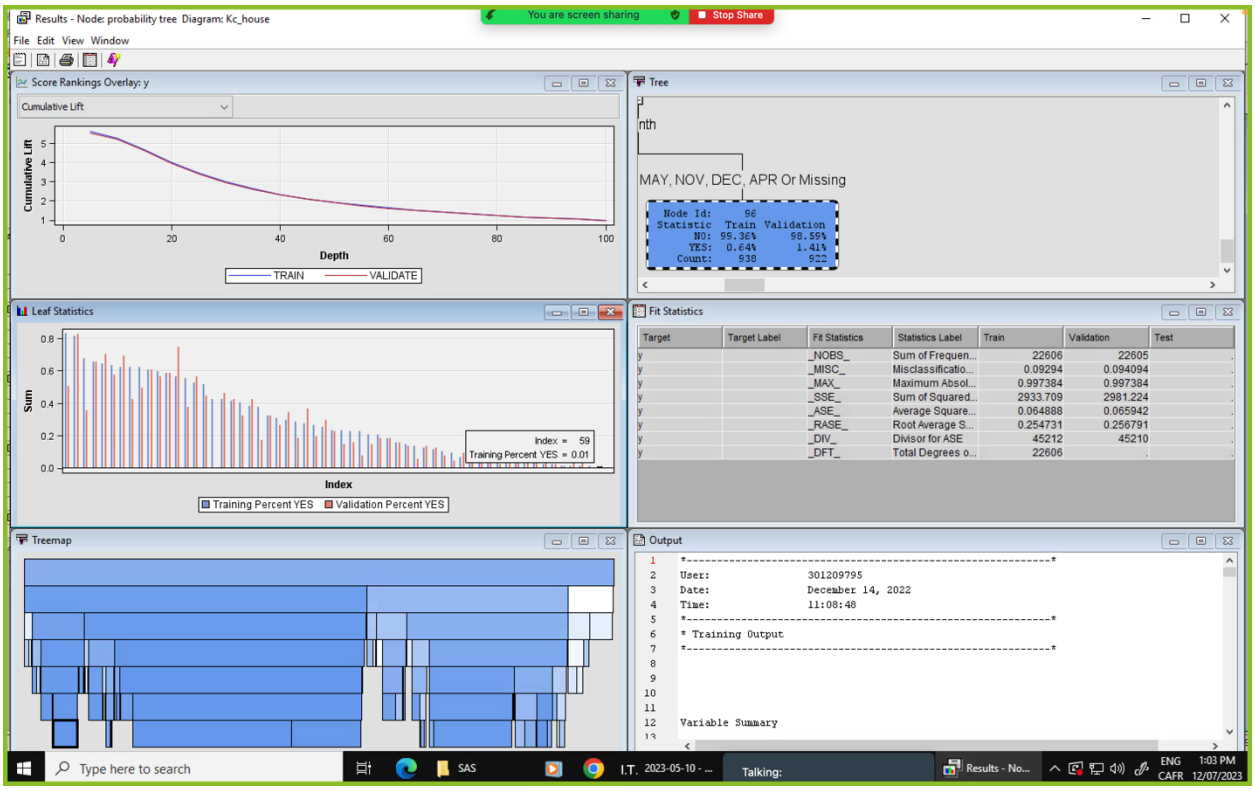
Class Decision Tree.

Our second decision tree is the Class decision tree using misclassification as an assessment measure. The subtree method used for Class decision tree is assessment. The maximum number of branches is three (3). Our second decision tree has 21 leaf statistics and an ASE of **0.069907**



Probability Decision Tree.

Our third decision tree is the Probability decision tree using average squared error as an assessment measure. The subtree method used for Probability decision tree is assessment. The maximum number of branches is three (3). Our third decision tree has 59 leaf statistics and an ASE of **0.065942**



Our key variables under the Probability decision tree are:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
duration		9	1.0000	1.0000	1.0000
poutcome		2	0.6905	0.7440	1.0775
month		10	0.6041	0.6049	1.0013
day		9	0.3903	0.3034	0.7773
contact		3	0.2100	0.1907	0.9080
REP_age	Replacement: age	3	0.1961	0.0827	0.4216
housing		4	0.1891	0.1640	0.8675

Tree Leaf Report

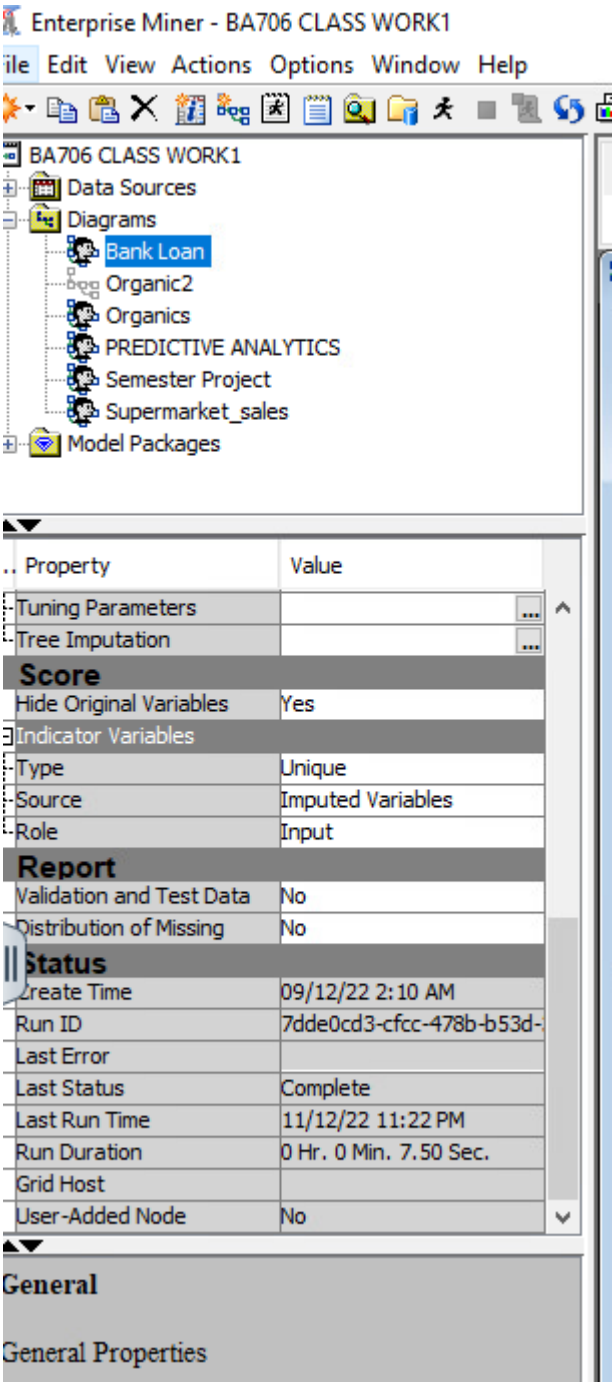
Node Id	Depth	Training Observations	Training Percent YES	Validation Observations	Validation Percent YES
106	6	6116	0.00	6108	0.00
118	6	3010	0.03	3061	0.03
107	6	2679	0.01	2756	0.01

Decision Tree Model Conclusion:

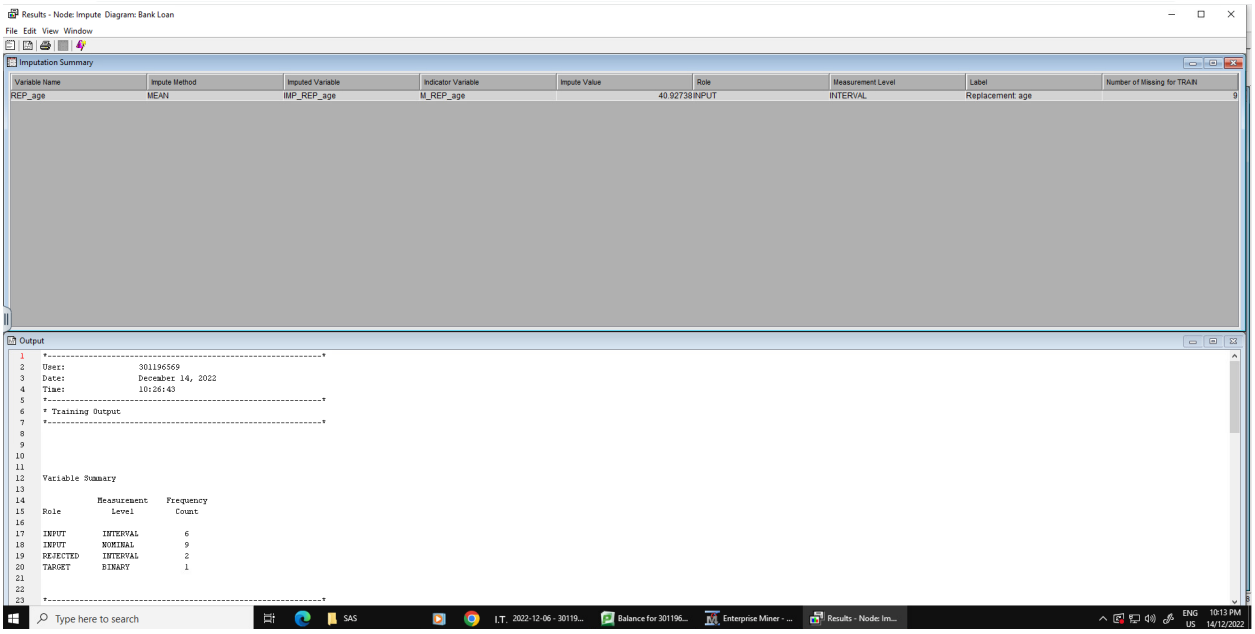
Using average squared error (ASE) as our model assessment measures, our best decision tree is the Probability decision tree with an ASE of **0.065942**. This is the best because it has the lowest ASE.

Imputation:

Under imputation, we changed the indicator variable type to unique and the role to input. See the screenshot below. This step was performed to impute the missing value(s) in preparation for regression and neural networks. The only missing value for age was replaced with the mean of the non-missing variables.

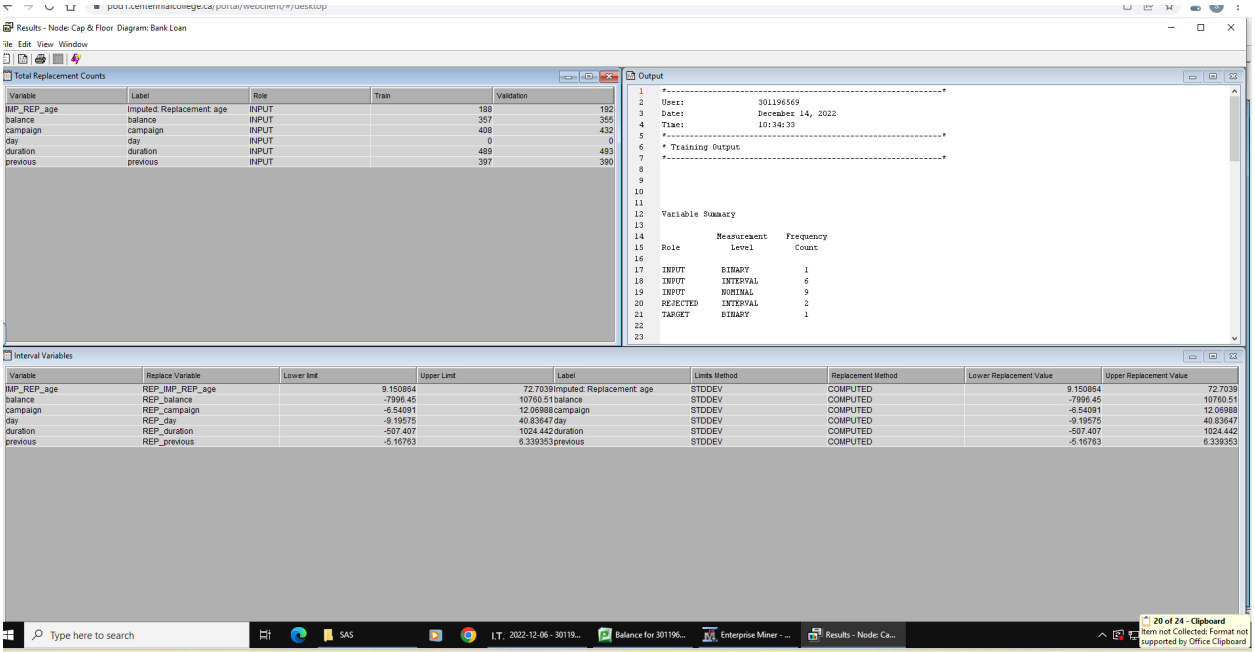


We then imputed the missing value for the age to **40.9**. See the screenshot below.



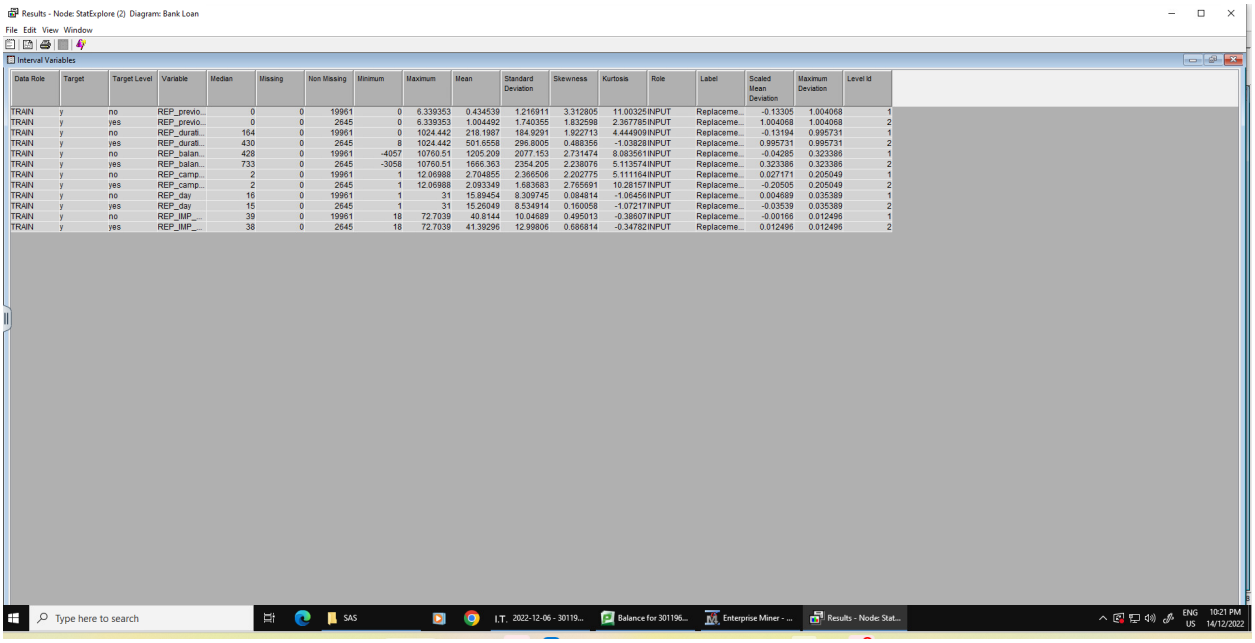
Cap and Floor:

In order to fix our skewed variables, we introduced cap and floor. However, we noticed that even after running the Cap and Floor, we still have some skewed variables. We, therefore, introduced StatExplore to ascertain the skewed variables.



StatExplore:

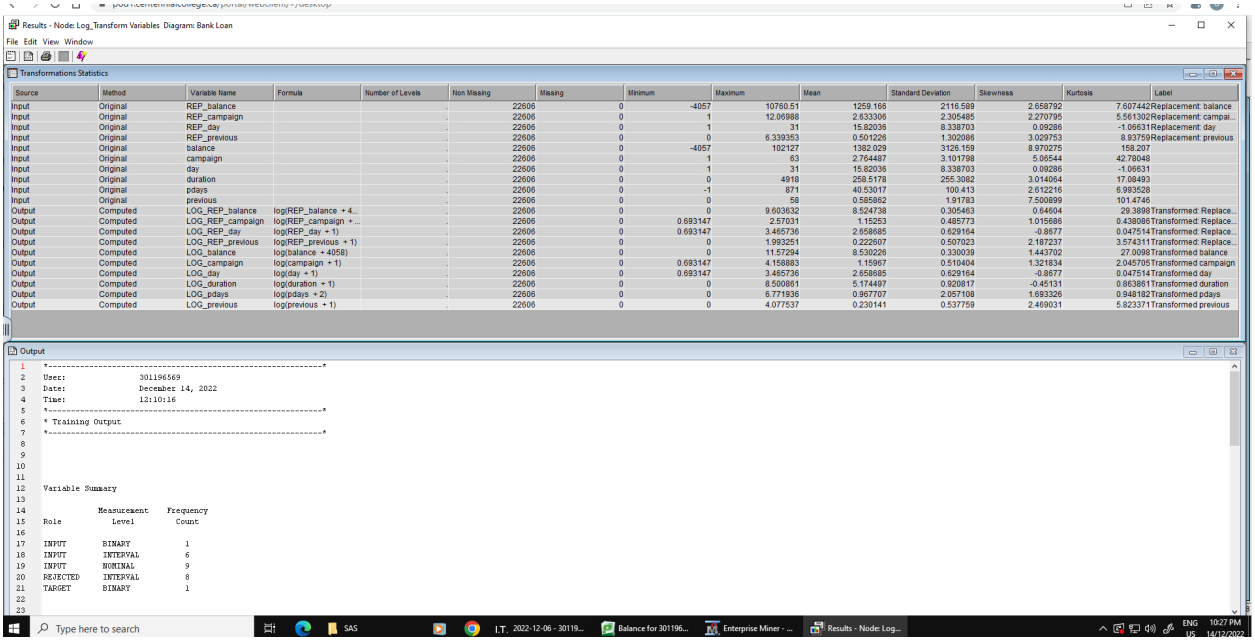
After Cap and Floor, we reintroduced StatExplore to check if the skewed variables have been fixed. However, the result as can be seen below shows that some of the variables are still skewed.

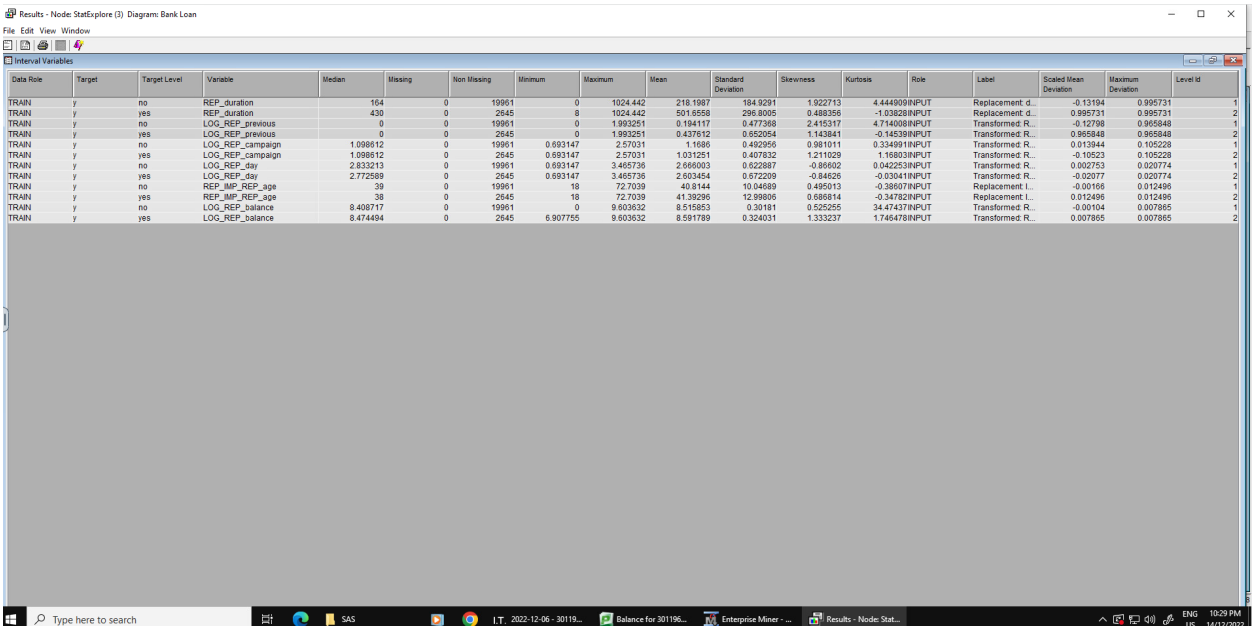


Log_Transform Variables:

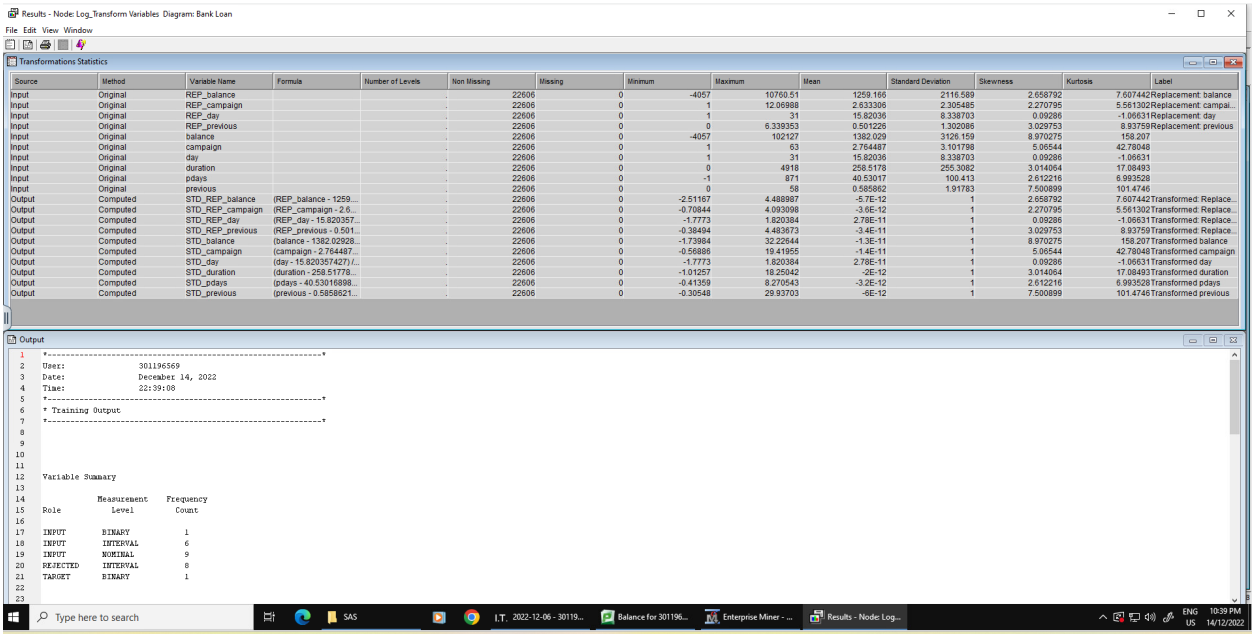
We further introduced Log_Transform Variables to fix variables that are skewed after Cap and Floor. However, we still have four of the variables that are still skewed after Log_Transform Variables as can be seen below in the screenshot. We also tried the standardized transform variables method to fix the skewed variables. However, the result got worst and so, we have to keep Log_Transform Variables.

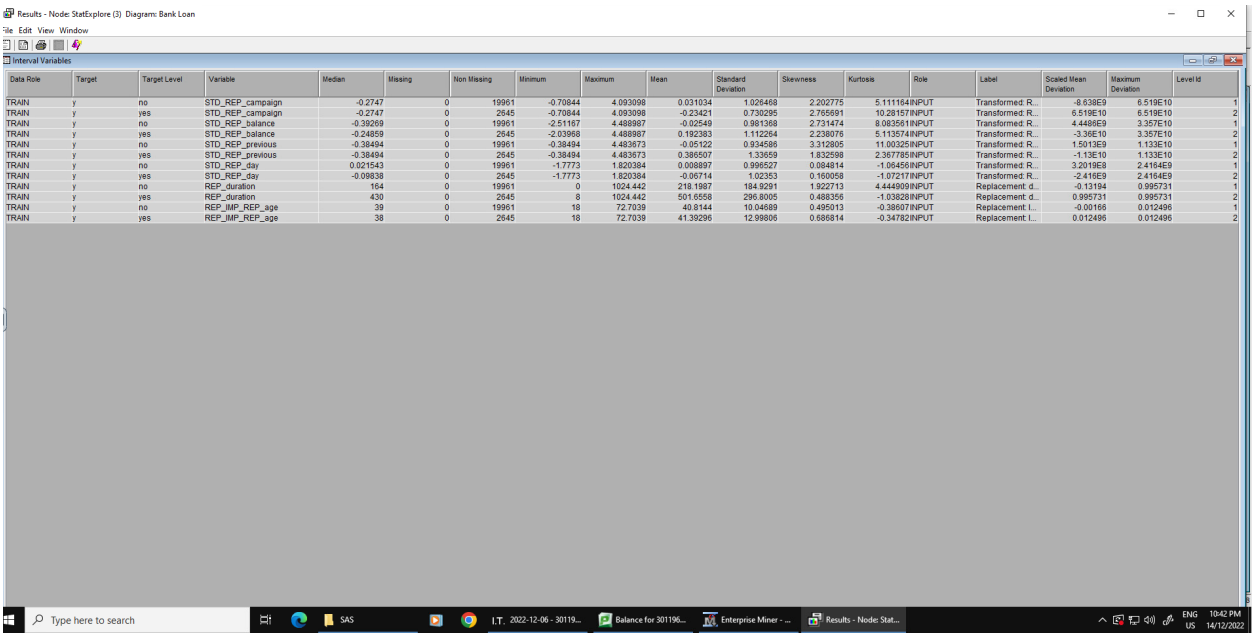
See below the results of standardized transform variables.





Results of Standardized Transform Variables.


Results - Node: Log_Transform Variables Diagram: Bank Loan
File Edit View Window
Transformations Statistics
Source Method Variable Name Formula Number of Levels Non Missing Missing Minimum Maximum Mean Standard Deviation Skewness Kurtosis Label
Input Original REP_balance 22606 0 -4057 10760.51 1259.166 2116.589 2.658792 7.607442 Replacement: balance
Input Original REP_campaign 22606 0 1 12.06988 2.633306 2.305485 2.270795 5.561302 Replacement: campai...
Input Original REP_day 22606 0 1 31 15.82036 8.338703 0.09286 -1.06631 Replacement: day
Input Original REP_previous 22606 0 0 6.339353 0.501226 1.302086 3.029753 8.93759 Replacement: previous
Input Original balance 22606 0 -4057 102127 1382.029 3126.159 8.970275 158.207
Input Original campaign 22606 0 1 63 2.764487 3.101798 5.06544 42.78048
Input Original day 22606 0 1 31 15.82036 8.338703 0.09286 -1.06631
Input Original duration 22606 0 0 4918 258.5178 255.3082 3.014064 17.08493
Input Original pdays 22606 0 -1 871 40.53017 100.413 2.612216 6.993528
Input Original previous 22606 0 0 58 0.585862 1.91783 7.500899 101.4746
Output Computed STD_REP_balance (REP_balance - 1259.... 22606 0 -2.51167 4.488987 -5.7E-12 1 2.658792 7.607442 Transformed: Replace...
Output Computed STD_REP_campaign (REP_campaign - 2.6... 22606 0 -0.70844 4.093098 -3.6E-12 1 2.270795 5.561302 Transformed: Replace...
Output Computed STD_REP_day (REP_day - 15.820357... 22606 0 -1.7773 1.820384 2.78E-11 1 0.09286 -1.06631 Transformed: Replace...
Output Computed STD_REP_previous (REP_previous - 0.501... 22606 0 -0.38494 4.483673 -3.4E-11 1 3.029753 8.93759 Transformed: Replace...
Output Computed STD_balance (balance - 1382.02928... 22606 0 -1.73984 32.22644 -1.3E-11 1 8.970275 158.207 Transformed: balance
Output Computed STD_campaign (campaign - 2.764487... 22606 0 -0.56886 19.41955 -1.4E-11 1 5.06544 42.78048 Transformed: campaign
Output Computed STD_day (day - 15.820357427) /... 22606 0 -1.7773 1.820384 2.78E-11 1 0.09286 -1.06631 Transformed: day
Output Computed STD_duration (duration - 258.51778... 22606 0 -1.01257 18.25042 -2E-12 1 3.014064 17.08493 Transformed: duration
Output Computed STD_pdays (pdays - 40.53016898... 22606 0 -0.41359 8.270543 -3.2E-12 1 2.612216 6.993528 Transformed: pdays
Output Computed STD_previous (previous - 0.5858621... 22606 0 -0.30548 29.93703 -6E-12 1 7.500899 101.4746 Transformed: previous
Output
1 *--*
2 User: 301196569
3 Date: December 14, 2022
4 Time: 22:39:08
5 *--*
6 * Training Output
7 *--*
12 Variable Summary
14 Measurement Frequency
15 Role Level Count
17 INPUT BINARY 1
18 INPUT INTERVAL 6
19 INPUT NOMINAL 9
20 REJECTED INTERVAL 8
21 TARGET BINARY 1
Type here to search
SAS
I.T. 2022-12-06 - 30119...
Balance for 301196...
Enterprise Miner - ...
Results - Node: Log...
ENG US
10:39 PM 14/12/2022

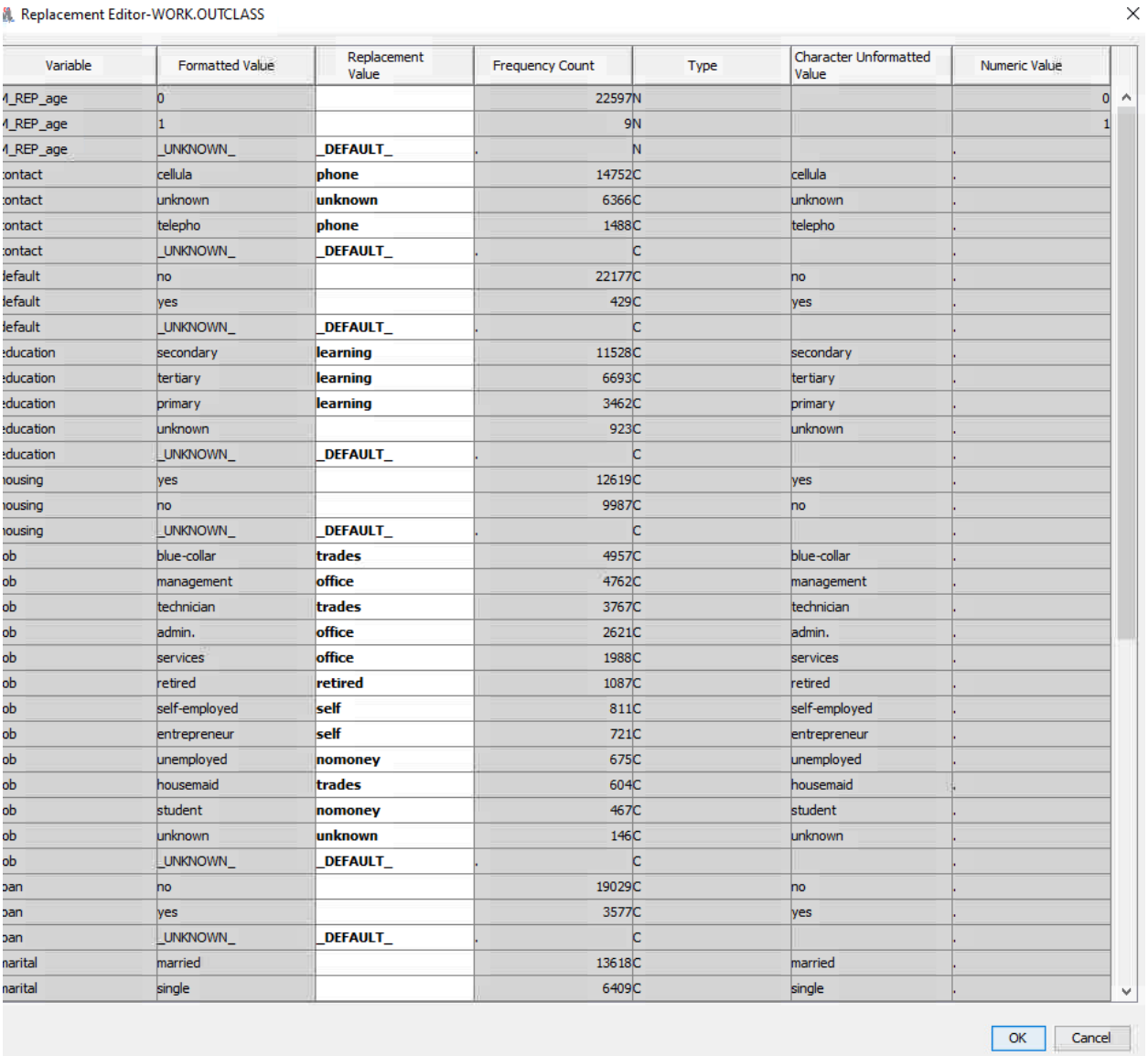


Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	y	no	STD_REP_campaign	-0.2747	0	19961	-0.70844	4.093098	0.031034	1.026468	2.202775	5.111164	INPUT	Transformed: R...	-8.638E9	6.519E10	1
TRAIN	y	yes	STD_REP_campaign	-0.2747	0	2645	-0.70844	4.093098	-0.23421	0.730295	2.765691	10.28157	INPUT	Transformed: R...	6.519E10	6.519E10	2
TRAIN	y	no	STD_REP_balance	-0.39269	0	19961	-2.51167	4.488987	-0.02549	0.981368	2.731474	8.083561	INPUT	Transformed: R...	4.4486E9	3.357E10	1
TRAIN	y	yes	STD_REP_balance	-0.24859	0	2645	-2.03968	4.488987	0.192383	1.112264	2.238076	5.113574	INPUT	Transformed: R...	-3.36E10	3.357E10	2
TRAIN	y	no	STD_REP_previous	-0.38494	0	19961	-0.38494	4.483673	-0.05122	0.934586	3.312805	11.00325	INPUT	Transformed: R...	1.5013E9	1.133E10	1
TRAIN	y	yes	STD_REP_previous	-0.38494	0	2645	-0.38494	4.483673	0.386507	1.33659	1.832598	2.367785	INPUT	Transformed: R...	-1.13E10	1.133E10	2
TRAIN	y	no	STD_REP_day	0.021543	0	19961	-1.7773	1.820384	0.008897	0.996527	0.084814	-1.06456	INPUT	Transformed: R...	3.2019E8	2.4164E9	1
TRAIN	y	yes	STD_REP_day	-0.09838	0	2645	-1.7773	1.820384	-0.06714	1.02353	0.160058	-1.07217	INPUT	Transformed: R...	-2.415E9	2.4164E9	2
TRAIN	y	no	REP_duration	164	0	19961	0	1024.442	218.1987	184.8291	1.922713	4.444909	INPUT	Replacement: d...	-0.13194	0.995731	1
TRAIN	y	yes	REP_duration	430	0	2645	8	1024.442	501.6558	296.8005	0.488356	-1.03828	INPUT	Replacement: d...	0.995731	0.995731	2
TRAIN	y	no	REP_IMP_REP_age	39	0	19961	18	72.7039	40.8144	10.04689	0.495013	-0.38607	INPUT	Replacement: I...	-0.00166	0.012496	1
TRAIN	y	yes	REP_IMP_REP_age	38	0	2645	18	72.7039	41.39296	12.99806	0.686814	-0.34782	INPUT	Replacement: I...	0.012496	0.012496	2

Dummies Recode:

We introduced dummies recode in our model to avoid overfitting. However, the results of our models' assessment measures became worst with dummies recode (replacement values). The results below show the effect of dummies recodes on the models.

We will retain the dummies recode node, however, we will not replace the dummy variables because our assessment measures will be optimized without dummies recode.

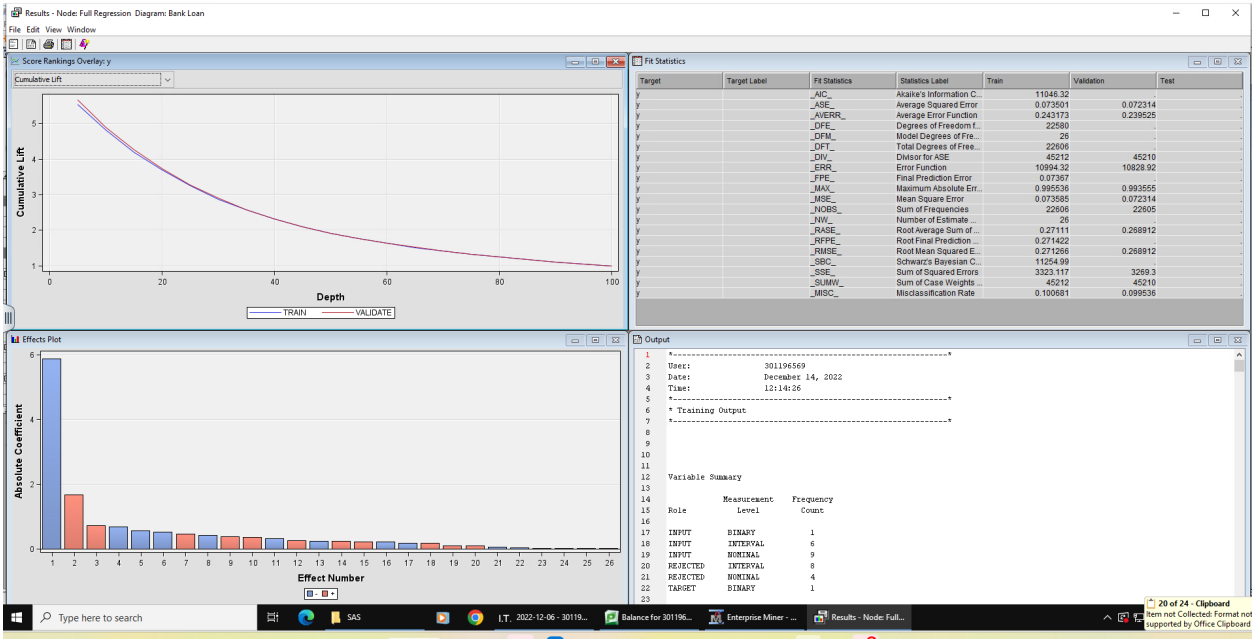


Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
M_REP_age	0		22597	N		0
M_REP_age	1		9	N		1
M_REP_age	_UNKNOWN_	**_DEFAULT_**	.	N		.
contact	cellula	**phone**	14752	C	cellula	.
contact	unknown	**unknown**	6366	C	unknown	.
contact	telepho	**phone**	1488	C	telepho	.
contact	_UNKNOWN_	**_DEFAULT_**	.	C		.
default	no		22177	C	no	.
default	yes		429	C	yes	.
default	_UNKNOWN_	**_DEFAULT_**	.	C		.
education	secondary	**learning**	11528	C	secondary	.
education	tertiary	**learning**	6693	C	tertiary	.
education	primary	**learning**	3462	C	primary	.
education	unknown		923	C	unknown	.
education	_UNKNOWN_	**_DEFAULT_**	.	C		.
housing	yes		12619	C	yes	.
housing	no		9987	C	no	.
housing	_UNKNOWN_	**_DEFAULT_**	.	C		.
job	blue-collar	**trades**	4957	C	blue-collar	.
job	management	**office**	4762	C	management	.
job	technician	**trades**	3767	C	technician	.
job	admin.	**office**	2621	C	admin.	.
job	services	**office**	1988	C	services	.
job	retired	**retired**	1087	C	retired	.
job	self-employed	**self**	811	C	self-employed	.
job	entrepreneur	**self**	721	C	entrepreneur	.
job	unemployed	**nomoney**	675	C	unemployed	.
job	housemaid	**trades**	604	C	housemaid	.
job	student	**nomoney**	467	C	student	.
job	unknown	**unknown**	146	C	unknown	.
job	_UNKNOWN_	**_DEFAULT_**	.	C		.
loan	no		19029	C	no	.
loan	yes		3577	C	yes	.
loan	_UNKNOWN_	**_DEFAULT_**	.	C		.
marital	married		13618	C	married	.
marital	single		6409	C	single	.

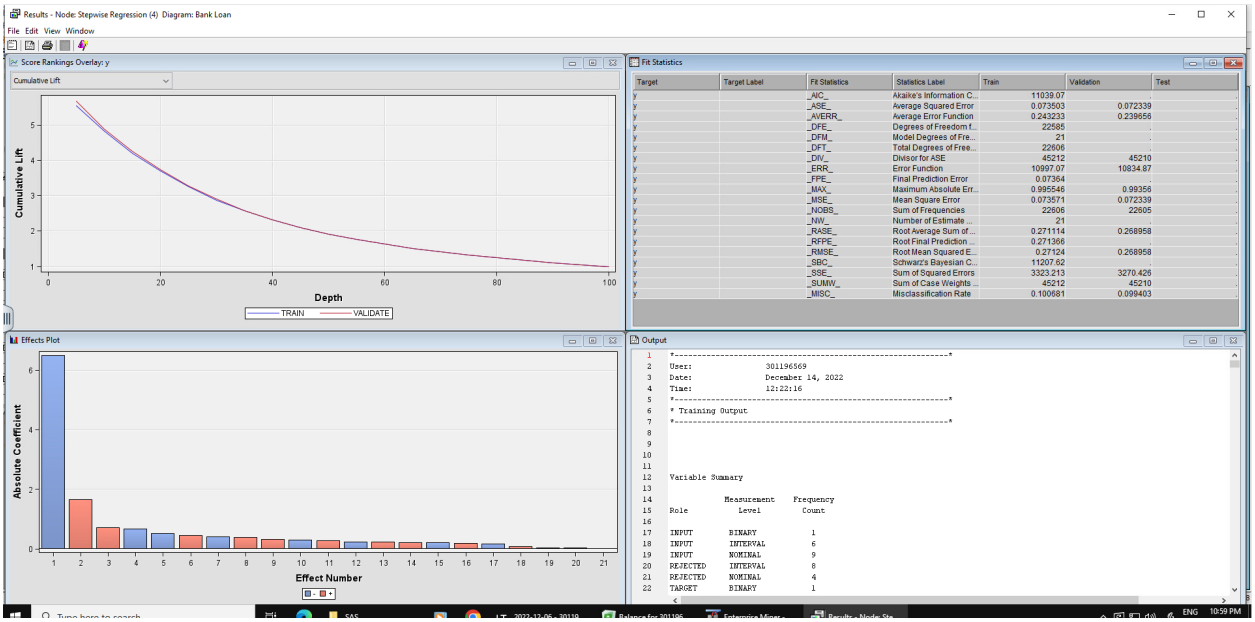
n	yes		3577	C	yes	.
n	_UNKNOWN_	_DEFAULT_	.	C		.
rital	married		13618	C	married	.
rital	single		6409	C	single	.
rital	divorced		2579	C	divorced	.
rital	_UNKNOWN_	_DEFAULT_	.	C		.
nth	may	q2	6918	C	may	.
nth	jul	q3	3452	C	jul	.
nth	aug	q3	3171	C	aug	.
nth	jun	q2	2591	C	jun	.
nth	nov	q4	1928	C	nov	.
nth	apr	q2	1460	C	apr	.
nth	feb	q1	1357	C	feb	.
nth	jan	q1	718	C	jan	.
nth	oct	q4	372	C	oct	.
nth	sep	q3	294	C	sep	.
nth	mar	q1	237	C	mar	.
nth	dec	q4	108	C	dec	.
nth	_UNKNOWN_	_DEFAULT_	.	C		.
utcome	unknown		18440	C	unknown	.
utcome	failure		2463	C	failure	.
utcome	other		948	C	other	.
utcome	success		755	C	success	.
utcome	_UNKNOWN_	_DEFAULT_	.	C		.
	no		19961	C	no	.
	yes		2645	C	yes	.
	UNKNOWN	_DEFAULT_	.	C		.

OK Cancel

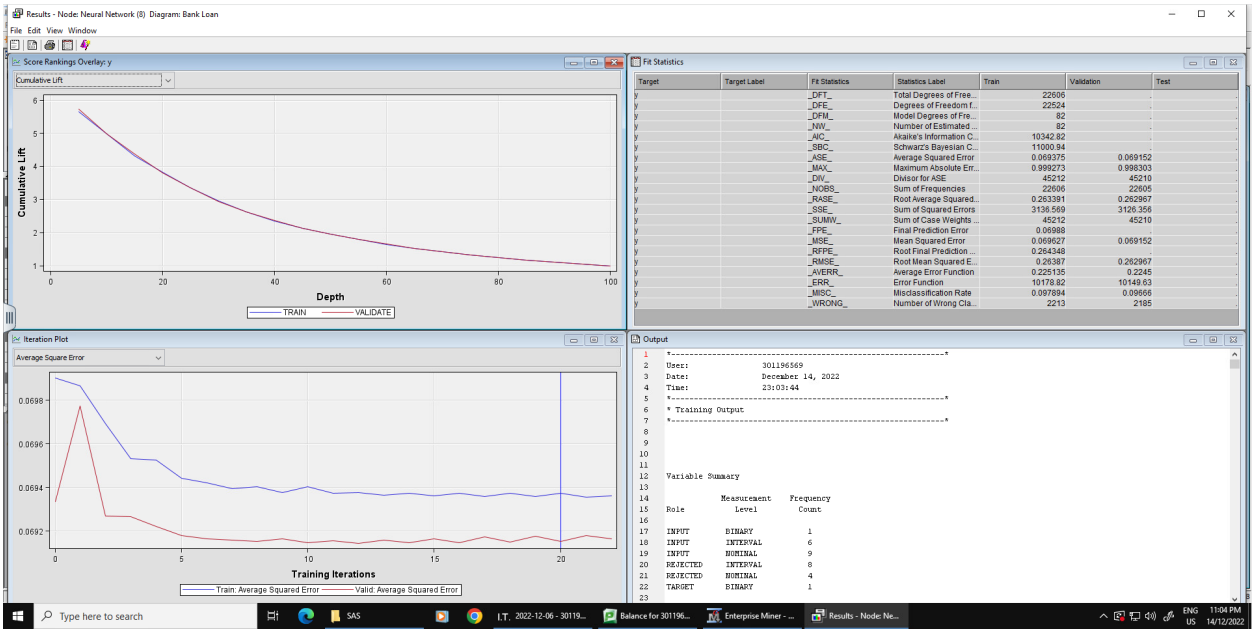
Full Regression with Dummies Recode (with replacement values). The ASE is 0.072314.



Stepwise Regression with Dummies Recode (with replacement values). The ASE is 0.072329.

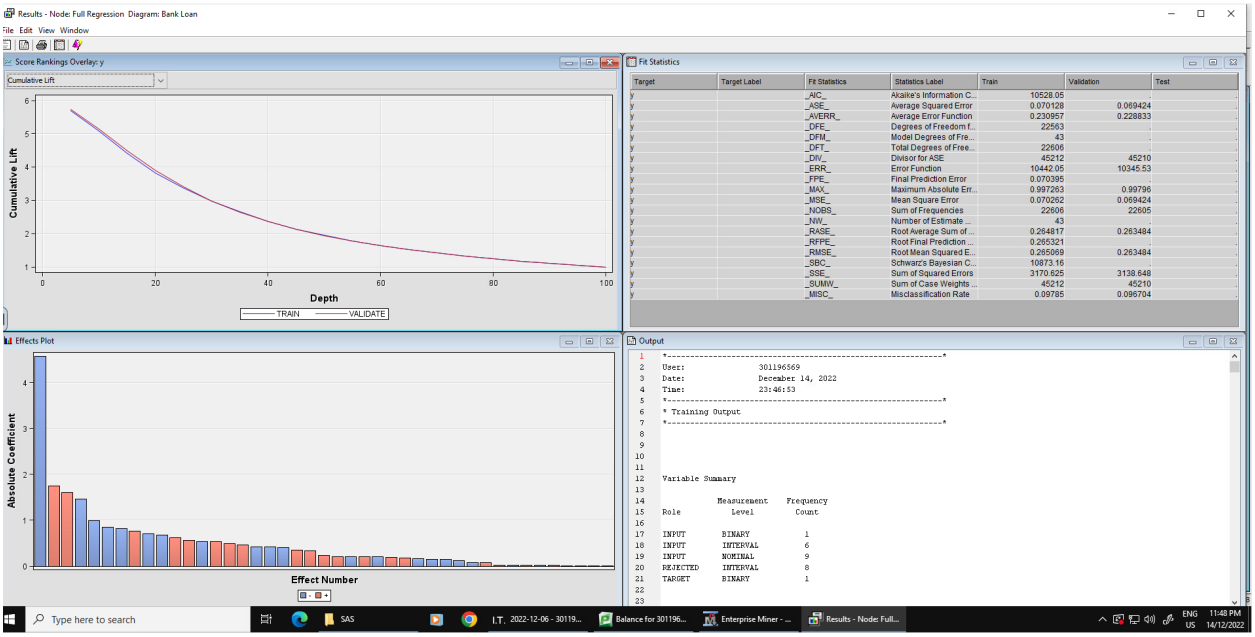


Neural Network with Dummies Recode (replacement values). The ASE is 0.069152.

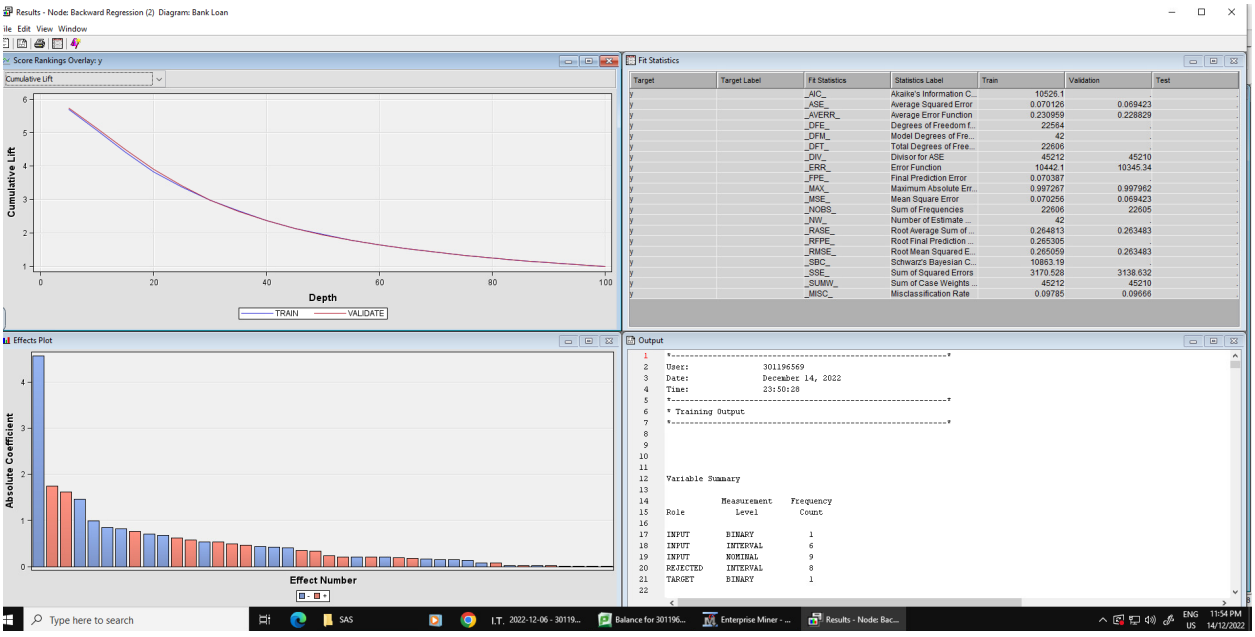


Regression without Dummies Recode

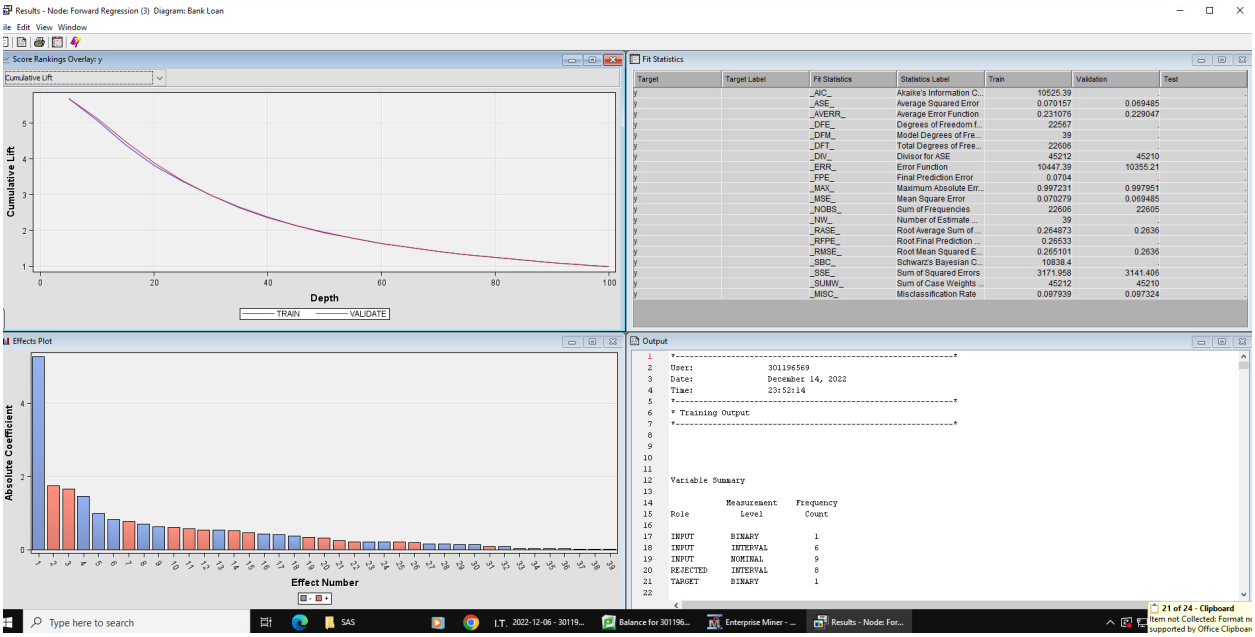
Full Regression: In our full regression, we did not change anything in the property value, including the selection model. Our ASE for full regression is **0.069424**. The result of the full without dummies recode is better than with dummies recode.



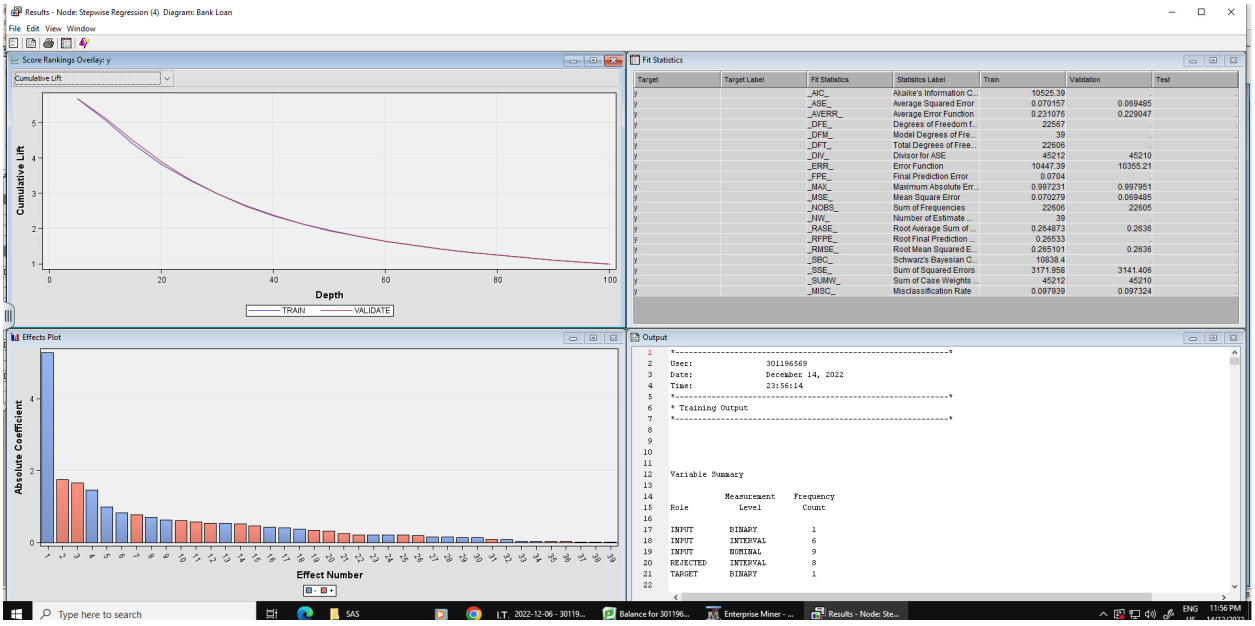
Backward Regression: In our regression, we made some changes. We changed the selection model to Backward and the selection criterion to validation error. Our ASE for backward regression is **0.069423**.



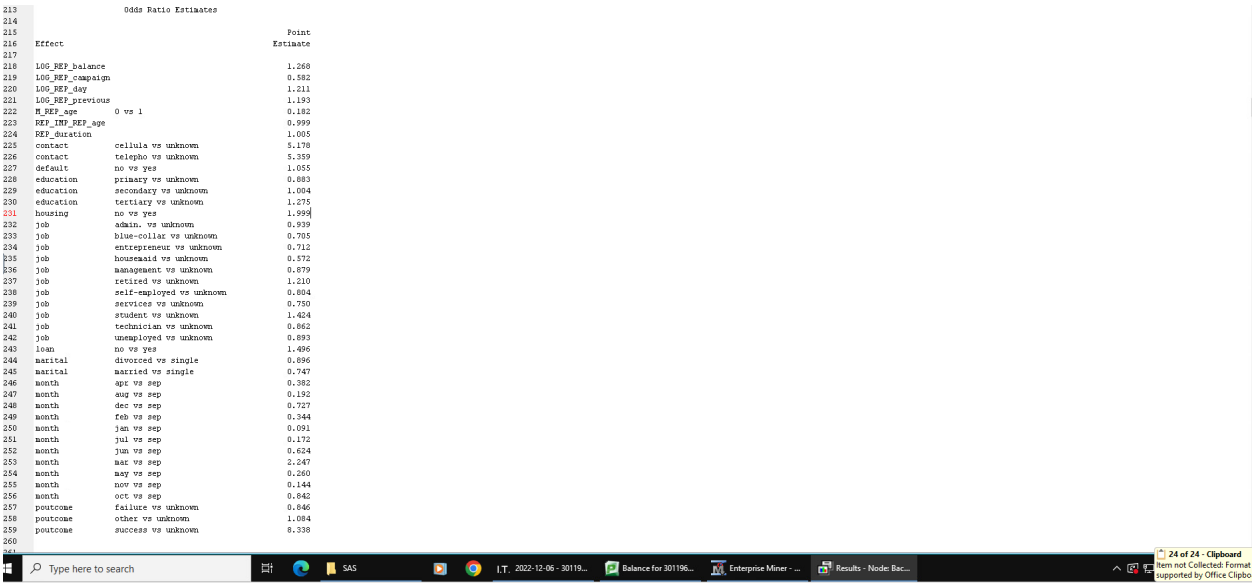
Forward Regression: For forward regression, we changed our selection model to forward and the selection criterion to validation error. Our ASE for forward regression is **0.069485**.



Stepwise Regression: In stepwise regression, we changed the selection model to stepwise while the selection criterion was validation error. The ASE for stepwise regression is **0.069485**.



Odds Ratio and interpretation.



Odds Ratio Estimates

Effect		Point Estimate
LOG_REP_balance		1.268
LOG_REP_campaign		0.582
LOG_REP_day		1.211
LOG_REP_previous		1.193
M_REP_age	0 vs 1	0.182
REP_IMP_REP_age		0.999
REP_duration		1.005
contact	cellula vs unknown	5.178
contact	telepho vs unknown	5.359
default	no vs yes	1.055
education	primary vs unknown	0.883
education	secondary vs unknown	1.004
education	tertiary vs unknown	1.275
housing	no vs yes	1.999
job	admin. vs unknown	0.939
job	blue-collar vs unknown	0.705
job	entrepreneur vs unknown	0.712
job	housemaid vs unknown	0.572
job	management vs unknown	0.879
job	retired vs unknown	1.210
job	self-employed vs unknown	0.804
job	services vs unknown	0.750
job	student vs unknown	1.424
job	technician vs unknown	0.862
job	unemployed vs unknown	0.893
loan	no vs yes	1.496
marital	divorced vs single	0.896
marital	married vs single	0.747
month	apr vs sep	0.382
month	aug vs sep	0.192
month	dec vs sep	0.727
month	feb vs sep	0.344
month	jan vs sep	0.091
month	jul vs sep	0.172
month	jun vs sep	0.624
month	mar vs sep	2.247
month	may vs sep	0.260
month	nov vs sep	0.144
month	oct vs sep	0.842
poutcome	failure vs unknown	0.846
poutcome	other vs unknown	1.084
poutcome	success vs unknown	8.338

Odds Ratio

Variables	**Ratios**	**% of Change**
LOG_REP_balance	**1.268**	**27%**
LOG_REP_camp	**0.582**	**-42%**
LOG_REP_REP_day	**1.211**	**21%**
LOG_REP_previous	**1.193**	**19%**
M_REP_age 0 v 1	**0.182**	**-82%**
REP_IMP_REP_age	**0.999**	**0%**
REP_duration	**1.005**	**0%**
contact cellula vs unknown	**5.178**	**418%**
contact telephone vs unknown	**5.359**	**436%**
default no vs yes	**1.055**	**5%**
education primary vs unknown	**0.883**	**-12%**
education secondary vs unknown	**1.004**	**0%**
education tertiary vs unknown	**1.275**	**28%**
housing no vs yes	**1.999**	**100%**
job admin. Vs unknown	**0.939**	**-6%**
job blue-collar vs unknown	**0.705**	**-30%**
job entrepreneur vs unknown	**0.712**	**-29%**
job housemaid vs unknown	**0.572**	**-43%**
job management vs unknown	**0.879**	**-12%**
job retired vs unknown	**1.21**	**21%**
job self-employed vs unknown	**0.804**	**-20%**
job services vs unknown	**0.75**	**-25%**

job student vs unknown	**1.424**	**42%**
job technician vs unknown	**0.862**	**-14%**
job unemployed vs unknown	**0.893**	**-11%**
loan no vs yes	**1.496**	**50%**
divorced vs single	**0.896**	**-10%**
married vs single	**0.747**	**-25%**
month apr vs sep	**0.382**	**-62%**
month aug vs sep	**0.192**	**-81%**
month dec vs sep	**0.727**	**-27%**
month feb vs sep	**0.344**	**-66%**
month jan vs sep	**0.091**	**-91%**
month jul vs sep	**0.172**	**-83%**
month jun vs sep	**0.624**	**-38%**
month mar vs sep	**0.624**	**-38%**
month may vs sep	**0.26**	**-74%**
month nov vs sep	**0.144**	**-86%**
month oct vs sep	**0.842**	**-16%**
poutcome failure vs unknown	**0.846**	**-15%**
poutcome other vs unknown	**1.084**	**8%**
poutcome success vs unknown	**8.338**	**734%**

From the odds ratio, we have the following variables to interpret.

- LOG_REP_balance: This means that for every log(euro) increase in the customers' log_balances, they are 27% more likely to accept the loan campaign of the bank. Alternatively, for every increase in euro of 2.74, the probability of accepting the loan campaign goes up 27%.
- LOG_REP_camp: For every log(campaign), the number of customers contacted is likely to reduce by 42%. Alternatively, for every decrease in campaign of 2.74, the probability of rejecting the loan campaign goes up 42%.
- LOG_REP_REP_day: Depending on the log(day) of the month the customers are contacted, there is a 21% chance that the customer will accept the loan campaign. Alternatively, for every increase in day of 2.74, the probability of accepting the loan campaign goes up 21%.
- LOG_REP_previous: For every log(previous) marketing of the bank loan to the customers there is a 19% chance of influencing the customers' decision to accept the bank loan. Alternatively, for every increase in previous marketing campaign of 2.74, the probability of accepting the loan campaign goes up 19%.
- M_REP_age 0 v 1: For every addition in the known age, there is an 82% chance of not being likely to accept the loan campaign.

- REP_IMP_REP_age: For every increase in the imputed age, there are unlikely that the loan campaign will influence the customers' decision to accept a bank loan.
- REP_duration: For every increase in the number of seconds spent explaining bank loans to customers, there is a 0.5% of chance of the customers accepting the loan campaign.
- Contact - cellula vs unknown: Customers contacted via cellula are 5 times more likely to accept the bank loan campaign than those contacted via unknown means of communication.
- Contact - telephone vs unknown: Customers contacted through telephone are also 5 times more likely to accept loan campaigns than the unknown.
- default - no vs yes: For every customer contacted for the bank loan, there is a 5% chance that they have no credit default.
- education - primary vs unknown: For every customer, there is a 12% chance that they do not have primary education than the unknown.
- education - secondary vs unknown: For every customer, there is a 0.4% chance that they do have secondary education than the unknown.
- education - tertiary vs unknown: Every customer contacted is 28% likely to have tertiary education.
- housing - no vs yes: Every customer is 100% likely to not have a housing loan.
- job - admin. Vs unknown: Every customer contacted is 6% more unlikely to be in an admin job than the unknown.
- Job - blue-collar vs unknown: 30% of the customers are more unlikely to have blue-collar than the unknown.
- job - entrepreneur vs unknown: There is a 29% chance that the customers are unlikely to be entrepreneurs.
- job - housemaid vs unknown: There is a 43% chance that the customers are more unlikely to be housemaids than the unknown.
- job - management vs unknown: The customers are 12% unlikely to be in a management cadre in their jobs.
- job - retired vs unknown: Retired customers are 21% more likely to accept loan campaigns.
- job - self-employed vs unknown: Self-employed customers are 20% more unlikely to accept loan campaigns than the unknown.
- job - services vs unknown: Customers in services jobs are 25% more likely not to accept loan campaigns than the unknown.
- job - student vs unknown: Students are 45% most likely to accept loan campaigns than the unknown.
- job - technician vs unknown: Customers who are technicians are 14% more likely not to accept loan campaigns than the unknown.
- job - unemployed vs unknown: Customers who are unemployed are 11% unlikely to accept loan campaigns.

- loan - no vs yes: Customers who have existing personal loans are 50% more likely to say no loan campaigns than yes.
- divorced - vs single: Divorced customers are 10% more unlikely to accept loan campaigns than single.
- married - vs single: Bank customers who are single are 25% more likely to accept loan campaigns than married.
- month - apr vs sep: Customers contacted in April are 62% more unlikely to accept loan campaigns than in September.
- month - aug vs sep: Customers contacted in August are 81% more unlikely to accept loan campaigns than in September.
- month - dec vs sep: Customers contacted in December are 27% more unlikely to accept loan campaigns than in September.
- month - feb vs sep: Customers contacted in February are 66% more unlikely to accept loan campaigns than in September.
- month - jan vs sep: Customers contacted in January are 91% more unlikely to accept loan campaigns than in September.
- month - jul vs sep: Customers contacted in July are 83% more unlikely to accept loan campaigns than in September.
- month - jun vs sep: Customers contacted in June are 38% more unlikely to accept loan campaigns than in September.
- month mar - vs sep: Customers contacted in March are 38% more unlikely to accept loan campaigns than in September.
- month - may vs sep: Customers contacted in May are 74% more unlikely to accept loan campaigns than in September.
- month - nov vs sep: Customers contacted in November are 86% more unlikely to accept loan campaigns than in September.
- month - oct vs sep: Customers contacted in October are 16% more unlikely to accept loan campaigns than in September.
- poutcome - failure vs unknown: There is a likelihood of 15% failure of the outcome of the previous campaign than the unknown.

- poutcome - other vs unknown: There is an 8% chance of other outcomes of the previous loan campaigns than the unknown.
- poutcome - success vs unknown: The outcome of the previous loan campaign is 8 times more likely to be successful than the unknown.

Summary of our regression models

S/N	Regression	ASE
1	Full regression	0.069424
2	Backward regression	0.069423
3	Forward and stepwise regression	0.069485

The most important variables in the regression are those variables with the least Ch-square of 0.0001 as below.

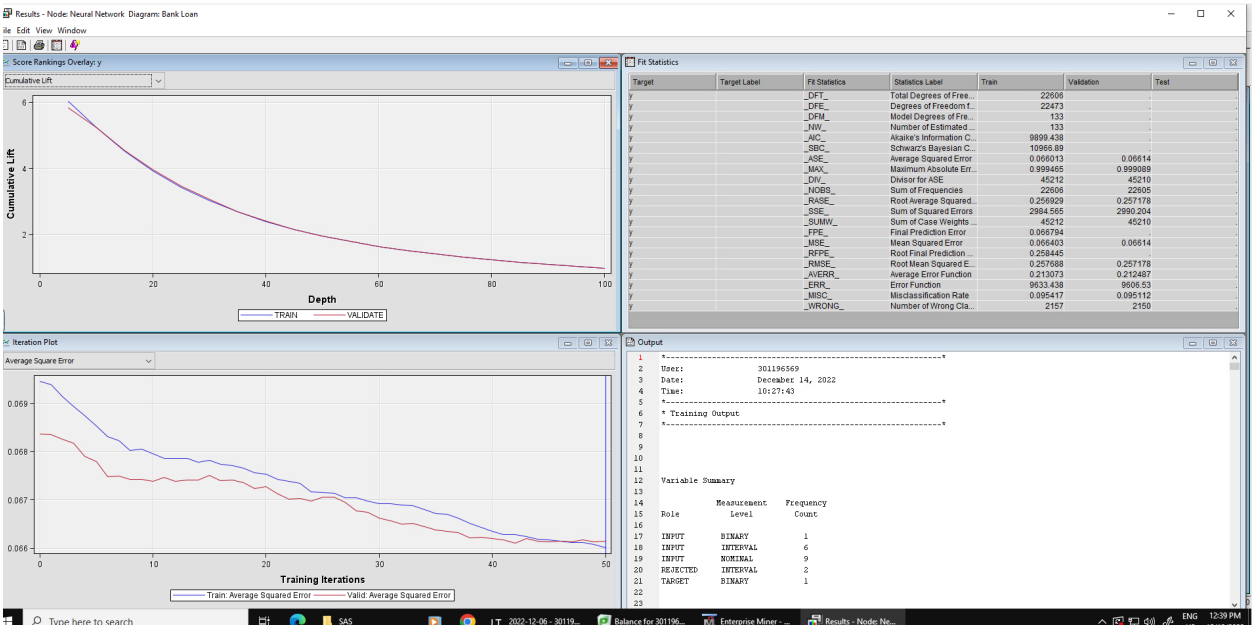
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LOG_REP_balance	1	8.3401	0.0039
LOG_REP_campaign	1	67.2839	<.0001
LOG_REP_day	1	17.1160	<.0001
LOG_REP_previous	1	2.6306	0.1048
M_REP_age	1	2.7811	0.0954
REP_IMP_REP_age	1	0.1083	0.7421
REP_duration	1	2572.8082	<.0001
contact	2	254.3552	<.0001
education	3	14.7029	0.0021
housing	1	122.3450	<.0001
job	11	41.7323	<.0001
loan	1	22.6922	<.0001
marital	2	21.2814	<.0001
month	11	534.6409	<.0001
poutcome	3	437.2939	<.0001

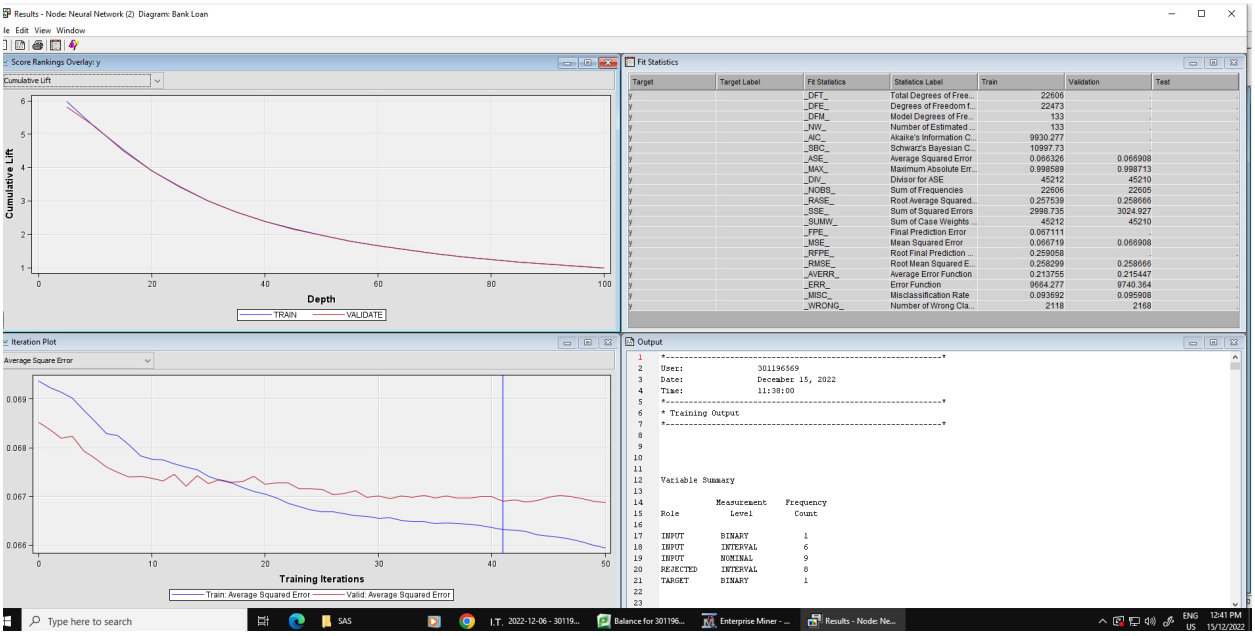
Conclusion: From the above, our best regression model is backward regression with an ASE of **0.069423**.

Neural Network:
We built neural network regression models using profit and loss as the selection criterion. The following are the neural network models and the results.

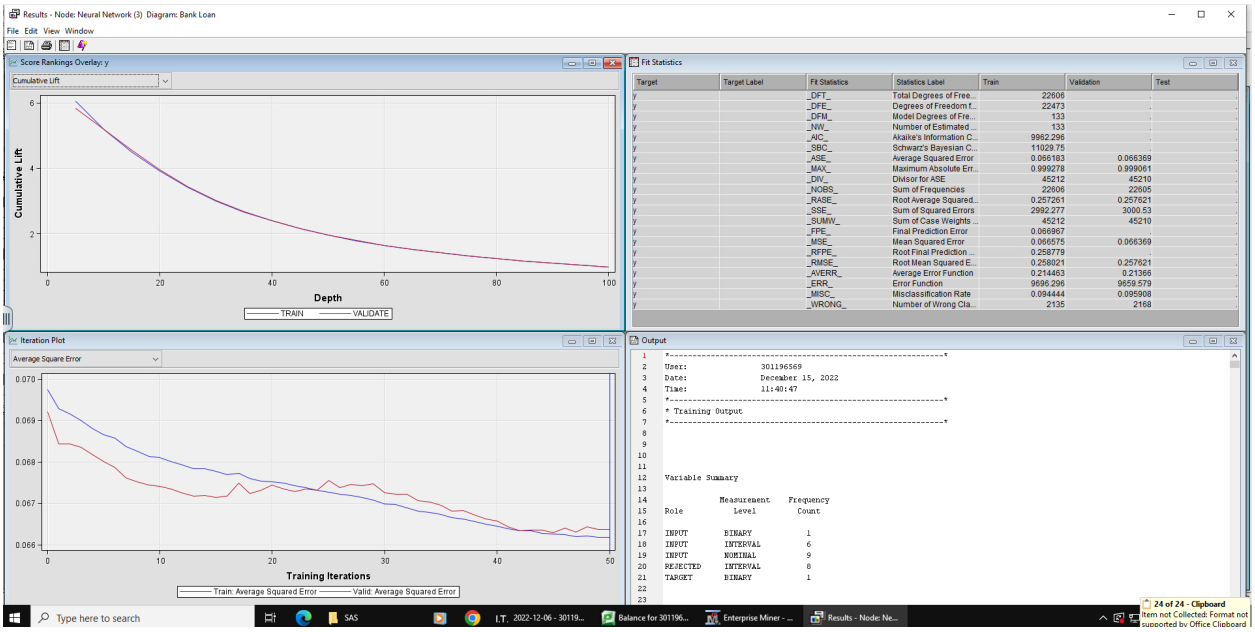
1. Neural network: The first network is connected to the impute without changing anything in the model optimization. The ASE for the first neural network is **0.06614**.



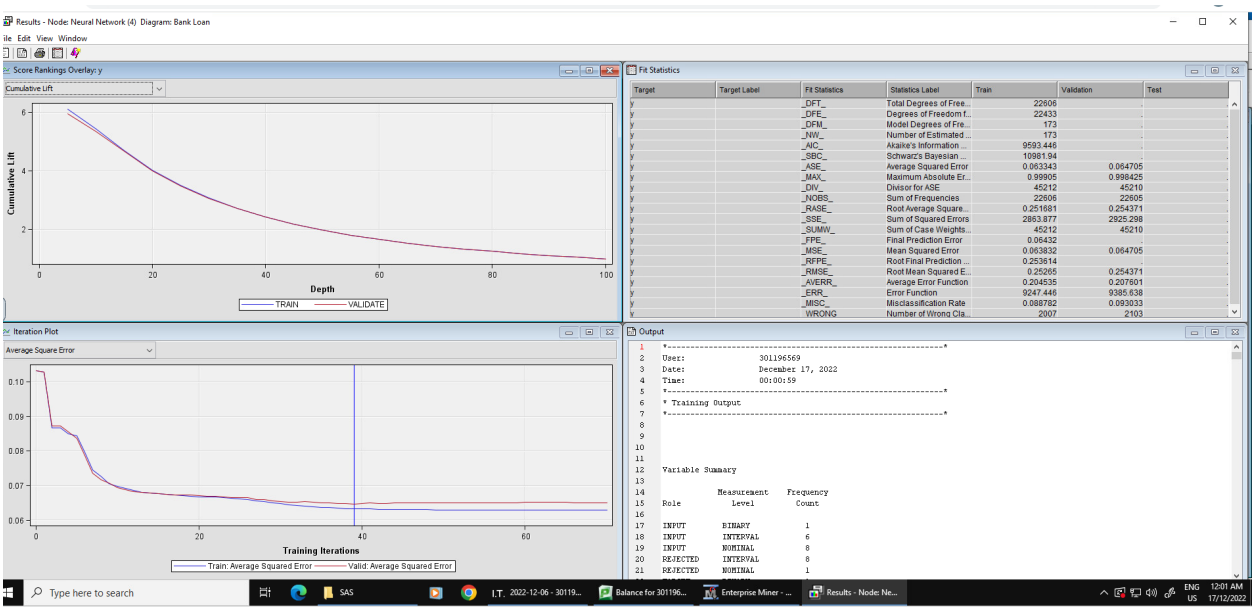
2. Neural network (2): The second neural network is connected Cap and Floor. The cap and floor are intended to treat variables that are skewed. Hence, the difference between the ASE of the first neural network and the second. The ASE of the second neural network is **0.066908**.



3. Neural network (3): The third to connected to the Log_Tranform Variables which is also intended to treat the skewed variables after cap and floor. However, the ASE of the third neural network is 0.066369.



4. Neural network (4): The neural network (4) is connected to the backward regression because it is our best regression model. However, we made some changes to optimize the model by changing the maximum iteration to 100 and preliminary training to no. We also changed the hidden unit number of the network from 3 to 4 and this is the optimal hidden unit number. The ASE for the 4th neural work is 0.064705.



Event Classification:

Our event classification refers to the classification of the impact of a bank loan campaign. The classification is below.

- **True negative**: This represents customers who were not targeted (contacted) by the bank for the loan campaign and did not subscribe to the bank loan. They represent a total of 19, 265 customers, or 85.2245% of the customers.
- **False negative**: This represents bank customers who were targeted (contacted) for the bank loan but did not subscribe to the loan. They are 1, 407 representing 6.2243% of the customers.
- **False positive**: These are the customers not targeted but subscribed to the loan. They are 696 representing 3.079% of the customers.
- **True positive**: These are bank customers targeted and who subscribed to the bank loan. They are 1, 237 representing 5.4722% of the customers.

See the event classification table below.

Classification Table

Data Role=TRAIN Target Variable=y Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	93.5734	96.5783	19278	85.2782
YES	NO	6.4266	50.0567	1324	5.8569
NO	YES	34.0818	3.4217	683	3.0213
YES	YES	65.9182	49.9433	1321	5.8436

Data Role=VALIDATE Target Variable=y Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	93.1937	96.5132	19265	85.2245
YES	NO	6.8063	53.2148	1407	6.2243
NO	YES	36.0062	3.4868	696	3.0790
YES	YES	63.9938	46.7852	1237	5.4722

Event Classification Table

Data Role=TRAIN Target=y Target Label=' '

False Negative	True Negative	False Positive	True Positive
1324	19278	683	1321

Summary of neural network

S/N	Neural Network	ASE
1	Neural Network (connected to impute).	0.06614
2	Neural network (2) connected to cap & floor.	0.066908
3	Neural network (3) connected to log_transform variable.	0.066369
4	Neural network (4) connected backward regression.	0.064705

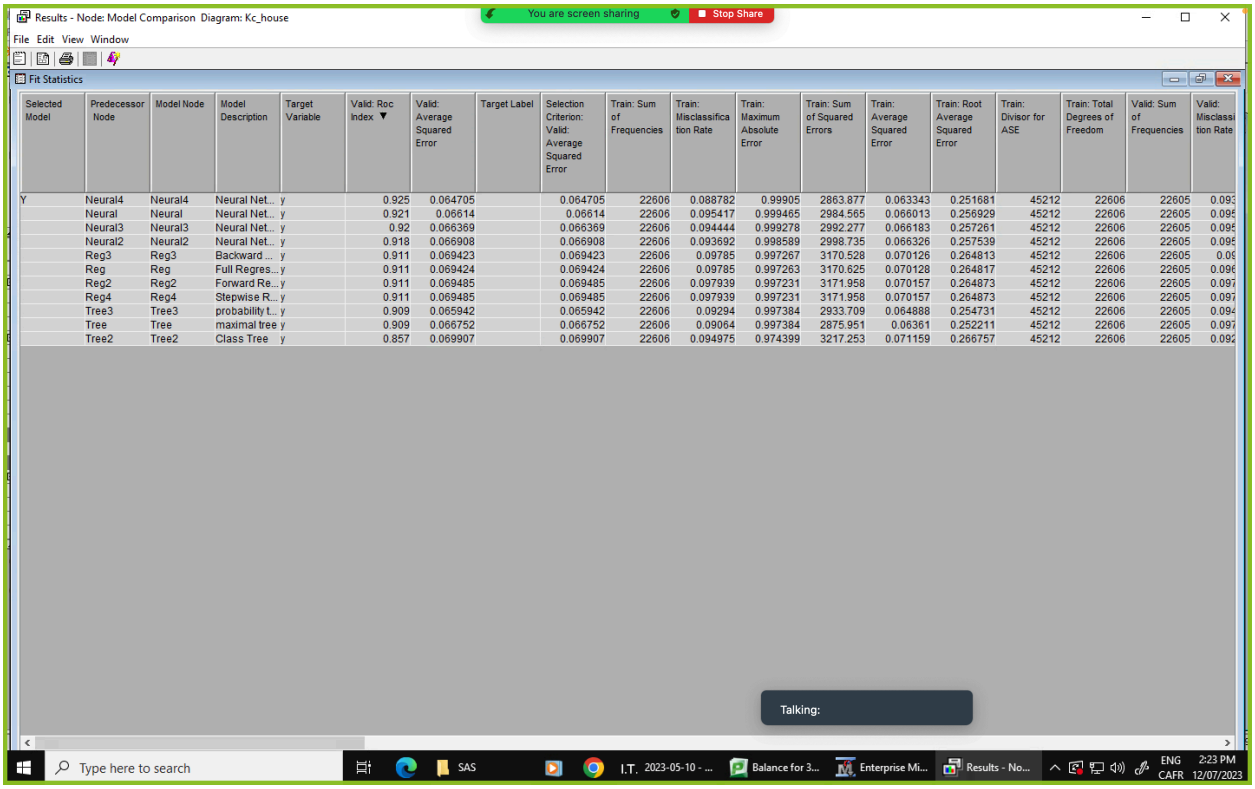
From the above summary table, our best neural network model is the one with the lowest ASE of **0.064705** which is the neural network (4) connected to backward regression.

Model comparison:

For our model comparison, use the ROC index and ASE to compare the three models, decision tree, regression, and neural network. Both the ROC index and the ASE are in agreements on the best model. The best model has the highest ROC index and the lowest ASE. Below are the screenshots of the model comparison, ROC curve and Fit statistics.

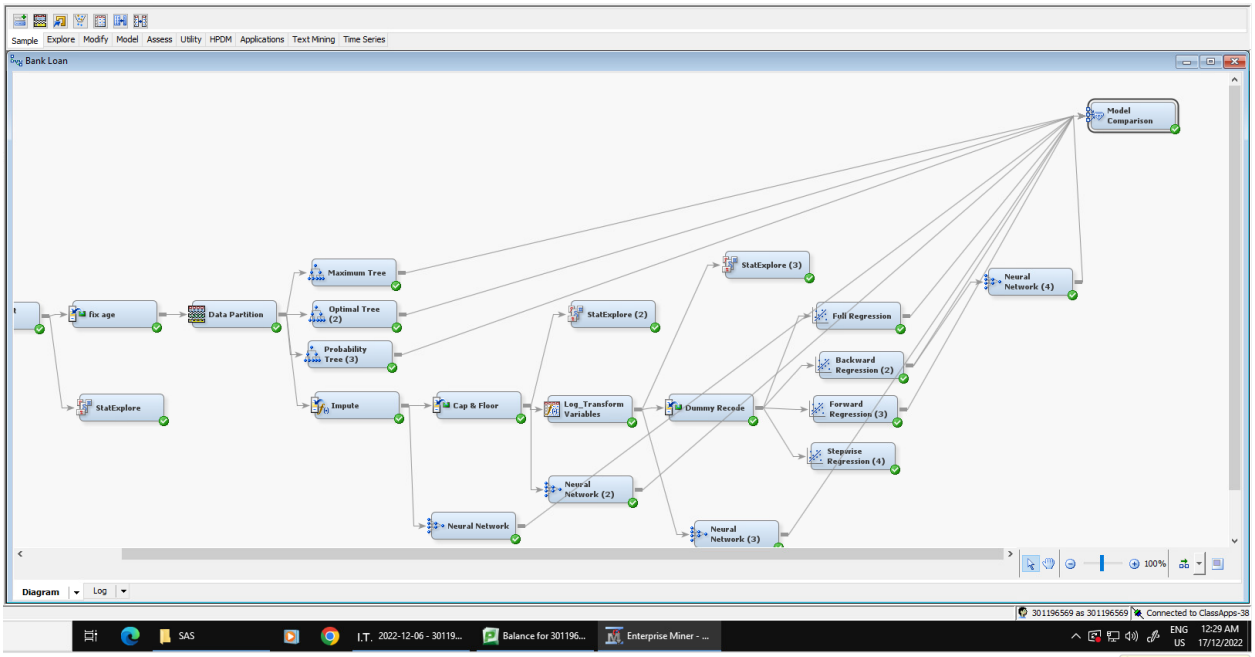
The specificity and sensitivity are measures used to evaluate the performance of the model in a binary classification task. Sensitivity also known as the True Positive rate or recall measures the portion of actual true positive case correctly identified. Thus, our sensitivity of 90.96% implies that the model correctly identifies 90.96% of the customers who said yes to the bank loan.

On the other hand, a specificity of 21.03% correctly identified the proportion of the customers who said no to the bank loan campaign. This implies that model only identifies 21% of actual customers who said no to the bank loan campaign as can be seen from the ROC curve above.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Roc Index ▼	Valid: Average Squared Error	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate
Y	Neural4	Neural4	Neural Net...	y	0.925	0.064705		0.064705	22606	0.088782	0.99905	2863.877	0.063343	0.251681	45212	22606	22605	0.093
	Neural	Neural	Neural Net...	y	0.921	0.06614		0.06614	22606	0.095417	0.999465	2984.565	0.066013	0.256929	45212	22606	22605	0.095
	Neural3	Neural3	Neural Net...	y	0.92	0.066369		0.066369	22606	0.094444	0.999278	2992.277	0.066183	0.257261	45212	22606	22605	0.095
	Neural2	Neural2	Neural Net...	y	0.918	0.066908		0.066908	22606	0.093692	0.998589	2998.735	0.066326	0.257539	45212	22606	22605	0.095
	Reg3	Reg3	Backward ...	y	0.911	0.069423		0.069423	22606	0.09785	0.997267	3170.528	0.070126	0.264813	45212	22606	22605	0.09
	Reg	Reg	Full Regres...	y	0.911	0.069424		0.069424	22606	0.09785	0.997263	3170.625	0.070128	0.264817	45212	22606	22605	0.096
	Reg2	Reg2	Forward Re...	y	0.911	0.069485		0.069485	22606	0.097939	0.997231	3171.958	0.070157	0.264873	45212	22606	22605	0.097
	Reg4	Reg4	Stepwise R...	y	0.911	0.069485		0.069485	22606	0.097939	0.997231	3171.958	0.070157	0.264873	45212	22606	22605	0.097
	Tree3	Tree3	probability t...	y	0.909	0.065942		0.065942	22606	0.09294	0.997384	2933.709	0.064888	0.254731	45212	22606	22605	0.094
	Tree	Tree	maximal tree	y	0.909	0.066752		0.066752	22606	0.09064	0.997384	2875.951	0.06361	0.252211	45212	22606	22605	0.097
	Tree2	Tree2	Class Tree	y	0.857	0.069907		0.069907	22606	0.094975	0.974399	3217.253	0.071159	0.266757	45212	22606	22605	0.092


Results - Node: Model Comparison Diagram: Kc_house
You are screen sharing
Stop Share
File Edit View Window
ROC Chart : y
Data Role = TRAIN
Data Role = VALIDATE
1 - Specificity = 0.21036
Sensitivity = 0.909607
Model Description = Neural Network 4h
Sensitivity
1 - Specificity
Talking:
probability tree
Class Tree
maximal tree
Stepwise Regres
Forward Regression
Full Regression
Neural Network 4h
Neural Network (3)
Neural Network (2)
Neural Network
Type here to search
SAS
I.T. 2023-05-10 - ...
Balance for 3...
Enterprise Mi...
Results - No...
ENG
CAFR
2:25 PM
12/07/2023


Sample
Explore
Modify
Model
Assess
Utility
HPDM
Applications
Text Mining
Time Series
Bank Loan
Model Comparison
fix age
Data Partition
Maximum Tree
Optimal Tree (2)
Probability Tree (3)
StatExplore
Impute
Cap & Floor
Log_Transform Variables
Dummy Recode
StatExplore (2)
StatExplore (3)
Full Regression
Backward Regression (2)
Forward Regression (3)
Stepwise Regression (4)
Neural Network (4)
Neural Network (2)
Neural Network
Neural Network (3)
100%
Diagram
Log
301196569 as 301196569
Connected to ClassApps-38
SAS
I.T. 2022-12-06 - 30119...
Balance for 301196...
Enterprise Miner - ...
ENG
US
12:29 AM
17/12/2022

Conclusion and recommendations

In summary, we performed model analytics on the bank dataset to predict the effectiveness of a marketing campaign by a bank using three different analytics models. The dataset consists of 45,211 observations and 17 variables. Our important variables are y (target variable), contact, housing and job. All our important variables have Chi-squares less than 0.0001.

In this project, we built decision trees, regressions, and neural networks to ascertain model optimization. Below is the summary of our best models with the model assessment measures.

S/N	MODELS	ASE	VALIDATION ROC INDEX
1	Probability Decision Tree	0.065942	0.909
2	Backward Regression	0.069423	0.911
3	Neural Network 4	0.064705	0.925

From the above table, our best model is the neural network model with the highest validation ROC index and least average squared error of 0.925 and 0.064705 respectively.

Limitation:

In the course of performing the analytics, we countered some challenges. First, we have skewed variables. We used cap and floor, standardize and finally log transformation but still have about three of the variables skewed. We were not able to transform all the variables. Secondly, we have dummy variables and we tried to use a dummy recode. However, our regression and neural network models' optimization were worse off with dummy recode replacement values. We, therefore, realized that our model without dummy recode have the optimal results, and it was difficult to interpret the odds ratio estimates of the model.

Recommendation:

We recommend that as analysts, there is a need to thoroughly review the data to ensure that irrelevant and redundant variables are rejected to ensure model fit.

From the business point of view, we will advise the client to pay attention to these variables in targeting customers to market the loan campaign. They are housing, loan, poutcome, and contact. These variables have been interpreted in the odds ratio interpretation section. Please refer to this section for better understanding of these variables and their impacts on the success of the loan campaign.

References

Camilosierra. (2020, June 30). *Bankdataset*. Kaggle. https://www.kaggle.com/datasets/juancamilosierra/bankdataset