Final Project

GROUP 7

Tu Nguyen

Chaw Hnin Nandar

Yuko Kurokawa

BA706 - Applied Analytic Modeling

Professor David Parent

July 15th, 2022

I. Introduction

Workforce is a crucial key for companies and organizations to be successful. Analyzing employees' records could improve the environment to keep excellent human resources and avoid employee attrition. Employee attrition could adversely affect organizations' productivity because they lose productive employees and resource for hiring and training new people.

This report focuses on predicting employee attrition using SAS Enterprise Miner, where decision trees, logistic regressions, and neural networks have been applied on the dataset found on Kaggle. The dataset has 35 columns and 1,029 samples of current and former employees. The steps for the prediction of employee attrition is the following;

1. Import the data
2. Explore and select the inputs
3. Perform decision trees
4. Build logistic regressions
5. Develop neural networks
6. Compare the models

II. Variable inputs

Dataset Description: The original dataset includes 35 columns and 1029 observations. The features' description is in Table 1.

Table.1 Dataset Description

Column Name	**Role**	**Type**	**Description**
Age	Input	Interval	Age
BusinessTravel	Input	Nominal	1=No Travel, 2=Travel Frequently, 3=Travel Rarely
DailyRate	Rejected	Interval	Salary Level
Department	Input	Nominal	1=HR, 2=R&D, 3=Sales
DistanceFromHome	Input	Interval	The distance from work to home
Education	Input	Nominal	1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
EducationField	Rejected	Nominal	1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TECHNICAL
EmployeeCount	Rejected	Interval	Employee or not
EmployeeNumber	ID	Nominal	Employee ID
EnvironmentSatisfaction	Input	Nominal	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Gender	Input	Nominal	(1=FEMALE, 2=MALE)
HourlyRate	Rejected	Interval	Hourly salary

JobInvolvement	Rejected	Nominal	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
JobLevel	Rejected	Nominal	Level of job
JobRole	Input	Nominal	1=HR REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= RESEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIVE, 9= SALES REPRESENTATIVE
JobSatisfaction	Input	Nominal	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
MaritalStatus	Input	Nominal	1=DIVORCED, 2=MARRIED, 3=SINGLE
MonthlyIncome	Input	Interval	Monthly Salary
MonthlyRate	Rejected	Interval	Monthly Rate
NumCompaniesWorked	Input	Interval	Number of companies worked at
Over18	Rejected	Binary	1=YES, 2=NO
OverTime	Input	Binary	1=NO, 2=YES
PercentSalaryHike	Rejected	Interval	Percentage increase in salary
PerformanceRating	Input	Nominal	Performance rating
RelationshipSatisfaction	Input	Nominal	Relations satisfaction
StandardHours	Rejected	Interval	Standard hours

StockOptionLevel	Input	Interval	Higher the number, the more stock option an employee has
TotalWorkingYears	Rejected	Interval	Total years of working
TrainingTimesLastYear	Input	Interval	Hours spent training
WorkLifeBalance	Input	Interval	Time spent between work and outside
YearsAtCompany	Input	Interval	Total number of years at the company
YearsInCurrentRole	Rejected	Interval	Years in current role
YearsSinceLastPromotion	Rejected	Interval	Last promotion
YearsWithCurrentManager	Rejected	Interval	Years spent with current manager
Attrition	Target	Binary	Employee leaving the company (0=No, 1=Yes)

The target variable, “Attrition”, describes the employee decision of either leaving the company (Yes) or staying in the company (No). We decided to make the attrition as the target variable depending on the factors having impact on whether the employees are leaving their current jobs. The selected variables to predict the target variable are shown above and some variables which are highly correlated and not necessarily important are rejected.

After analyzing, we can see that the percentage of employees resigning is less than the percentage staying in the same positions.

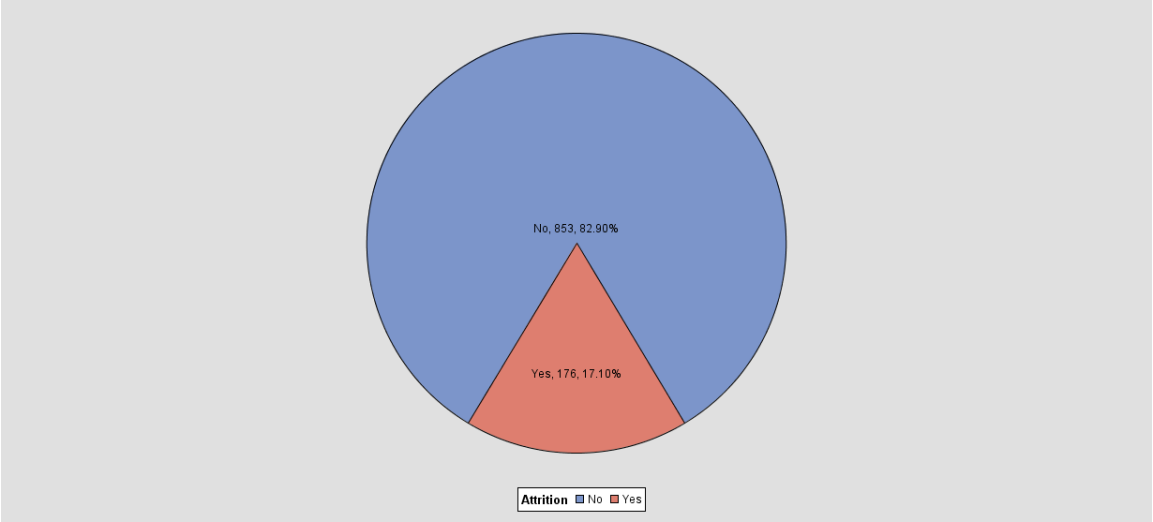


Figure.1 Attrition

Our data cleaning revealed that the frequency for the variables, ‘EmployeeCount’, ‘over18’, and ‘StandardHours’ are identical as shown below so that they were rejected as they are highly correlated.

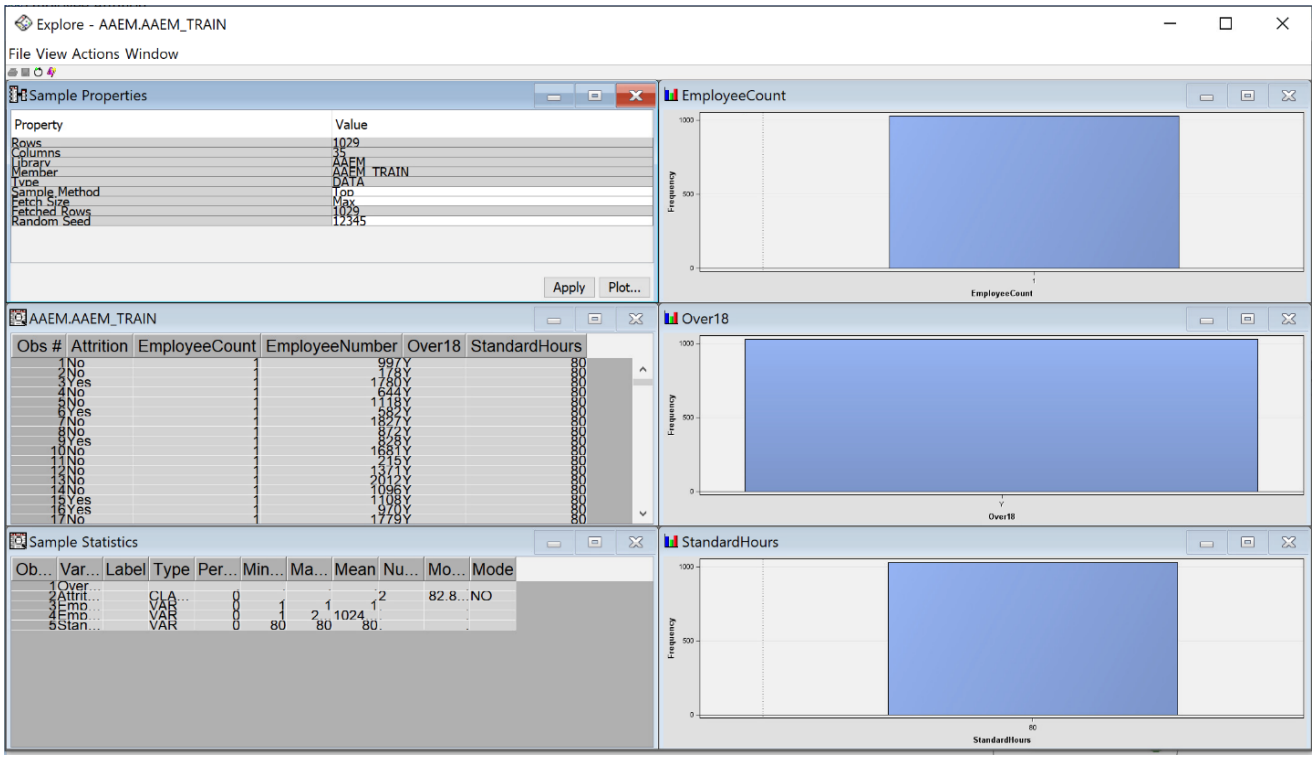


Figure.2 EmployeeCount, over18, and StandardHours

We also rejected other 11 variables considered as highly correlated:

- We rejected Educationfield column because it is highly correlated with Education apparently
- We rejected dailyrate, hourlyrate, monthlyrate and joblevel because they are correlated with MonthlyIncome
- We rejected JobInvolvement and PercentSalaryHike because they are correlated with PerformanceRate
- We rejected TotalWorkingYears, YearsInCurrManager, YearsWithCurrManager, YearsSinceLastPromotion because they are also correlated with YearsAtCompany

III. Modelings

A. DECISION TREE

1. Partition the data

For an honest assessment of model performance, we split the data into training data set and validation data set in half (50-50) as shown below. The training data set was used to fit the model, and the validation data was used to monitor and test the model. Data partition can avoid overfitting the model as it divides the dataset into training and valid parts to give the balance result of the model.

.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0

Figure.3 Data Partition Property Panel

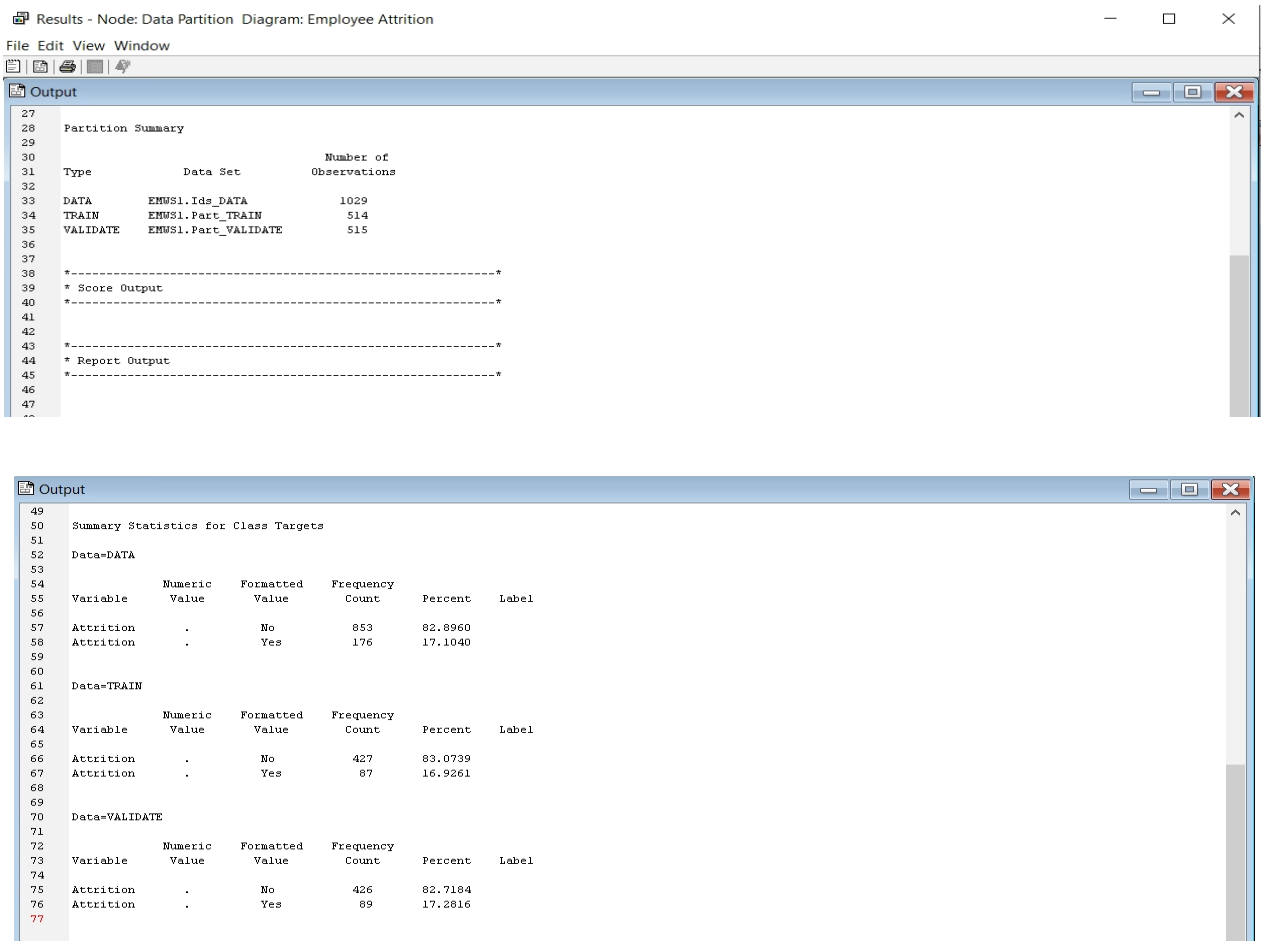


Figure.4 Data Partition output

Due to the odd number of the “Yes” and “No” of the target variable, The data proportions in both train and validate are not precisely the same.

2. Maximal Tree

Firstly, we built a Maximal Tree using an automated way. This tree has 13 leaves.

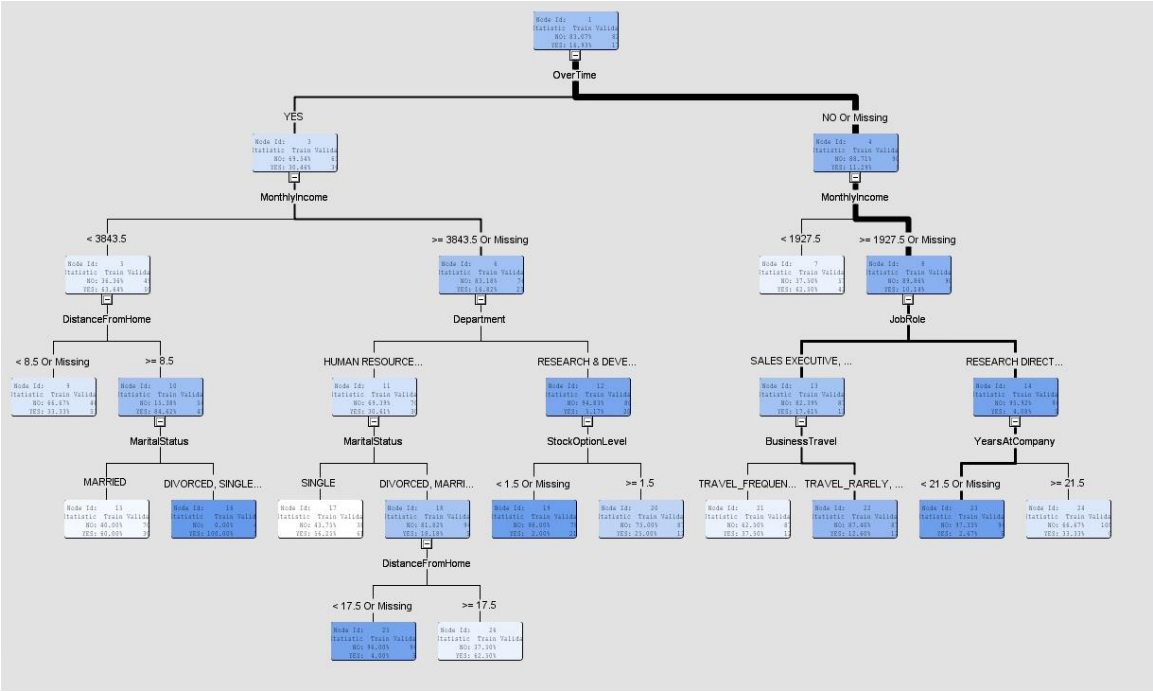


Figure.5 Maximal Tree

The subtree assessment in misclassification rate shows improvement in fit occurs in the train data set. However, we find this model overfitting when looking at the validation data set. Therefore, we built Misclassification Tree (MISC Tree) and Average Squared Error Tree (ASE Tree) to prune the model.

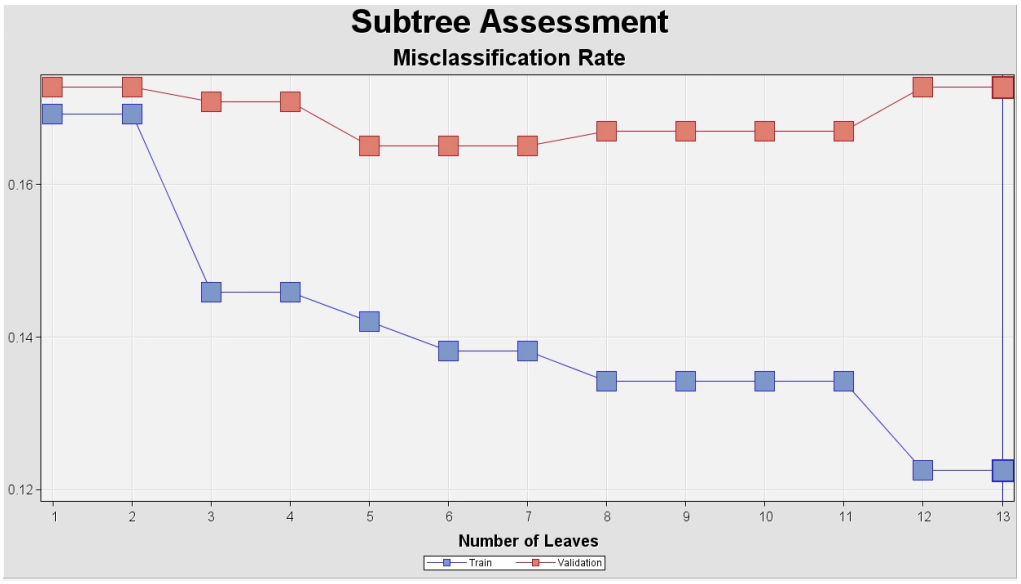


Figure.6 Maximal Tree Subtree Assessment

3. MISC Tree

Secondly, we developed Misclassification Tree (MISC Tree) using Assessment as a method and Misclassification as an assessment measure. Now, the number of leaves are reduced to 5 leaves.

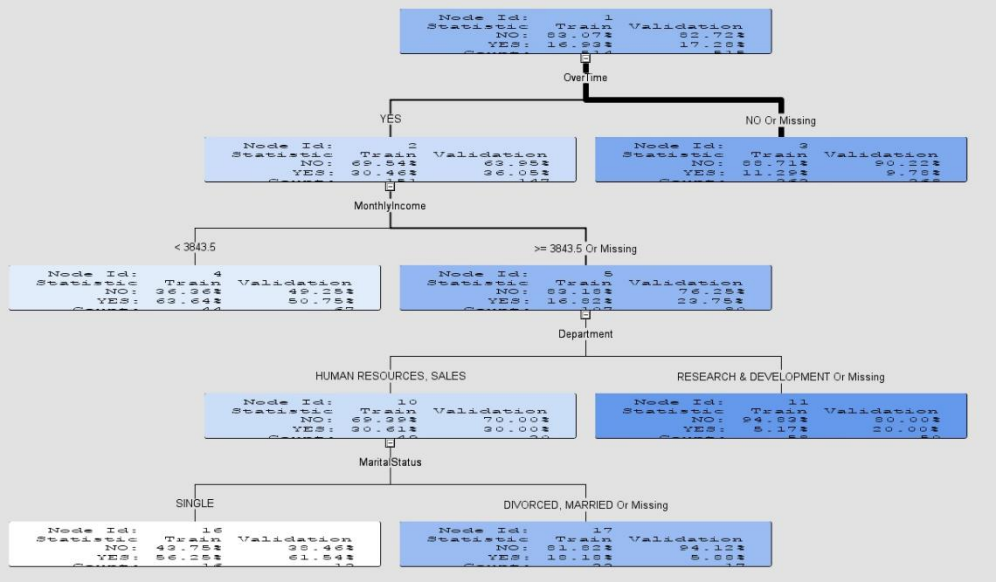


Figure.7 MISC Tree

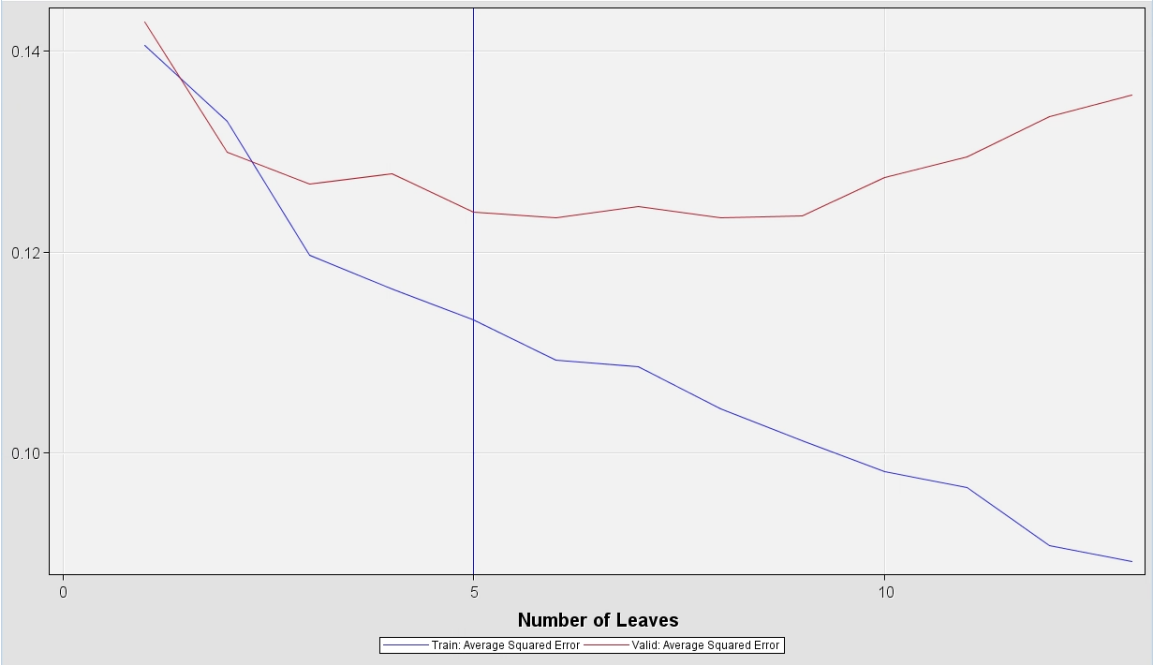


Figure.8 MISC Tree Subtree Assessment

Looking at the subtree assessment in Average Squared Error rate, it got slightly better than Maximal Tree.

4. ASE Tree

Lastly, we developed ASE Tree using Average Squared Error as an assessment measure to assess attrition probability. This tree has seven leaves.

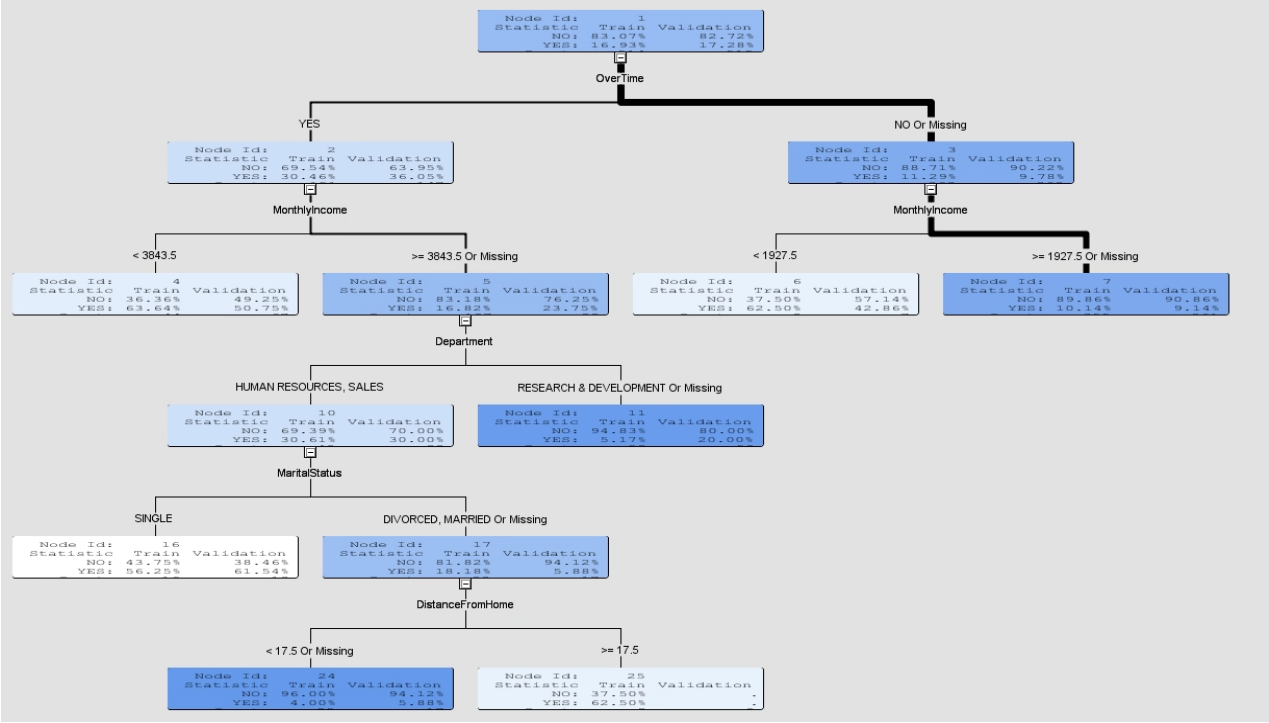


Figure.9 ASE Tree

In the subtree assessment in Average Square Error, a seven-leaf of validation data is optimal under the Average Squared Error criterion.

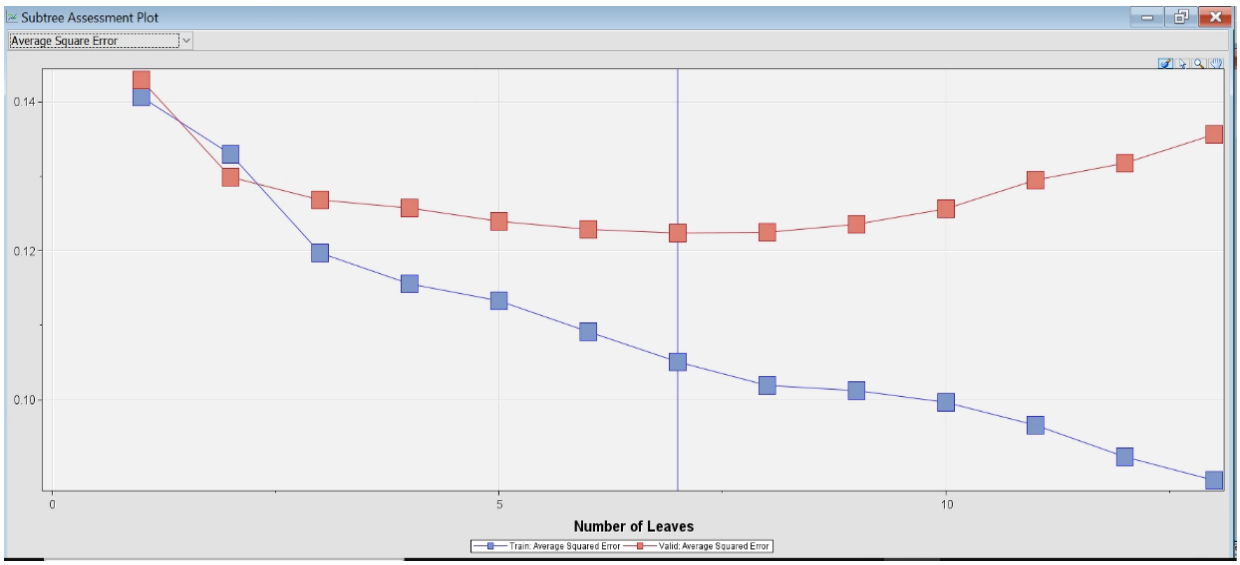


Figure.10 ASE Tree Subtree Assessment

5. Comparing the trees

Among the three trees, ASE Tree is the best model with the lowest Average Squared Error. The results are the following:

Models	**Selection Criterion: Average Squared Error**
Maximal Tree	**0.135631**
Misclassification Tree	**0.123947**
ASE Tree	**0.122375**

Table.2 Average Squared Error

According to ASE Tree, we can see that from the bold line that the employees who do not work overtime (overtime variable from table 1) and whose monthly salary (monthlyincome variable from table 1) is more than $1,925.7 would stay as their current jobs at the company.

B. REGRESSION

Regression is a parametric model which is different from decision tree models in predictions. Regression models assume the specific structure between inputs and the target. Model essentials of regression are to predict new cases, to select useful inputs and optimize complexity.

In every step, simple mathematical formula is used in different regression models and select the best model based on the fit statistics calculated with the least average squared error (ASE) of validation data. So, the first thing to do in regression is to manage the missing values as regression models use mathematical formula.

6. StatExplore

We connect the StatExplore node to the employee train node to find the missing values in the data we are using. In order to get the optimal model of regression and neural network, missing values have to be controlled well. First, we have to define if the trained data has missing values or not with StatExplore.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	BusinessTravel	INPUT	4	5	Travel_Rarely	70.26	Travel_Frequently	19.34
TRAIN	Department	INPUT	3	0	Research & Development	65.69	Sales	30.22
TRAIN	Gender	INPUT	2	0	Male	59.96	Female	40.04
TRAIN	JobRole	INPUT	9	0	Sales Executive	21.09	Research Scientist	20.80
TRAIN	MaritalStatus	INPUT	4	5	Married	46.06	Single	31.10
TRAIN	OverTime	INPUT	2	0	No	71.04	Yes	28.96
TRAIN	Attrition	TARGET	2	0	No	82.90	Yes	17.10

Figure.11 Missing values of interval variables

As the result, there are missing values in 4 variables: Business travel and marital status from class variable and age and distance from home from interval variable.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	37.93057	9.395978	893	136	18	37	60	0.237209	-0.55558
DistanceFromHome	INPUT	9.930407	8.421791	934	95	1	8	29	0.784726	-0.56551
Education	INPUT	2.892128	1.053541	1029	0	1	3	5	-0.27376	-0.64879
EnvironmentSatisfaction	INPUT	2.683188	1.096829	1029	0	1	3	4	-0.2658	-1.24234
HourlyRate	INPUT	66.68027	20.47409	1029	0	30	67	100	-0.08875	-1.21204
JobSatisfaction	INPUT	2.712342	1.096889	1029	0	1	3	4	-0.2896	-1.23404
NumCompaniesWorked	INPUT	2.653061	2.508186	1029	0	0	1	9	1.074606	0.083404
PerformanceRating	INPUT	3.159378	0.366206	1029	0	3	3	4	1.863897	1.476978
RelationshipSatisfaction	INPUT	2.68999	1.077767	1029	0	1	3	4	-0.26609	-1.1965
StockOptionLevel	INPUT	0.822157	0.874662	1029	0	0	1	3	0.956067	0.274039
TrainingTimesLastYear	INPUT	2.782313	1.283401	1029	0	0	3	6	0.558077	0.564671
WorkLifeBalance	INPUT	2.748299	0.697278	1029	0	1	3	4	-0.5439	0.420205
YearsAtCompany	INPUT	6.942663	6.068322	1029	0	0	5	37	1.670785	3.314189

Figure.12 Missing values of nominal variables

7. Imputation

Missing values in the working data are imputed to reduce the multiple collinearity. If there are still missing values, regression comes back and cannot compute. Missing values will not add anything to the model. So, we need to impute the missing values with the data we have, to get the optimal model.

The impute node from modify session is connected to the data partition node by setting “mean” as input method for interval variables.

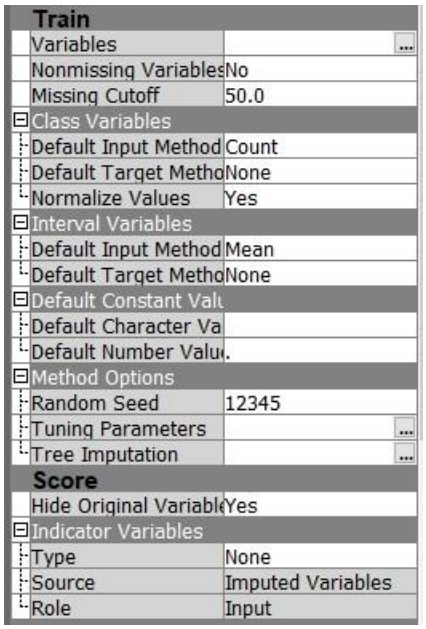


Figure.13 Impute node property panel

There are four imputed variables: age, business travel, distance from home and marital status. After imputation, the missing values are replaced by the average values of the current variables.

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP_Age	37.83632287	INPUT	INTERVAL		68
BusinessTravel	COUNT	IMP_BusinessTravel	Travel_Rarely	INPUT	NOMINAL		1
DistanceFromHome	MEAN	IMP_DistanceFromH...	10.133909287	INPUT	INTERVAL		51
MaritalStatus	COUNT	IMP_MaritalStatus	Married	INPUT	NOMINAL		4

Figure.14 Imputed variables

Then, we checked again with another StatExplore connected to the impute node if there were still missing values. We see there isn't any missing values in the trained variables.

8. Replacement (cap and floor)

We take the replacement node from the modify tab and connect to the transform variable to find confusing variables and to replace them or combine into one specific variable.

We set the default limit as none in the property panel of replacement which is also named as cap and floor. We found out that job role and business travel variables need to be replaced as shown in the screenshot.

We decided to group the upper management (UM) including job roles of research scientist, manufacturing director, manager, research director and human resources, sale group (S) as sales executive and sales representative, and group of health(H) as laboratory technician and healthcare representative and group of non travel (NT) which includes the variables of travel_rarely and non_travel.

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Attrition	No		427	C	No	.
Attrition	Yes		87	C	Yes	.
Attrition	_UNKNOWN_	**_DEFAULT_**	.	C		.
Department	Research & Development		340	C	Research & Development	.
Department	Sales		152	C	Sales	.
Department	Human Resources		22	C	Human Resources	.
Department	_UNKNOWN_	**_DEFAULT_**	.	C		.
Gender	Male		328	C	Male	.
Gender	Female		186	C	Female	.
Gender	_UNKNOWN_	**_DEFAULT_**	.	C		.
IMP_BusinessTravel	Travel_Rarely	**NT**	367	C	Travel_Rarely	.
IMP_BusinessTravel	Travel_Frequently		101	C	Travel_Frequently	.
IMP_BusinessTravel	Non-Travel	**NT**	46	C	Non-Travel	.
IMP_BusinessTravel	_UNKNOWN_	**_DEFAULT_**	.	C		.
IMP_MaritalStatus	Married		227	C	Married	.
IMP_MaritalStatus	Single		168	C	Single	.
IMP_MaritalStatus	Divorced		119	C	Divorced	.
IMP_MaritalStatus	_UNKNOWN_	**_DEFAULT_**	.	C		.
JobRole	Research Scientist	**UM**	109	C	Research Scientist	.
JobRole	Sales Executive	**S**	106	C	Sales Executive	.
JobRole	Laboratory Technician	**H**	86	C	Laboratory Technician	.
JobRole	Healthcare Representative	**H**	51	C	Healthcare Representative	.
JobRole	Manufacturing Director	**UM**	48	C	Manufacturing Director	.
JobRole	Manager	**UM**	38	C	Manager	.
JobRole	Research Director	**UM**	29	C	Research Director	.
JobRole	Sales Representative	**S**	28	C	Sales Representative	.
JobRole	Human Resources	**UM**	19	C	Human Resources	.
JobRole	_UNKNOWN_	**_DEFAULT_**	.	C		.
OverTime	No		363	C	No	.
OverTime	Yes		151	C	Yes	.
OverTime	_UNKNOWN_	**_DEFAULT_**	.	C		.

Figure.15 Changed Variables

9. Transform Variables

After that, we decided to add transform variable node from modify tab to check the skewness of the input variables are fit for the model building. The transform variable node is connected to the impute node.

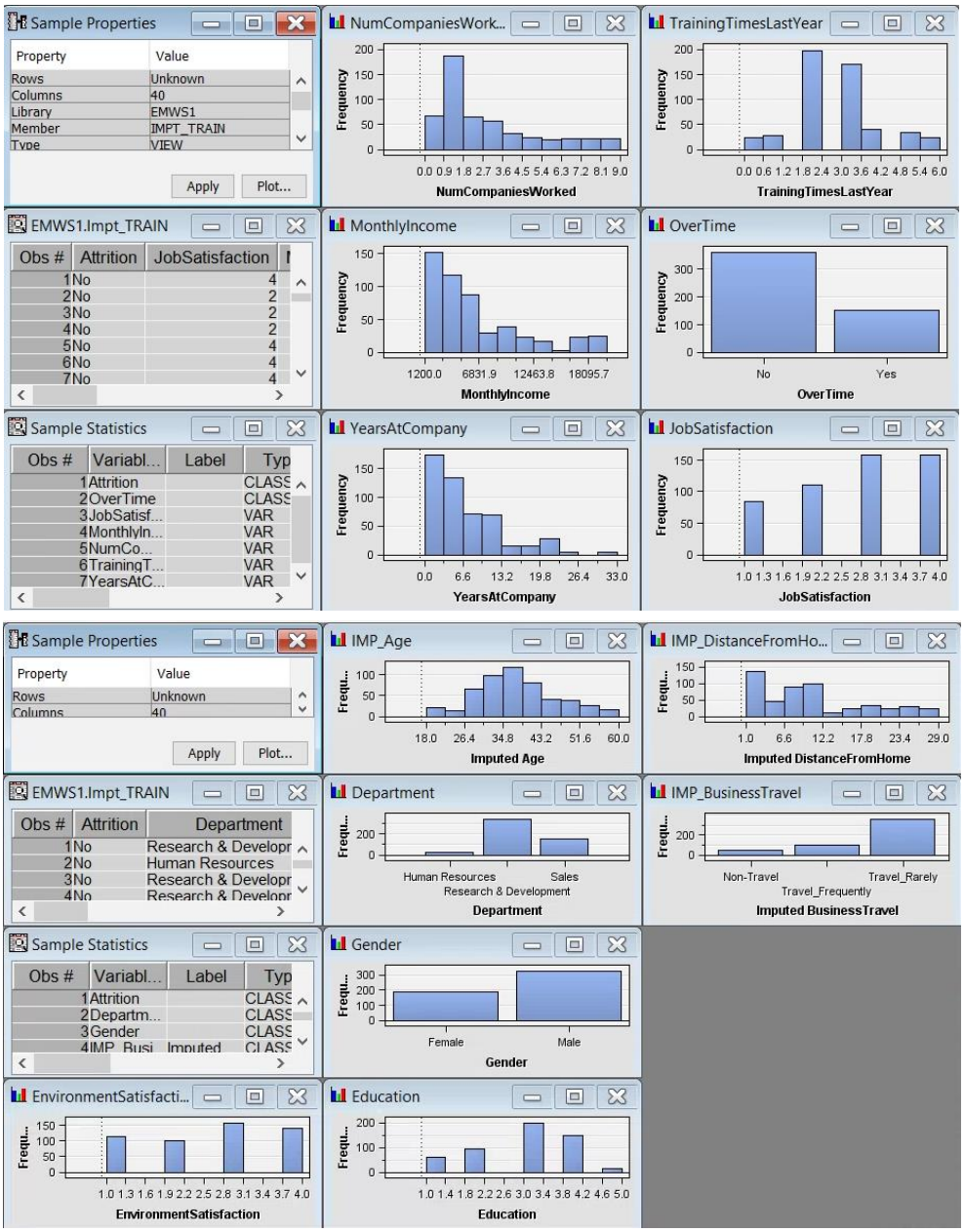
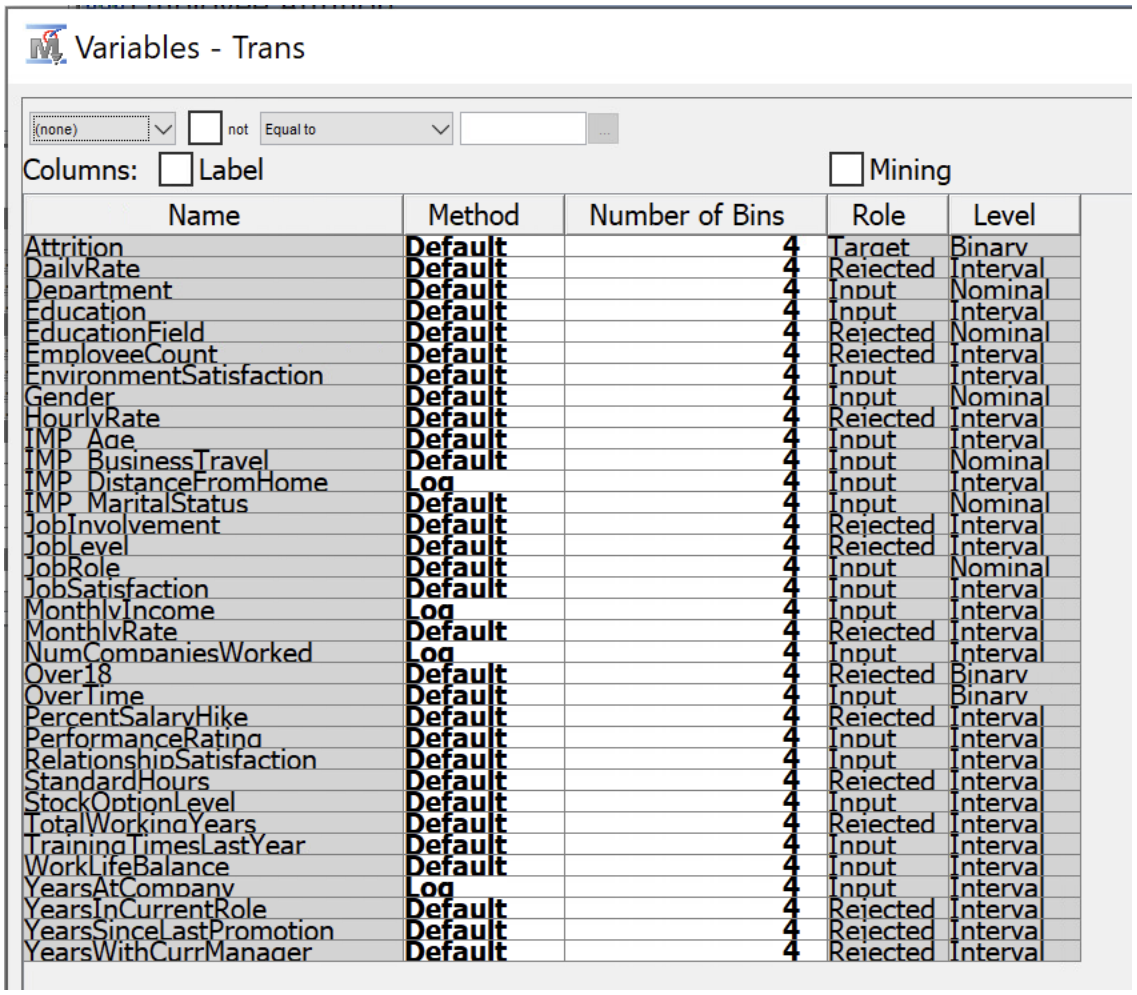


Figure.16 Skewed variables

After checking the skewness of the variables by selecting and exploring them, we saw the variables: distance from home, monthly income, number of companies worked, and years at the company need to be applied log transformation. Then, run the transformation node to complete the work.



Name	Method	Number of Bins	Role	Level
Attrition	Default	4	Target	Binary
DailyRate	Default	4	Rejected	Interval
Department	Default	4	Input	Nominal
Education	Default	4	Input	Interval
EducationField	Default	4	Rejected	Nominal
EmployeeCount	Default	4	Rejected	Interval
EnvironmentSatisfaction	Default	4	Input	Interval
Gender	Default	4	Input	Nominal
HourlyRate	Default	4	Rejected	Interval
IMP_Age	Default	4	Input	Interval
IMP_BusinessTravel	Default	4	Input	Nominal
IMP_DistanceFromHome	Log	4	Input	Interval
IMP_MaritalStatus	Default	4	Input	Nominal
JobInvolvement	Default	4	Rejected	Interval
JobLevel	Default	4	Rejected	Interval
JobRole	Default	4	Input	Nominal
JobSatisfaction	Default	4	Input	Interval
MonthlyIncome	Log	4	Input	Interval
MonthlyRate	Default	4	Rejected	Interval
NumCompaniesWorked	Log	4	Input	Interval
Over18	Default	4	Rejected	Binary
OverTime	Default	4	Input	Binary
PercentSalaryHike	Default	4	Rejected	Interval
PerformanceRating	Default	4	Input	Interval
RelationshipSatisfaction	Default	4	Input	Interval
StandardHours	Default	4	Rejected	Interval
StockOptionLevel	Default	4	Input	Interval
TotalWorkingYears	Default	4	Rejected	Interval
TrainingTimesLastYear	Default	4	Input	Interval
WorkLifeBalance	Default	4	Input	Interval
YearsAtCompany	Log	4	Input	Interval
YearsInCurrentRole	Default	4	Rejected	Interval
YearsSinceLastPromotion	Default	4	Rejected	Interval
YearsWithCurrManager	Default	4	Rejected	Interval

Figure.17 Transformation

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	IMP_DistanceFromHome		.	514	0	1	29	10.13391	7.961126	0.775746	-0.37867	Imputed Dist...
Input	Original	MonthlyIncome		.	514	0	1200	19973	6497.78	4799.732	1.336374	0.849219	
Input	Original	NumCompaniesWorked		.	514	0	0	9	2.63035	2.524641	1.155068	0.263645	
Input	Original	YearsAtCompany		.	514	0	0	33	6.974708	6.054121	1.474703	2.202319	
Output	Computed	LOG_IMP_DistanceFro...	log(IMP_DistanceFromHome + 1)	.	514	0	0.693147	3.401197	2.103396	0.843648	-0.35482	-1.04162	Transformed...
Output	Computed	LOG_MonthlyIncome	log(MonthlyIncome + 1)	.	514	0	7.09091	9.902187	8.542157	0.677102	0.336335	-0.82594	Transformed...
Output	Computed	LOG_NumCompanies...	log(NumCompaniesWorked + 1)	.	514	0	0	2.302585	1.067294	0.664528	0.187733	-0.80922	Transformed...
Output	Computed	LOG_YearsAtCompany	log(YearsAtCompany + 1)	.	514	0	0	3.526361	1.801509	0.765505	-0.16575	-0.44553	Transformed...

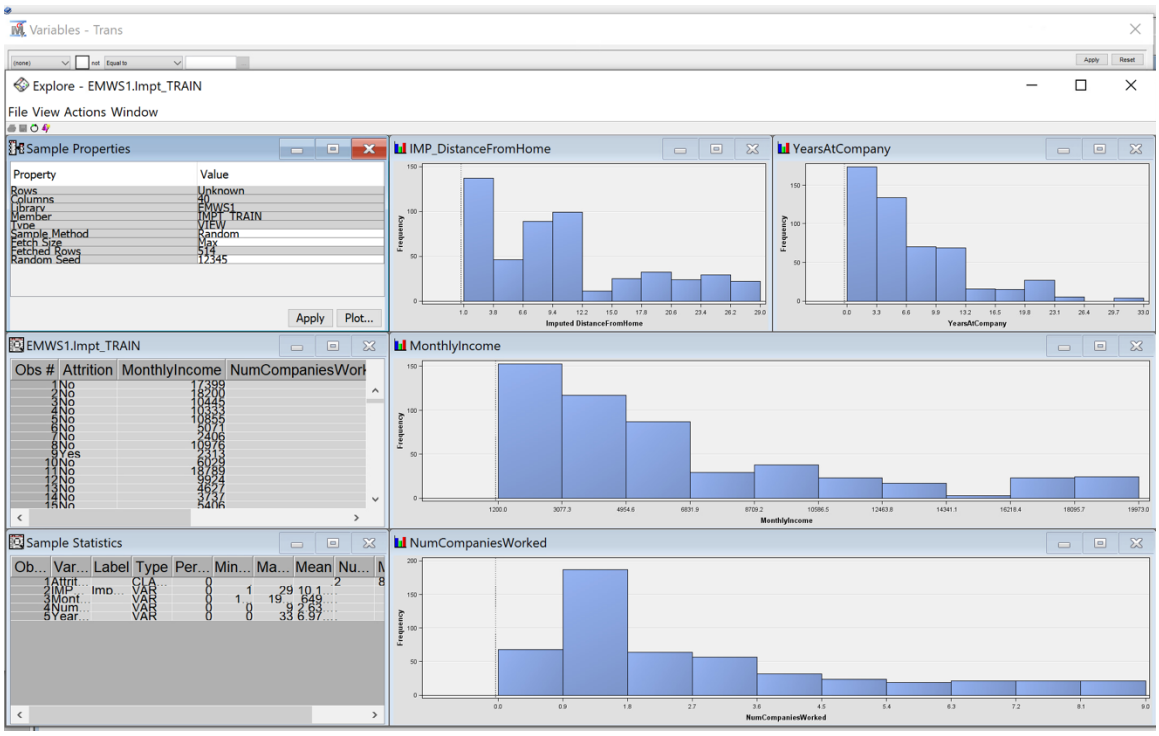
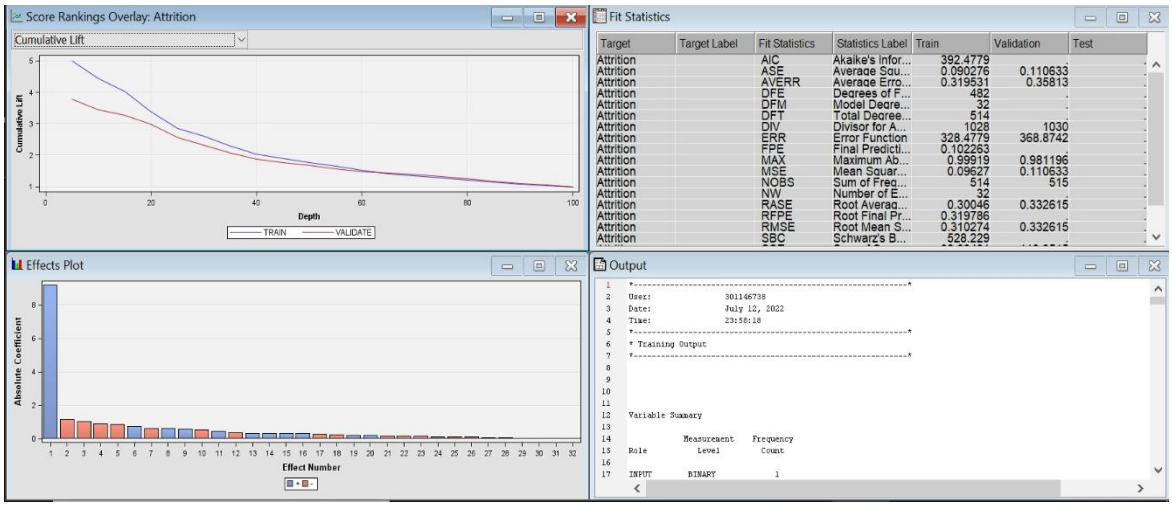


Figure.18 Transformed variables

10. Full Regression

The new regression node from the model tab is selected and connected to the cap and floor node. Then, set the selection model as none in its property panel.

Figure.19 Full Regression results



Results of full regression

- The ASE of validation is 0.110633.
- There are total 29 variables.

11. Forward Regression

The regression node from the model tab is selected and connected to the cap and floor node. Then, we set the selection model as forward and the selection criterion as validation error in its property panel.

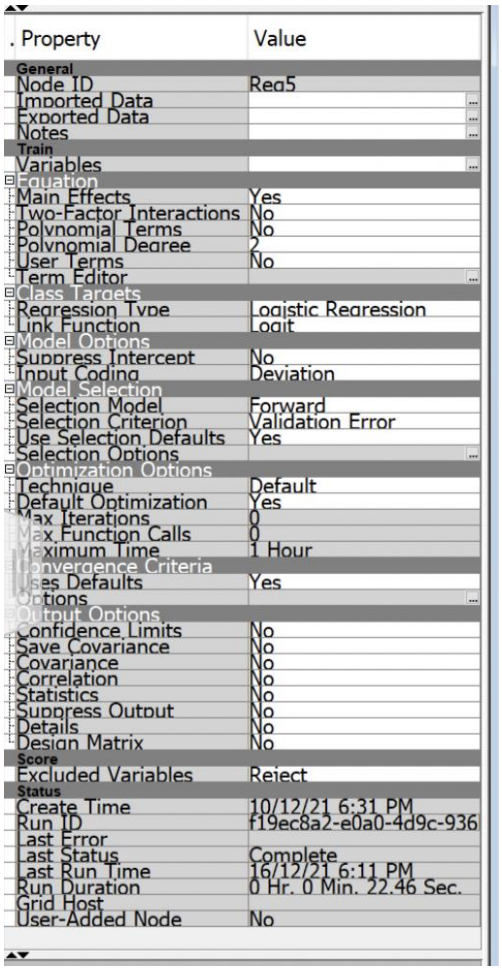


Figure.20 Forward Regression property panel

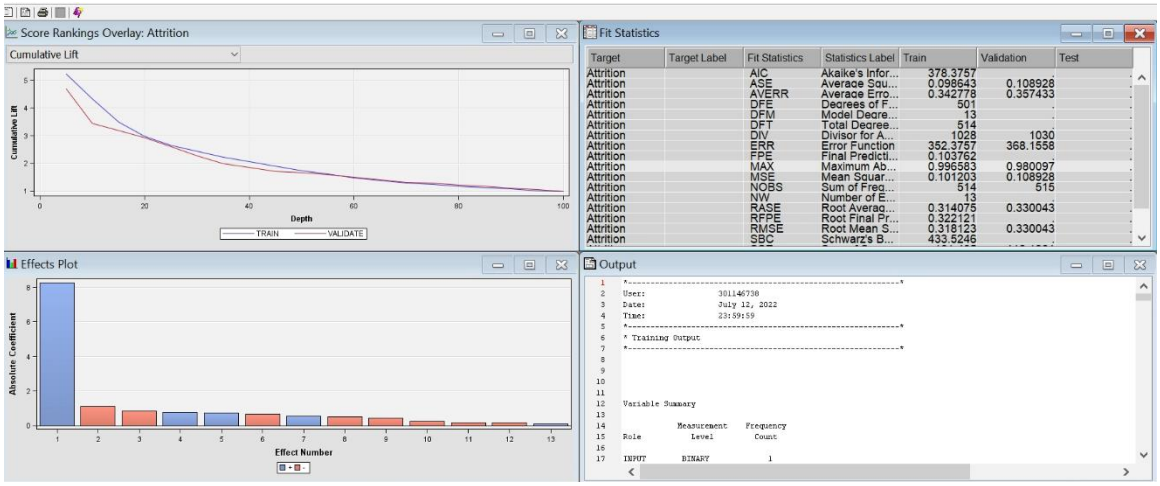


Figure.21 Forward Regression Results

Forward regression results:

- The ASE of Validation from fit statistics is 0.108928.
- 13 parameters are included in the final model. They are 'Intercept', 'EnvironmentStaisfaction1', 'EnvironmentStaisfaction2', 'EnvironmentStaisfaction3' 'IMP_MaritalStatus_Divorced', 'IMP_MaritalStatus_Married', 'LOG_MonthlyIncome', 'LOG_NumCompaniesWorked', 'LOG_YearsAtCompany', 'OverTime_N0', 'REP_IMP_BusinessTravel_NT', 'REP_JobRole_H', 'REP_JobRole_S'.

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	8.2829	2.2539	13.50	0.0002		999.000
EnvironmentSatisfaction	1	1	0.7850	0.2313	11.52	0.0007		2.192
EnvironmentSatisfaction	2	1	-0.1544	0.2744	0.32	0.5738		0.857
EnvironmentSatisfaction	3	1	-0.1621	0.2283	0.50	0.4778		0.850
IMP_MaritalStatus	Divorced	1	-0.6610	0.2511	6.93	0.0085		0.516
IMP_MaritalStatus	Married	1	0.0819	0.1948	0.18	0.6743		1.085
LOG_MonthlyIncome		1	-1.1195	0.2908	14.82	0.0001	-0.4179	0.326
LOG_NumCompaniesWorked		1	0.5542	0.2194	6.38	0.0115	0.2030	1.741
LOG_YearsAtCompany		1	-0.4359	0.2107	4.28	0.0386	-0.1840	0.647
OverTime	No	1	-0.8482	0.1460	33.75	<.0001		0.428
REP_IMP_BusinessTravel	NT	1	-0.5130	0.1577	10.58	0.0011		0.599
REP_JobRole	H	1	-0.2623	0.2051	1.64	0.2010		0.769
REP_JobRole	S	1	0.7436	0.1970	14.25	0.0002		2.104

Figure.22 Analysis of Maximum Likelihood Estimates

Odds ratio estimates interpretation

Odds Ratio Estimates

Effect		Point Estimate
EnvironmentSatisfaction	1 vs 4	3.503
EnvironmentSatisfaction	2 vs 4	1.369
EnvironmentSatisfaction	3 vs 4	1.359
IMP_MaritalStatus	Divorced vs Single	0.289
IMP_MaritalStatus	Married vs Single	0.608
LOG_MonthlyIncome		0.326
LOG_NumCompaniesWorked		1.741
LOG_YearsAtCompany		0.647
OverTime	No vs Yes	0.183
REP_IMP_BusinessTravel	NT vs Travel_Frequently	0.358
REP_JobRole	H vs UM	1.245
REP_JobRole	S vs UM	3.404

- People with low environment satisfaction have 3 times more likely to leave the job compared to those with high environment satisfaction.
- People with medium and high environment satisfaction are 37 and 36 percent respectively more likely to leave the job compared to those with highest environment satisfaction.
- “Divorced” and “Married” are less likely to leave the job compared with “Singles” at 71% and 39%, respectively.
- For “Log monthly income”, natural log 2.71 raised 1.1195 times of change, there is 67% less likely to leave the job.
- For “Log number of company work”, natural log 2.71 raised 0.5542 times of change, there is 74%more likely to leave the job.
- For “Log Years at company”, natural log 2.71 raised 0.4359 times of change, there is 35% less likely to leave the job.
- Those with no “Overtime” are 82% less likely to resign from the job than with overtime.
- Those with “non-travel” are 64% less likely to leave the job compared to those who travel frequently.
- Employees with health job roles are 25% more likely and those with sales position are more than 3 times chance to leave the job compared to those from upper management level.

12. Backward Regression

The new regression node from the model tab is selected and connected to the cap and floor node. Then, set the selection model as backward in its property panel.

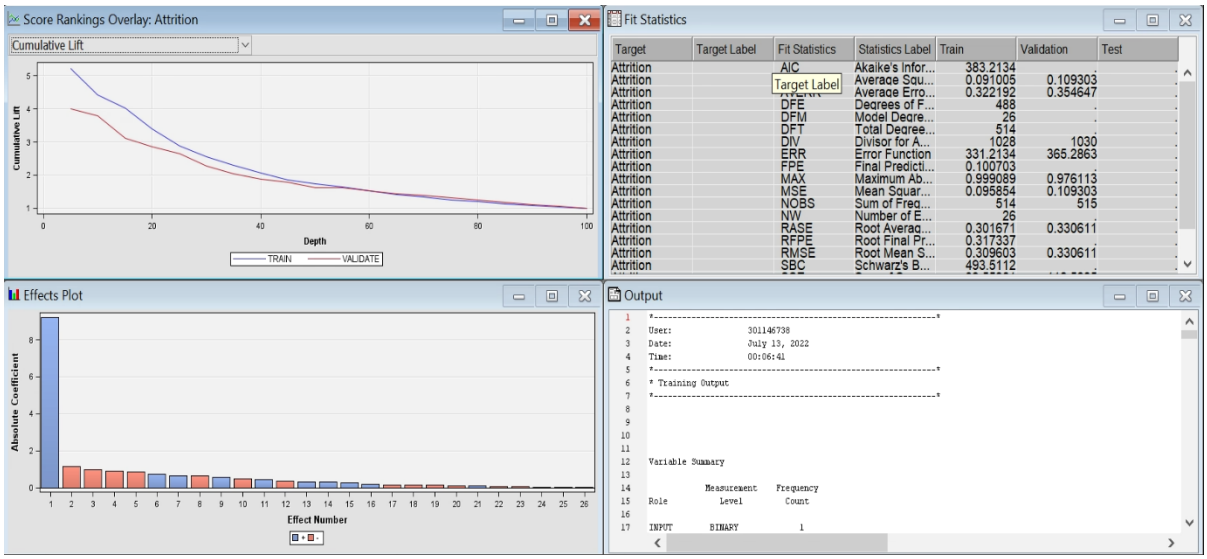


Figure.24 Backward Regression results

Results of backward regression

- The ASE of validation is 0.109303.
- There are total 26 variables.

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	9.2191	2.6367	12.23	0.0005		999.000
Department	Human Resources	1	0.5509	0.5662	0.95	0.3306		1.735
Department	Research & Development	1	-0.9950	0.4778	4.34	0.0373		0.370
EnvironmentSatisfaction	1	1	0.7426	0.2455	9.15	0.0025		2.101
EnvironmentSatisfaction	2	1	-0.1415	0.2852	0.25	0.6198		0.868
EnvironmentSatisfaction	3	1	-0.1167	0.2370	0.24	0.6225		0.890
IMP_Age		1	-0.0135	0.0194	0.48	0.4888	-0.0648	0.987
IMP_MaritalStatus	Divorced	1	-0.8768	0.3132	7.84	0.0051		0.416
IMP_MaritalStatus	Married	1	0.00861	0.2116	0.00	0.9675		1.009
JobSatisfaction	1	1	0.3062	0.2799	1.20	0.2741		1.358
JobSatisfaction	2	1	0.1115	0.2688	0.17	0.6783		1.118
JobSatisfaction	3	1	-0.0555	0.2305	0.06	0.8098		0.946
LOG_IMP_DistanceFromHome		1	0.3261	0.1856	3.09	0.0789	0.1517	1.386
LOG_MonthlyIncome		1	-1.1531	0.3327	12.01	0.0005	-0.4304	0.316
LOG_NumCompaniesWorked		1	0.6557	0.2409	7.41	0.0065	0.2402	1.927
LOG_YearsAtCompany		1	-0.3729	0.2250	2.75	0.0974	-0.1574	0.689
OverTime	No	1	-0.8603	0.1536	31.37	<.0001		0.423
REP_IMP_BusinessTravel	NT	1	-0.6324	0.1698	13.86	0.0002		0.531
REP_JobRole	H	1	0.2874	0.4621	0.39	0.5340		1.333
REP_JobRole	S	1	-0.0782	0.7782	0.01	0.9200		0.925
RelationshipSatisfaction	1	1	0.4203	0.2580	2.65	0.1033		1.522
RelationshipSatisfaction	2	1	0.0255	0.2619	0.01	0.9224		1.026
RelationshipSatisfaction	3	1	-0.4828	0.2467	3.83	0.0503		0.617
StockOptionLevel		1	0.2085	0.2177	0.92	0.3382	0.1017	1.232
TrainingTimesLastYear		1	-0.1362	0.1199	1.29	0.2561	-0.0973	0.873
WorkLifeBalance		1	-0.1440	0.2075	0.48	0.4877	-0.0563	0.866

Figure.25 Analysis of Maximum Likelihood Estimates

Figure.25 Analysis of Maximum Likelihood Estimates

Odds ratios estimates interpretation

Odds Ratio Estimates

Effect		Point Estimate
Department	Human Resources vs Sales	1.113
Department	Research & Development vs Sales	0.237
EnvironmentSatisfaction	1 vs 4	3.411
EnvironmentSatisfaction	2 vs 4	1.409
EnvironmentSatisfaction	3 vs 4	1.444
IMP_Age		0.987
IMP_MaritalStatus	Divorced vs Single	0.175
IMP_MaritalStatus	Married vs Single	0.423
JobSatisfaction	1 vs 4	1.951
JobSatisfaction	2 vs 4	1.606
JobSatisfaction	3 vs 4	1.359
LOG_IMP_DistanceFromHome		1.386
LOG_MonthlyIncome		0.316
LOG_NumCompaniesWorked		1.927
LOG_YearsAtCompany		0.689
OverTime	No vs Yes	0.179
REP_IMP_BusinessTravel	NT vs Travel_Frequently	0.282
REP_JobRole	H vs UM	1.643
REP_JobRole	S vs UM	1.140
RelationshipSatisfaction	1 vs 4	1.467
RelationshipSatisfaction	2 vs 4	0.989
RelationshipSatisfaction	3 vs 4	0.595
StockOptionLevel		1.232
TrainingTimesLastYear		0.873
WorkLifeBalance		0.866

Figure.26 Odds Ratio Estimates

Odd ratio estimates interpretation

- Human resource and research & development department are 11% more likely and 76% less likely to leave the job compared to sales department.
- People with low environment satisfaction have 3 times more likely to leave the job compared to those with high environment satisfaction.
- People with medium and high environment satisfaction are 41 and 44 percent respectively more likely to leave the job compared to those with highest environment satisfaction.

- People with older age are 1% less likely to leave the job
- “Divorced” are 82% and “Married” are 58% less likely to leave the job than singles.
- Employees with low, medium and high job satisfaction are 95%,60% and 36% more likely to resign from the job compared to those with highest job satisfaction.
- For “Log distance from home”, natural log, 2.71 raised 0.3261 times of change, there is 39% chance to leave the job.
- For “Log monthly income”, natural log 2.71 with 1.1531 times of change of income, there is 68% less likely to leave the job.
- For “Log number of company work”, 2.71 raised 0.3729 times of change of number of companies, there is 92% more likely to leave the job.
- For “Log year at company”, 2.71 raised 0.6557 times of change of year at company, there is 31% less likely to leave the job.
- Those with no “Overtime” are 82% less likely to leave the job than those with overtime.
- Those with “Non_travel” are 72% less likely to leave the job compared to “Travel_frequently”.
- Health job roles are 64% and sales are 14% more chance to leave the job compared to upper management level.
- Employees with low relationship satisfaction are 47% more likely to change the job compared to those with highest relationship satisfaction.
- Employees with medium and high relationship satisfaction are 1% and 40% less likely to leave the job compared to those with highest relationship satisfaction.
- The high stock option level makes the employees 23% more to leave the job.
- Employees with training times in the previous year are 12% less likely to leave the job.
- Employees with work life balance are 13% less likely to leave the job.

13. Stepwise Regression

The new regression node from the model tab is selected and connected to the cap and floor node.

Then, set the selection model as backward in its property panel.

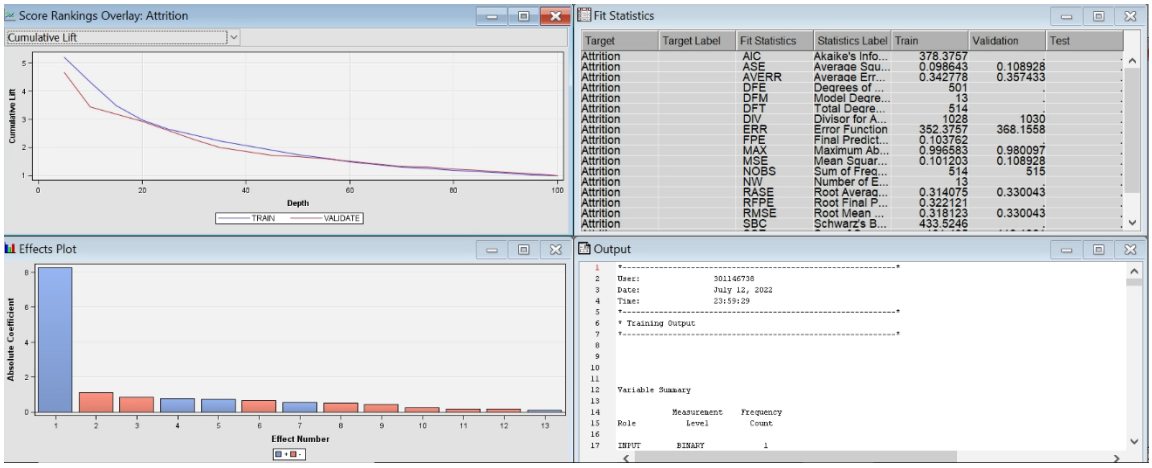


Figure.27 Stepwise Regression results

Results of stepwise regression

- The ASE of validation is 0.108928.
- There are total 13 variables.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	8.2829	2.2539	13.50	0.0002		999.000
EnvironmentSatisfaction	1	1	0.7850	0.2313	11.52	0.0007		2.192
EnvironmentSatisfaction	2	1	-0.1544	0.2744	0.32	0.5738		0.857
EnvironmentSatisfaction	3	1	-0.1621	0.2283	0.50	0.4778		0.850
IMP_MaritalStatus	Divorced	1	-0.6610	0.2511	6.93	0.0085		0.516
IMP_MaritalStatus	Married	1	0.0819	0.1948	0.18	0.6743		1.085
LOG_MonthlyIncome		1	-1.1195	0.2908	14.82	0.0001	-0.4179	0.326
LOG_NumCompaniesWorked		1	0.5542	0.2194	6.38	0.0115	0.2030	1.741
LOG_YearsAtCompany		1	-0.4359	0.2107	4.28	0.0386	-0.1840	0.647
OverTime	No	1	-0.8482	0.1460	33.75	<.0001		0.428
REP_IMP_BusinessTravel	NT	1	-0.5130	0.1577	10.58	0.0011		0.599
REP_JobRole	H	1	-0.2623	0.2051	1.64	0.2010		0.769
REP_JobRole	S	1	0.7436	0.1970	14.25	0.0002		2.104

Odds Ratio Estimates

Effect		Point Estimate
EnvironmentSatisfaction	1 vs 4	3.503
EnvironmentSatisfaction	2 vs 4	1.369
EnvironmentSatisfaction	3 vs 4	1.359
IMP_MaritalStatus	Divorced vs Single	0.289
IMP_MaritalStatus	Married vs Single	0.608
LOG_MonthlyIncome		0.326
LOG_NumCompaniesWorked		1.741
LOG_YearsAtCompany		0.647
OverTime	No vs Yes	0.183
REP_IMP_BusinessTravel	NT vs Travel_Frequently	0.358
REP_JobRole	H vs UM	1.245
REP_JobRole	S vs UM	3.404

Figure.28 Analysis of Maximum Likelihood Estimates and Odds Ratio Estimate

Odd ratio estimates interpretation

- Workers with low environment satisfaction have 3.5 times more likely to leave the job compared to those with high environment satisfaction.
- Workers with medium and high environment satisfaction are 37 and 36 percent respectively more likely to leave the job compared to those with highest environment satisfaction.
- "Divorced" are 71% and "Married" are 39% less likely to leave the job than singles.
- For "Log monthly income", natural log 2.71 with 1.1195 times of change of income, there is 67% less likely to leave the job.
- For "Log number of company work", 2.71 raised 0.5542 times of change of number of companies, there is 74% more likely to leave the job.
- For "Log year at company", 2.71 raised 0.4359 times of change of year at company, there is 35% less likely to leave the job.

- Employees with no “Overtime” are 81% less likely to leave the job than those with overtime.
- Those with “Non_business_travel” are 64% less likely to leave the job compared to “Travel_frequently”.
- Health job roles are 25% more and sales are 3 times more chances to leave the job compared to upper management level.

In conclusion of all regression models, the ASE of validation of forward and stepwise regression is the same. Therefore, they are considered to be the best regression models for this trained data.

C. NEURAL NETWORK

Neural networks require a complete record for estimation and scoring. We run up to 9 hidden units to get the best one of neural network after impute node. For each neural network, we developed each hidden unit starting from 3 to 9 with 50 and 100 interations. However, we saw that running with 50 interations always gave us the best results so the screenshots below only show 50 interations. Additionallly, we also run the same types of hidden units after transform variables node.

1. After imputation

    a) 3 hidden units

Figure. 30 Property panel

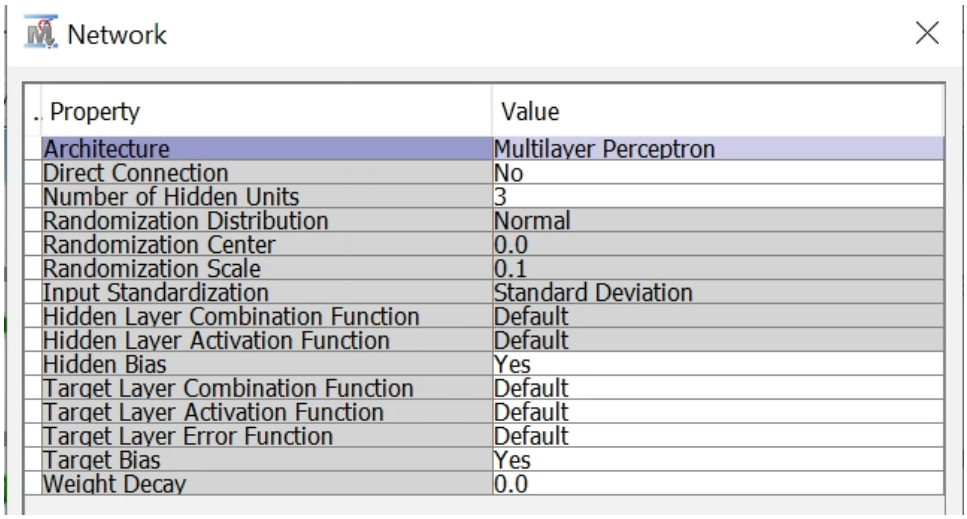
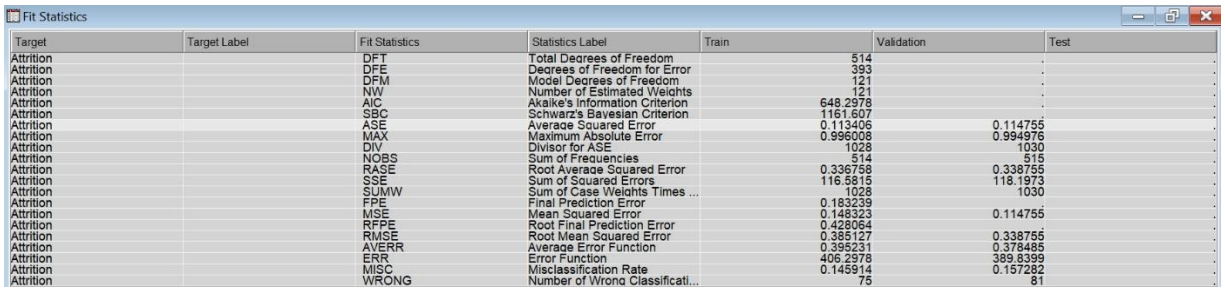


Figure. 31 Fit statistics



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	393	.	.
Attrition		DFM	Model Degrees of Freedom	121	.	.
Attrition		NW	Number of Estimated Weights	121	.	.
Attrition		AIC	Akaike's Information Criterion	648.2978	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1161.607	.	.
Attrition		ASE	Average Squared Error	0.113406	0.114755	.
Attrition		MAX	Maximum Absolute Error	0.996008	0.994976	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.336758	0.338755	.
Attrition		SSE	Sum of Squared Errors	116.5815	118.1973	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.183239	.	.
Attrition		MSE	Mean Squared Error	0.148323	0.114755	.
Attrition		RFPE	Root Final Prediction Error	0.428064	.	.
Attrition		RMSE	Root Mean Squared Error	0.385127	0.338755	.
Attrition		AVERR	Average Error Function	0.395231	0.378485	.
Attrition		ERR	Error Function	406.2978	389.8399	.
Attrition		MISC	Misclassification Rate	0.145914	0.157282	.
Attrition		WRONG	Number of Wrong Classificati...	75	81	.

Outcome: ASE of validation is 0.114755

b) 4 hidden units

Figure. 32 Property panel

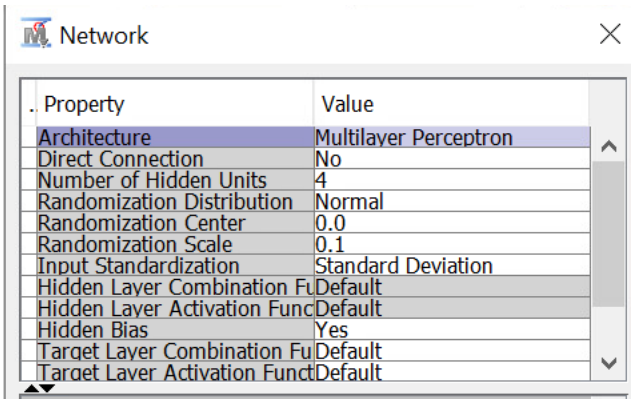
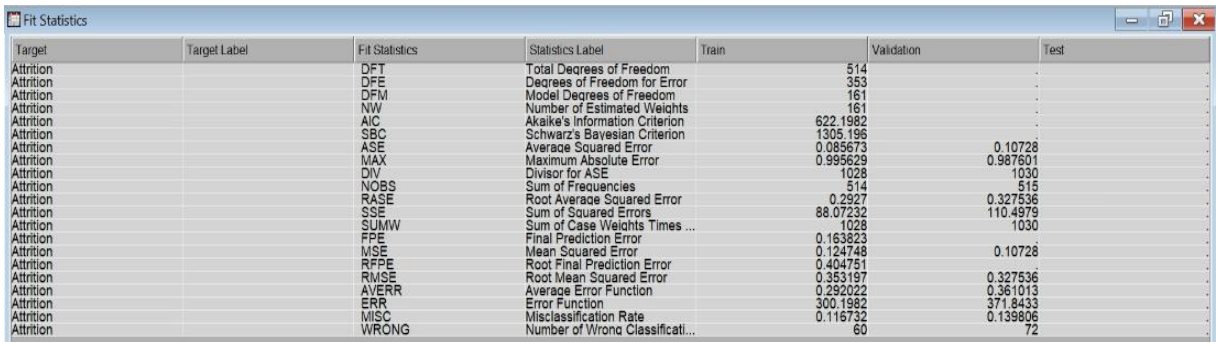


Figure. 33 Fit statistics



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	353	.	.
Attrition		DFM	Model Degrees of Freedom	161	.	.
Attrition		NW	Number of Estimated Weights	161	.	.
Attrition		AIC	Akaike's Information Criterion	622.1982	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1305.196	.	.
Attrition		ASE	Average Squared Error	0.085673	0.10728	.
Attrition		MAX	Maximum Absolute Error	0.995629	0.987601	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.2927	0.327536	.
Attrition		SSE	Sum of Squared Errors	88.07232	110.4979	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.163823	.	.
Attrition		MSE	Mean Squared Error	0.124748	0.10728	.
Attrition		RFPE	Root Final Prediction Error	0.404751	.	.
Attrition		RMSE	Root Mean Squared Error	0.353197	0.327536	.
Attrition		AVERR	Average Error Function	0.292022	0.361013	.
Attrition		ERR	Error Function	300.1982	371.8433	.
Attrition		MISC	Misclassification Rate	0.116732	0.139806	.
Attrition		WRONG	Number of Wrong Classificati...	60	72	.

Outcome: ASE of validation is 0.10728

c) 5 hidden units

Figure. 34 Property panel

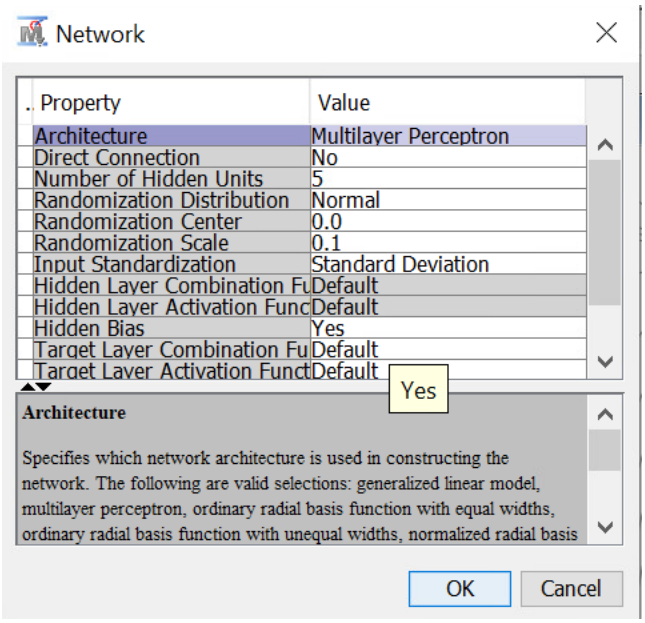


Figure. 35 Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	313	.	.
Attrition		DFM	Model Degrees of Freedom	201	.	.
Attrition		NW	Number of Estimated Weights	201	.	.
Attrition		AIC	Akaike's Information Criterion	714.5284	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1567.215	.	.
Attrition		ASE	Average Squared Error	0.086743	0.114009	.
Attrition		MAX	Maximum Absolute Error	0.999834	0.998289	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.294522	0.337652	.
Attrition		SSE	Sum of Squared Errors	89.17217	117.4291	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.198152	.	.
Attrition		MSE	Mean Squared Error	0.142448	0.114009	.
Attrition		RFPE	Root Final Prediction Error	0.445142	.	.
Attrition		RMSE	Root Mean Squared Error	0.377422	0.337652	.
Attrition		AVERR	Average Error Function	0.304016	0.399786	.
Attrition		ERR	Error Function	312.5284	411.7791	.
Attrition		MISC	Misclassification Rate	0.110895	0.139806	.
Attrition		WRONG	Number of Wrong Classificati...	57	72	.

Outcome: ASE is 0.114009

d) 6 hidden units

Figure. 36 Property panel

Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	6
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Fu	Default
Hidden Layer Activation Func	Default
Hidden Bias	Yes
Target Layer Combination Fu	Default
Target Layer Activation Funct	Default

Architecture

Specifies which network architecture is used in constructing the network. The following are valid selections: generalized linear model, multilayer perceptron, ordinary radial basis function with equal widths, ordinary radial basis function with unequal widths, normalized radial basis

OK Cancel

Figure. 37 Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	
Attrition		DFE	Degrees of Freedom for Error	273	.	
Attrition		DFM	Model Degrees of Freedom	241	.	
Attrition		NW	Number of Estimated Weights	241	.	
Attrition		AIC	Akaike's Information Criterion	876.22	.	
Attrition		SBC	Schwarz's Bayesian Criterion	1898.596	.	
Attrition		ASE	Average Squared Error	0.111734	0.12583	
Attrition		MAX	Maximum Absolute Error	0.998358	0.984982	
Attrition		DIV	Divisor for ASE	1028	1030	
Attrition		NOBS	Sum of Frequencies	514	515	
Attrition		RASE	Root Average Squared Error	0.334267	0.354725	
Attrition		SSE	Sum of Squared Errors	114.8627	129.605	
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	
Attrition		FPE	Final Prediction Error	0.309008	.	
Attrition		MSE	Mean Squared Error	0.210371	0.12583	
Attrition		RFPE	Root Final Prediction Error	0.555885	.	
Attrition		RMSE	Root Mean Squared Error	0.458662	0.354725	
Attrition		AVERR	Average Error Function	0.383482	0.396648	
Attrition		ERR	Error Function	394.22	408.5479	
Attrition		MISC	Misclassification Rate	0.151751	0.172816	
Attrition		WRONG	Number of Wrong Classificati...	78	89	

Outcome: ASE is 0.12583

e) 7 hidden units

Figure. 38 Property panel

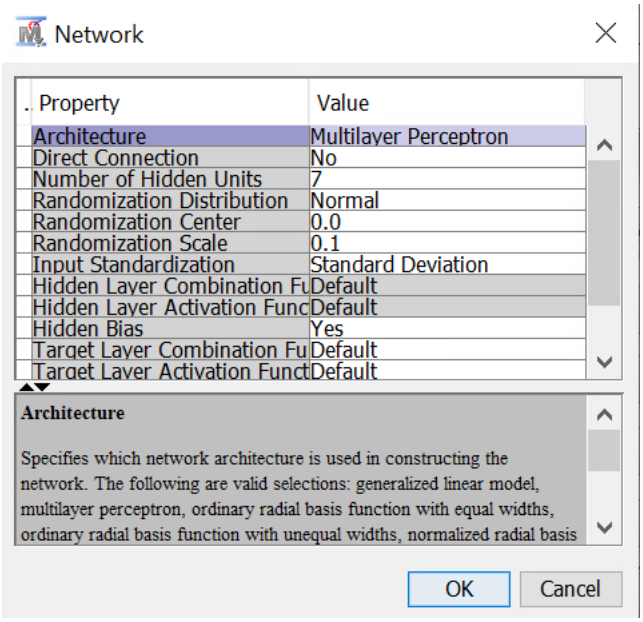


Figure. 39 Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	233	.	.
Attrition		DFM	Model Degrees of Freedom	281	.	.
Attrition		NW	Number of Estimated Weights	281	.	.
Attrition		AIC	Akaike's Information Criterion	905.86	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	2097.925	.	.
Attrition		ASE	Average Squared Error	0.099063	0.109919	.
Attrition		MAX	Maximum Absolute Error	0.981131	0.967328	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.314727	0.33154	.
Attrition		SSE	Sum of Squared Errors	101.8267	113.2161	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.337971	.	.
Attrition		MSE	Mean Squared Error	0.218512	0.109919	.
Attrition		RFPE	Root Final Prediction Error	0.581353	.	.
Attrition		RMSE	Root Mean Squared Error	0.467453	0.33154	.
Attrition		AVERR	Average Error Function	0.334494	0.35407	.
Attrition		ERR	Error Function	343.86	364.692	.
Attrition		MISC	Misclassification Rate	0.128405	0.151456	.
Attrition		WRONG	Number of Wrong Classificati...	66	78	.

Outcome: ASE is 0.109919

f) 8 hidden units

Figure. 40 Property panel

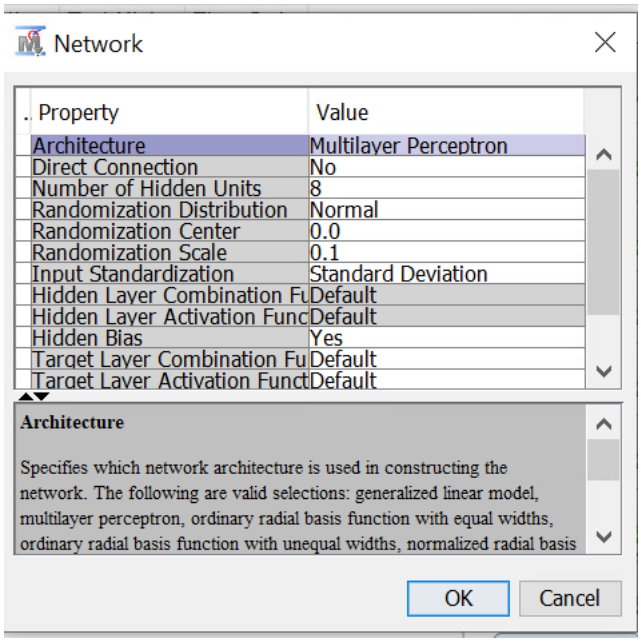
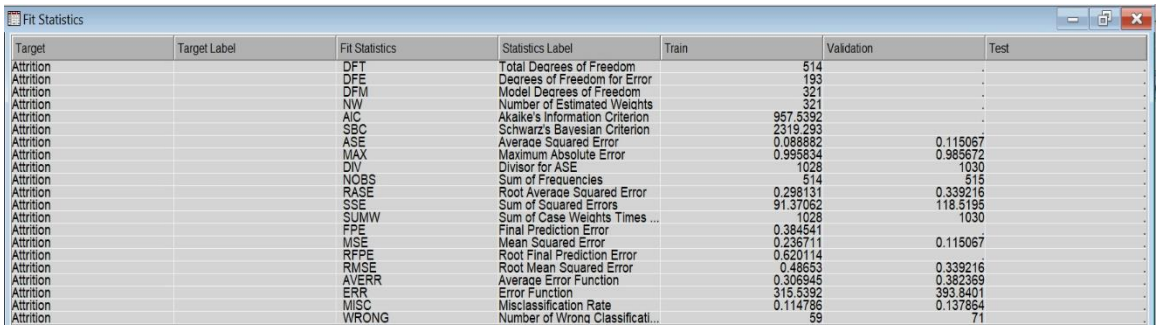


Figure. 41 Fit statistics

Outcome: ASE is 0.101124



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	193	.	.
Attrition		DFM	Model Degrees of Freedom	321	.	.
Attrition		NW	Number of Estimated Weights	321	.	.
Attrition		AIC	Akaike's Information Criterion	957.5392	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	2319.293	.	.
Attrition		ASE	Average Squared Error	0.088882	0.115067	.
Attrition		MAX	Maximum Absolute Error	0.995834	0.985672	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.298131	0.339216	.
Attrition		SSE	Sum of Squared Errors	91.37062	118.5195	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.384541	.	.
Attrition		MSE	Mean Squared Error	0.236711	0.115067	.
Attrition		RFPE	Root Final Prediction Error	0.620114	.	.
Attrition		RMSE	Root Mean Squared Error	0.48653	0.339216	.
Attrition		AVERR	Average Error Function	0.306945	0.382369	.
Attrition		ERR	Error Function	315.5392	393.8401	.
Attrition		MISC	Misclassification Rate	0.114786	0.137864	.
Attrition		WRONG	Number of Wrong Classificati...	59	71	.

Outcome: ASE is 0.115067

g) 9 hidden units

Figure. 42 Property panel

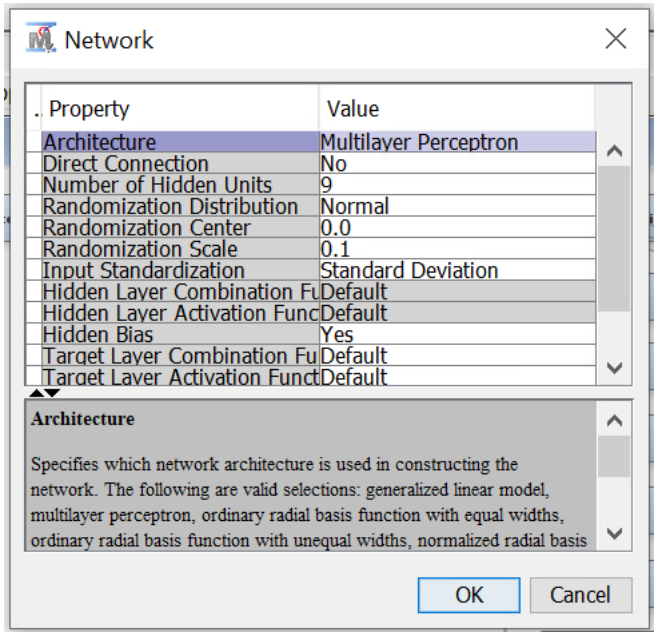
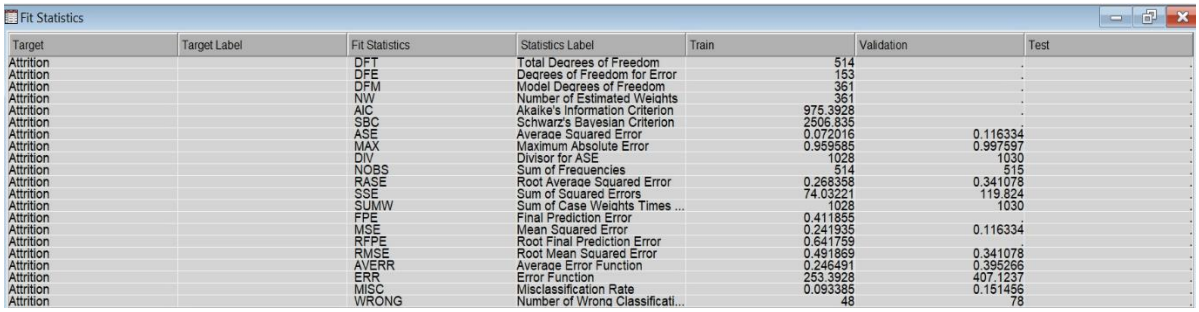


Figure.43 Fit statistics



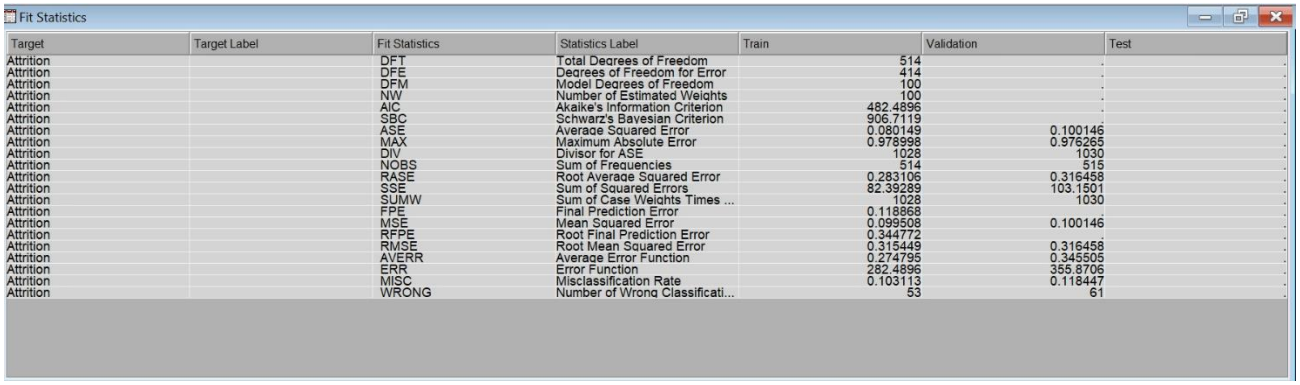
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	153	.	.
Attrition		DFM	Model Degrees of Freedom	361	.	.
Attrition		NW	Number of Estimated Weights	361	.	.
Attrition		AIC	Akaike's Information Criterion	975.3928	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	2506.835	.	.
Attrition		ASE	Average Squared Error	0.072016	0.116334	.
Attrition		MAX	Maximum Absolute Error	0.959585	0.997597	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.268358	0.341078	.
Attrition		SSE	Sum of Squared Errors	74.03221	119.824	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.411855	.	.
Attrition		MSE	Mean Squared Error	0.241935	0.116334	.
Attrition		RFPE	Root Final Prediction Error	0.641759	.	.
Attrition		RMSE	Root Mean Squared Error	0.491869	0.341078	.
Attrition		AVERR	Average Error Function	0.246491	0.395266	.
Attrition		ERR	Error Function	253.3928	407.1237	.
Attrition		MISC	Misclassification Rate	0.093385	0.151456	.
Attrition		WRONG	Number of Wrong Classificati...	48	78	.

Outcome: ASE is 0.116334

2. After transform variable

    a. *3 hidden units*

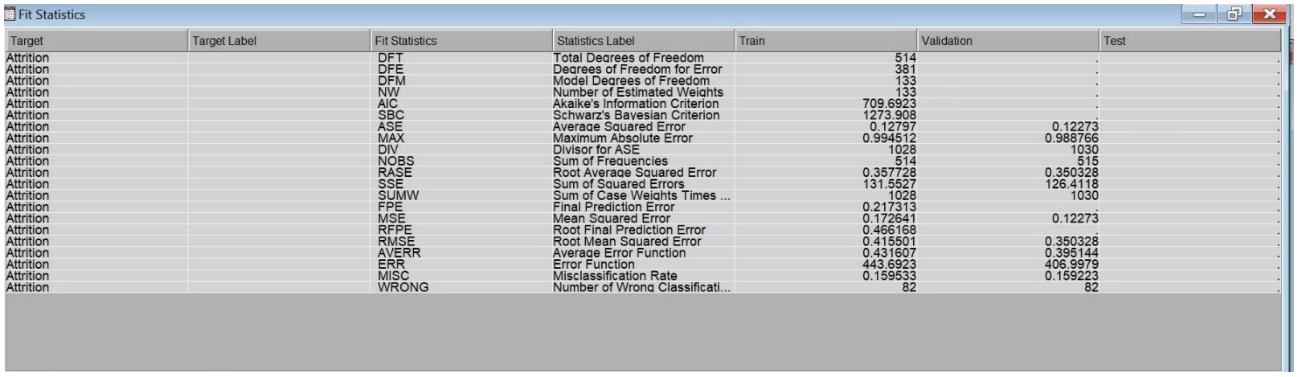


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	414	.	.
Attrition		DFM	Model Degrees of Freedom	100	.	.
Attrition		NW	Number of Estimated Weights	100	.	.
Attrition		AIC	Akaike's Information Criterion	482.4896	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	906.7119	.	.
Attrition		ASE	Average Squared Error	0.080149	0.100146	.
Attrition		MAX	Maximum Absolute Error	0.978998	0.976265	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.283106	0.316458	.
Attrition		SSE	Sum of Squared Errors	82.39289	103.1501	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.118868	.	.
Attrition		MSE	Mean Squared Error	0.099508	0.100146	.
Attrition		RFPE	Root Final Prediction Error	0.344772	.	.
Attrition		RMSE	Root Mean Squared Error	0.315449	0.316458	.
Attrition		AVERR	Average Error Function	0.274795	0.345505	.
Attrition		ERR	Error Function	282.4896	355.8706	.
Attrition		MISC	Misclassification Rate	0.103113	0.118447	.
Attrition		WRONG	Number of Wrong Classificati...	53	61	.

Outcome: ASE is 0.100146

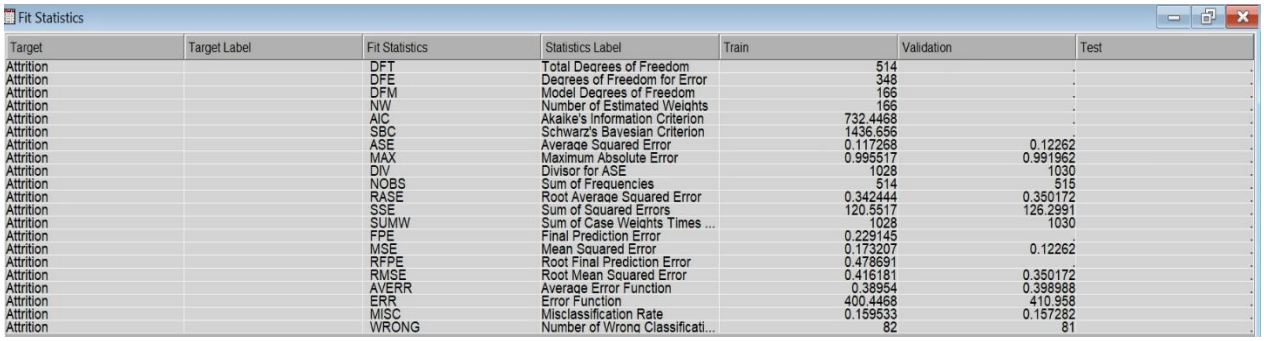
b. 4 hidden units


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	381	.	.
Attrition		DFM	Model Degrees of Freedom	133	.	.
Attrition		NW	Number of Estimated Weights	133	.	.
Attrition		AIC	Akaike's Information Criterion	709.6923	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1273.908	.	.
Attrition		ASE	Average Squared Error	0.12797	0.12273	.
Attrition		MAX	Maximum Absolute Error	0.994512	0.988766	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.357728	0.350328	.
Attrition		SSE	Sum of Squared Errors	131.5527	126.4118	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.217313	.	.
Attrition		MSE	Mean Squared Error	0.172641	0.12273	.
Attrition		RFPE	Root Final Prediction Error	0.466168	.	.
Attrition		RMSE	Root Mean Squared Error	0.415501	0.350328	.
Attrition		AVERR	Average Error Function	0.431607	0.395144	.
Attrition		ERR	Error Function	443.6923	406.9979	.
Attrition		MISC	Misclassification Rate	0.159533	0.159223	.
Attrition		WRONG	Number of Wrong Classificati...	82	82	.

Outcome: ASE is 0.12273

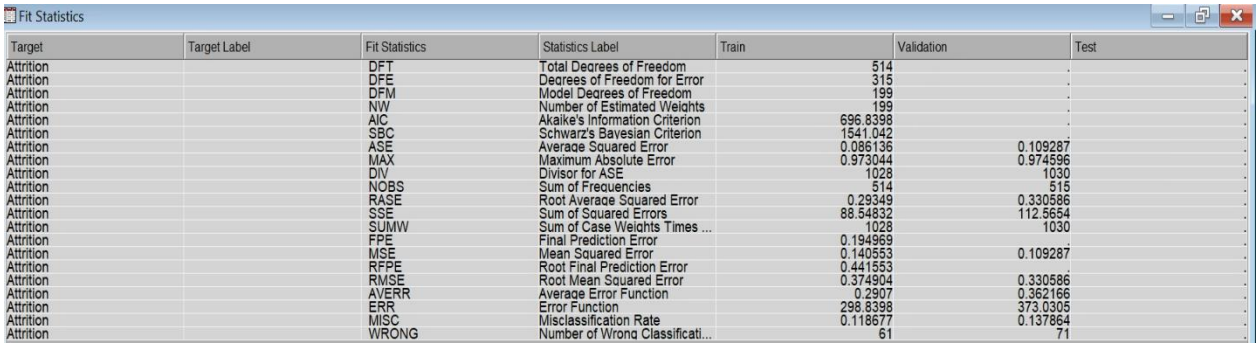
c. 5 hidden units


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	348	.	.
Attrition		DFM	Model Degrees of Freedom	166	.	.
Attrition		NW	Number of Estimated Weights	166	.	.
Attrition		AIC	Akaike's Information Criterion	732.4468	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1436.656	.	.
Attrition		ASE	Average Squared Error	0.117268	0.12262	.
Attrition		MAX	Maximum Absolute Error	0.995517	0.991962	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.342444	0.350172	.
Attrition		SSE	Sum of Squared Errors	120.5517	126.2991	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.229145	.	.
Attrition		MSE	Mean Squared Error	0.173207	0.12262	.
Attrition		RFPE	Root Final Prediction Error	0.478691	.	.
Attrition		RMSE	Root Mean Squared Error	0.416181	0.350172	.
Attrition		AVERR	Average Error Function	0.38954	0.398988	.
Attrition		ERR	Error Function	400.4468	410.958	.
Attrition		MISC	Misclassification Rate	0.159533	0.157282	.
Attrition		WRONG	Number of Wrong Classificati...	82	81	.

Outcome: ASE is 0.12262

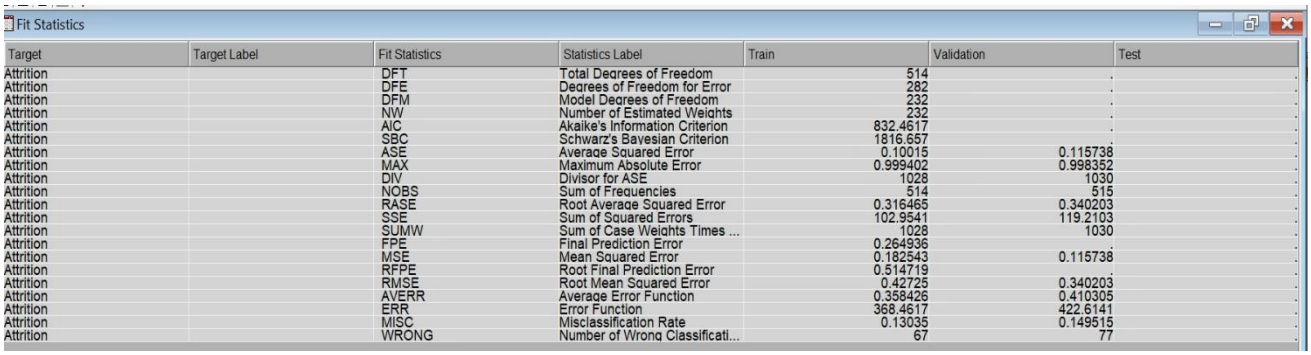
d. 6 hidden units


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	315	.	.
Attrition		DFM	Model Degrees of Freedom	199	.	.
Attrition		NW	Number of Estimated Weights	199	.	.
Attrition		AIC	Akaike's Information Criterion	696.8398	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1541.042	.	.
Attrition		ASE	Average Squared Error	0.086136	0.109287	.
Attrition		MAX	Maximum Absolute Error	0.973044	0.974596	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.29349	0.330586	.
Attrition		SSE	Sum of Squared Errors	88.54832	112.5654	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.194969	.	.
Attrition		MSE	Mean Squared Error	0.140553	0.109287	.
Attrition		RFPE	Root Final Prediction Error	0.441553	.	.
Attrition		RMSE	Root Mean Squared Error	0.374904	0.330586	.
Attrition		AVERR	Average Error Function	0.2907	0.362166	.
Attrition		ERR	Error Function	298.8398	373.0305	.
Attrition		MISC	Misclassification Rate	0.118677	0.137864	.
Attrition		WRONG	Number of Wrong Classificati...	61	71	.

Outcome: ASE is 0.109287

e. 7 hidden units

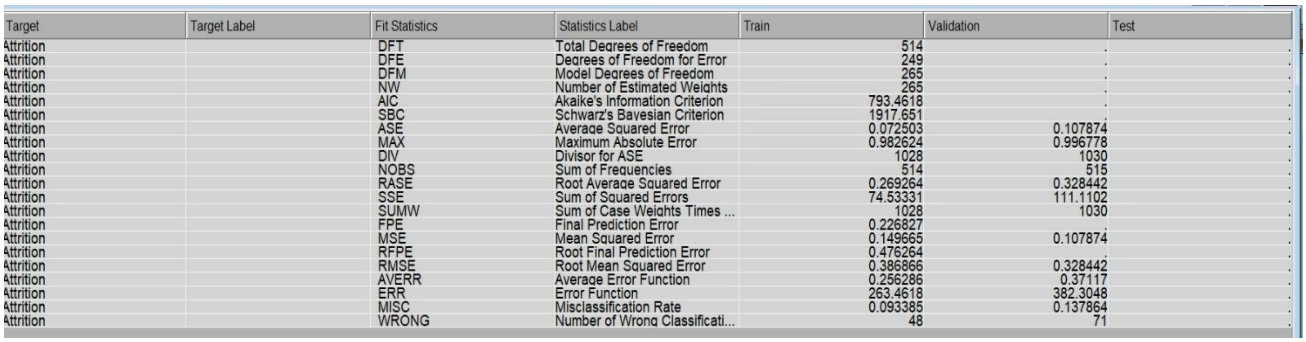


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	282	.	.
Attrition		DFM	Model Degrees of Freedom	232	.	.
Attrition		NW	Number of Estimated Weights	232	.	.
Attrition		AIC	Akaike's Information Criterion	832.4617	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1816.657	.	.
Attrition		ASE	Average Squared Error	0.10015	0.115738	.
Attrition		MAX	Maximum Absolute Error	0.999402	0.998352	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.316465	0.340203	.
Attrition		SSE	Sum of Squared Errors	102.9541	119.2103	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.264936	.	.
Attrition		MSE	Mean Squared Error	0.182543	0.115738	.
Attrition		RFPE	Root Final Prediction Error	0.514719	.	.
Attrition		RMSE	Root Mean Squared Error	0.42725	0.340203	.
Attrition		AVERR	Average Error Function	0.358426	0.410305	.
Attrition		ERR	Error Function	368.4617	422.6141	.
Attrition		MISC	Misclassification Rate	0.13035	0.149515	.
Attrition		WRONG	Number of Wrong Classificati...	67	77	.

Outcome: ASE is 0.115738

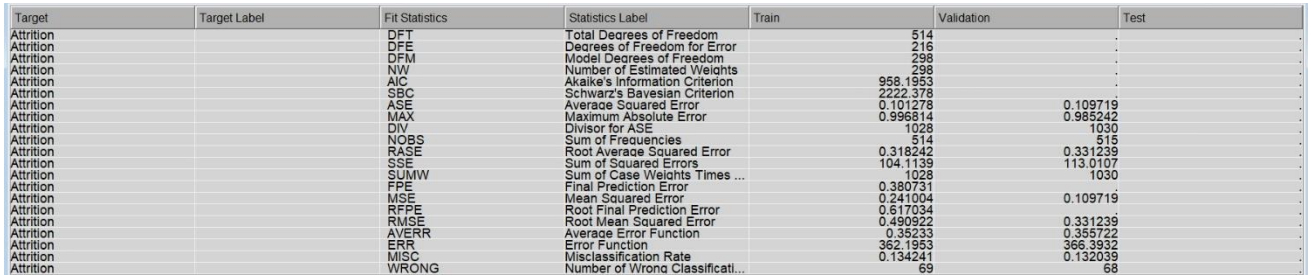
f. 8 hidden units



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	249	.	.
Attrition		DFM	Model Degrees of Freedom	265	.	.
Attrition		NW	Number of Estimated Weights	265	.	.
Attrition		AIC	Akaike's Information Criterion	793.4618	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	1917.651	.	.
Attrition		ASE	Average Squared Error	0.072503	0.107874	.
Attrition		MAX	Maximum Absolute Error	0.982624	0.996778	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.269264	0.328442	.
Attrition		SSE	Sum of Squared Errors	74.53331	111.1102	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.226827	.	.
Attrition		MSE	Mean Squared Error	0.149665	0.107874	.
Attrition		RFPE	Root Final Prediction Error	0.476264	.	.
Attrition		RMSE	Root Mean Squared Error	0.386866	0.328442	.
Attrition		AVERR	Average Error Function	0.256286	0.37117	.
Attrition		ERR	Error Function	263.4618	382.3048	.
Attrition		MISC	Misclassification Rate	0.093385	0.137864	.
Attrition		WRONG	Number of Wrong Classificati...	48	71	.

Outcome: ASE is 0.107874

g. 9 hidden units



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degrees of Freedom	514	.	.
Attrition		DFE	Degrees of Freedom for Error	216	.	.
Attrition		DFM	Model Degrees of Freedom	298	.	.
Attrition		NW	Number of Estimated Weights	298	.	.
Attrition		AIC	Akaike's Information Criterion	958.1953	.	.
Attrition		SBC	Schwarz's Bayesian Criterion	2222.378	.	.
Attrition		ASE	Average Squared Error	0.101278	0.109719	.
Attrition		MAX	Maximum Absolute Error	0.996814	0.985242	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequencies	514	515	.
Attrition		RASE	Root Average Squared Error	0.318242	0.331239	.
Attrition		SSE	Sum of Squared Errors	104.1139	113.0107	.
Attrition		SUMW	Sum of Case Weights Times ...	1028	1030	.
Attrition		FPE	Final Prediction Error	0.380731	.	.
Attrition		MSE	Mean Squared Error	0.241004	0.109719	.
Attrition		RFPE	Root Final Prediction Error	0.617034	.	.
Attrition		RMSE	Root Mean Squared Error	0.490922	0.331239	.
Attrition		AVERR	Average Error Function	0.35233	0.355722	.
Attrition		ERR	Error Function	362.1953	366.3932	.
Attrition		MISC	Misclassification Rate	0.134241	0.132039	.
Attrition		WRONG	Number of Wrong Classificati...	69	68	.

Outcome: ASE is 0.109719

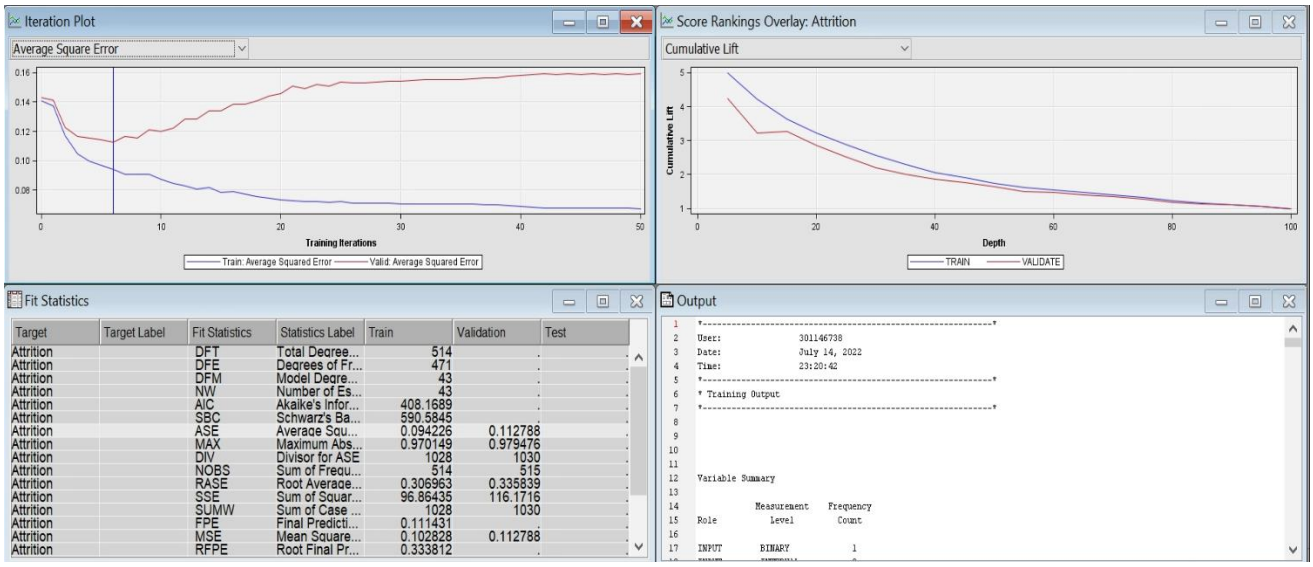
In short, we have a quick summary for ASE from 3 to 9 units as follows:

Table.3 ASE

Hidden units	ASE: impute	ASE: after transform variables
3	0.114755	0.100146
4	0.10728	0.12273
5	0.114009	0.12262
6	0.12583	0.109287
7	0.109919	0.115738
8	0.115067	0.107874
9	0.116334	0.109719

It can be seen that ASE of 3 hidden units after transform variables get the least error compared to other neural networks (after impute node and after transform variables node) so it is considered to be the best one.

On the other hand, we connect the best neural network which is 3 hidden units to the best regression model which is stepwise regression because we want to try the neural network with reduced variables. However, as a result, the 3-hidden unit neural network is still the best one. The result could be seen as below:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition		DFT	Total Degree...	514	.	.
Attrition		DFE	Degrees of Fr...	471	.	.
Attrition		DFM	Model Degre...	43	.	.
Attrition		NW	Number of Es...	43	.	.
Attrition		AIC	Akaike's Infor...	408.1689	.	.
Attrition		SBC	Schwarz's Ba...	590.5845	.	.
Attrition		ASE	Average Squ...	0.094226	0.112788	.
Attrition		MAX	Maximum Abs...	0.970149	0.979476	.
Attrition		DIV	Divisor for ASE	1028	1030	.
Attrition		NOBS	Sum of Frequ...	514	515	.
Attrition		RASE	Root Average...	0.306963	0.335839	.
Attrition		SSE	Sum of Squar...	96.86435	116.1716	.
Attrition		SUMW	Sum of Case ...	1028	1030	.
Attrition		FPE	Final Predicti...	0.111431	.	.
Attrition		MSE	Mean Square...	0.102828	0.112788	.
Attrition		RFPE	Root Final Pr...	0.333812	.	.

```
*------------------------------------------------------------*
User:                  301146738
Date:                  July 14, 2022
Time:                  23:20:42
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

                    Measurement    Frequency
Role                   Level          Count

INPUT               BINARY               1
```

D. MODEL COMPARISON

Among those 3 models, we need to find out which model is the best. As a result, based on Average Square Error criteria, neural network with 3 hidden units is the best model.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Neural	NN 3HU 50I	0.10015	0.08015	0.10311	0.11845
	Neural2	NN 4HU 50I	0.10728	0.08567	0.11673	0.13981
	Neural8	NN 8HU 50I	0.10787	0.07250	0.09339	0.13786
	Reg	Stepwise Regression	0.10893	0.09864	0.12451	0.15146
	Reg5	Forward Regression	0.10893	0.09864	0.12451	0.15146
	Neural5	NN 6HU 50I	0.10929	0.08614	0.11868	0.13786
	Reg4	Backward Regression	0.10930	0.09101	0.11479	0.14369
	Neural9	NN 9HU 50I	0.10972	0.10128	0.13424	0.13204
	Neural13	NN 7HU 50I	0.10992	0.09905	0.12840	0.15146
	Reg6	Full Regression	0.11063	0.09028	0.11284	0.14563
	Neural10	NN 3HU 50I	0.11279	0.09423	0.12451	0.15340
	Neural11	NN 5HU 50I	0.11401	0.08674	0.11089	0.13981
	Neural7	NN 3HU 50I	0.11475	0.11341	0.14591	0.15728
	Neural14	NN 8HU 50I	0.11507	0.08888	0.11479	0.13786
	Neural6	NN 7HU 50I	0.11574	0.10015	0.13035	0.14951
	Neural15	NN 9HU 50I	0.11633	0.07202	0.09339	0.15146
	Tree2	ASE Tree	0.12237	0.10510	0.13424	0.16699
	Neural4	NN 5HU 50I	0.12262	0.11727	0.15953	0.15728
	Neural3	NN 4HU 50I	0.12273	0.12797	0.15953	0.15922
	Tree	MISC Tree	0.12395	0.11331	0.14202	0.16505
	Neural12	NN 6HU 50I	0.12583	0.11173	0.15175	0.17282
	Tree3	Maximal Tree	0.13563	0.08926	0.12257	0.17282

Figure. 44 Fit statistics

We also looked at ROC chart of all the models in model comparison.

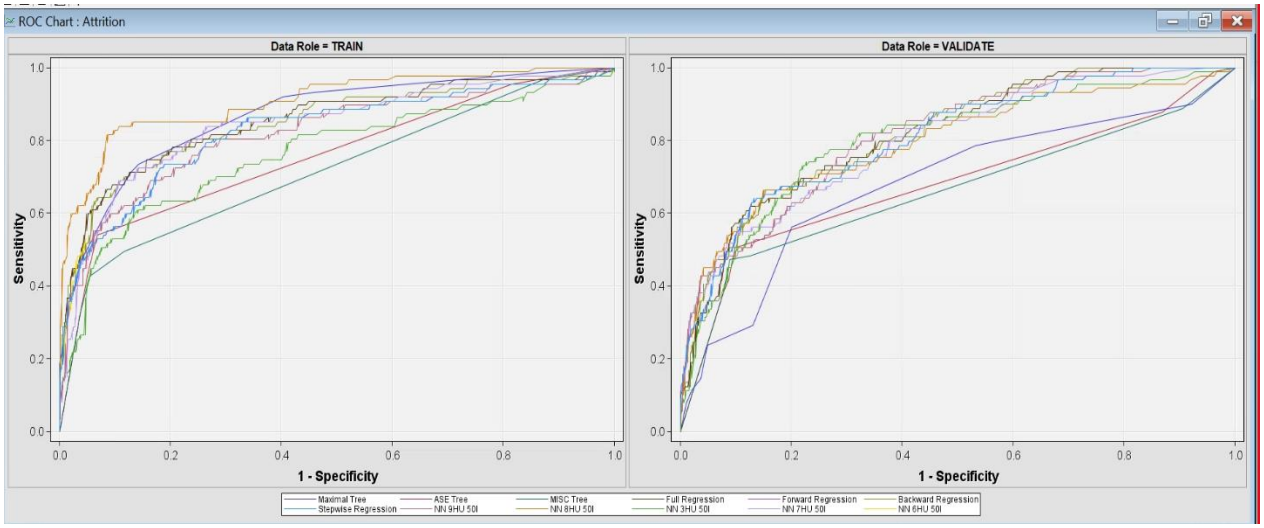


Figure. 45 Results

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	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	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights
Y	Neural	Neural	NN 3H...	Attrition	0.1001...		0.1001...	514	0.1031...	0.9789...	82.392...	0.0801...	0.2831...	1028	514	515	0.1184...	0.9762...	103.15...	0.3164...	1030	414	100	100
	Neural2	Neural2	NN 4H...	Attrition	0.10728		0.10728	514	0.1167...	0.9956...	88.072...	0.0856...	0.2927	1028	514	515	0.1398...	0.9876...	110.49...	0.3275...	1030	353	161	161
	Neural8	Neural8	NN 8H...	Attrition	0.1078...		0.1078...	514	0.0933...	0.9826...	74.533...	0.0725...	0.2692...	1028	514	515	0.1378...	0.9967...	111.11...	0.3284...	1030	249	265	265
	Reg	Reg	Stepwi...	Attrition	0.1089...		0.1089...	514	0.1245...	0.9965...	101.405	0.0986...	0.3140...	1028	514	515	0.1514...	0.9800...	112.19...	0.3300...	1030	501	13	13
	Reg5	Reg5	Forwar...	Attrition	0.1089...		0.1089...	514	0.1245...	0.9965...	101.405	0.0986...	0.3140...	1028	514	515	0.1514...	0.9800...	112.19...	0.3300...	1030	501	13	13
	Neural5	Neural5	NN 6H...	Attrition	0.1092...		0.1092...	514	0.1186...	0.9730...	88.548...	0.0861...	0.29349	1028	514	515	0.1378...	0.9745...	112.56...	0.3305...	1030	315	199	199
	Reg4	Reg4	Backw...	Attrition	0.1093...		0.1093...	514	0.1147...	0.9990...	93.553...	0.0910...	0.3016...	1028	514	515	0.1436...	0.9761...	112.58...	0.3306...	1030	488	26	26
	Neural9	Neural9	NN 9H...	Attrition	0.1097...		0.1097...	514	0.1342...	0.9968...	104.11...	0.1012...	0.3182...	1028	514	515	0.1320...	0.9852...	113.01...	0.3312...	1030	216	298	298
	Neural13	Neural13	NN 7H...	Attrition	0.1099...		0.1099...	514	0.1284...	0.9811...	101.82...	0.0990...	0.3147...	1028	514	515	0.1514...	0.9673...	113.21...	0.33154	1030	233	281	281
	Reg6	Reg6	Full Re...	Attrition	0.1106...		0.1106...	514	0.11284	0.99919	92.804...	0.0902...	0.30046	1028	514	515	0.1456...	0.9811...	113.95...	0.3326...	1030	482	32	32
	Neural10	Neural10	NN 3H...	Attrition	0.1127...		0.1127...	514	0.1245...	0.9701...	96.864...	0.0942...	0.3069...	1028	514	515	0.1533...	0.9794...	116.17...	0.3358...	1030	471	43	43
	Neural11	Neural11	NN 5H...	Attrition	0.1140...		0.1140...	514	0.1108...	0.9998...	89.172...	0.0867...	0.2945...	1028	514	515	0.1398...	0.9982...	117.42...	0.3376...	1030	313	201	201
	Neural7	Neural7	NN 3H...	Attrition	0.1147...		0.1147...	514	0.1459...	0.9960...	116.58...	0.1134...	0.3367...	1028	514	515	0.1572...	0.9949...	118.19...	0.3387...	1030	393	121	121
	Neural14	Neural14	NN 8H...	Attrition	0.1150...		0.1150...	514	0.1147...	0.9958...	91.370...	0.0888...	0.2981...	1028	514	515	0.1378...	0.9856...	118.51...	0.3392...	1030	193	321	321
	Neural6	Neural6	NN 7H...	Attrition	0.1157...		0.1157...	514	0.13035	0.9994...	102.95...	0.10015	0.3164...	1028	514	515	0.1495...	0.9983...	119.21...	0.3402...	1030	282	232	232
	Neural15	Neural15	NN 9H...	Attrition	0.1163...		0.1163...	514	0.0933...	0.9595...	74.032...	0.0720...	0.2683...	1028	514	515	0.1514...	0.9975...	119.824	0.3410...	1030	153	361	361
	Tree2	Tree2	ASE Tr...	Attrition	0.1223...		0.1223...	514	0.1342...	0.96	108.04...	0.1051...	0.3241...	1028	514	515	0.16699	0.96	126.04...	0.3498...	1030	.	.	.
	Neural4	Neural4	NN 5H...	Attrition	0.12262		0.12262	514	0.1595...	0.9955...	120.55...	0.1172...	0.3424...	1028	514	515	0.1572...	0.9919...	126.29...	0.3501...	1030	348	166	166
	Neural3	Neural3	NN 4H...	Attrition	0.12273		0.12273	514	0.1595...	0.9945...	131.55...	0.12797	0.3577...	1028	514	515	0.1592...	0.9887...	126.41...	0.3503...	1030	381	133	133
	Tree	Tree	MISC ...	Attrition	0.1239...		0.1239...	514	0.1420...	0.9482...	116.48...	0.1133...	0.3366...	1028	514	515	0.1650...	0.9482...	127.66...	0.3520...	1030	.	.	.
	Neural12	Neural12	NN 6H...	Attrition	0.12583		0.12583	514	0.1517...	0.9983...	114.86...	0.1117...	0.3342...	1028	514	515	0.1728...	0.9849...	129.605	0.3547...	1030	273	241	241
	Tree3	Tree3	Maxim...	Attrition	0.1356...		0.1356...	514	0.1225...	0.98	91.756...	0.0892...	0.2987...	1028	514	515	0.1728...	1	139.70...	0.3682...	1030	.	.	.

IV. Recommendation

After analyzing different regression models and neural networks, we come the conclustion that 3-hidden unit neural network model is the best. However, based on the results of decision tree and regression models, ASE tree and stepwise regression are also the good ones. There are several factors for the company to determine to keep the talented people.

We suggest the company reduce the amount of working overtime for the employees so that they are likely to show their work more efficiently and have more time for their personal life. Thus, the employees are willing to stay at the company with a balanced schedule. Moreover, another factor that should be considered is adjusting the salary of the low-level employees where they are encouraged to continue contributing their work towards the company even with their small tasks.

The other factor that we would like to mention here is environmental satisfaction among the working environment and the employees. Since they spend the entire day at work,

they prefer being in the sparkling and well-maintenance office, diverse company culture, remarkable policies for vacation, paternity leave and paid time off.

V. Conclusion

In conclusion, we performed the predictive models to see which factors could affect attrition, such as Decision tree, Regression models, and Neural Networks. After analyzing all the models, Neural Network with 3 hidden units after Transform variables node was the best model. We also made recommendations for the company to retain good employees by reducing overtime work, considering the salary adjustment, and improving work environment.

Appendix: Employee Attrition Diagram