Predictive Analysis: Factors Affecting IBM Employee Attrition

Mitzie Irene P. Conchada, 301258577

Mary Claire C. Doña, 301323966

Karen Ann H. Francisco, 301238093

Jason S. Yap, 301293413

Centennial College

BA 706–Applied Analytic Modeling - 001

David Parent

December 15, 2023

Introduction

Employee attrition can be an issue for companies since this translates to higher cost when it has to hire and train new employees. Companies take on different strategies to keep their employees, especially the talented ones and those who can potentially contribute to the growth of the company. Another important aspect of employee attrition is the impact on employee's morale, productivity and overall organization culture (Conchada et al., 2023). This study tries to look into the various factors affecting the employee attrition, specifically at IBM.

Business Problem Statement and Objectives

Our group chose the IBM database from Kaggle. The IBM database (2023) has 1,470 observations, 12 inputs/predictors/independent variables, and 1 target variable. Our study is an extension of our research paper in our BA701 Analytic Lifecycle Management class, where the authors Conchada, Doña and Francisco (2023) performed a simple descriptive study by applying the different theories and methods on analytic lifecycle management.

For this study, we focused on the same problem, but this time we employ the various methodologies in analytic modelling. The research question of our study delves on: What are the factors affecting employee attrition at IBM in order to answer the research question, our study tried to achieve the following objectives:

1. Determine the factors affecting employee attrition at IBM using predictive modelling that will represent the relationship between inputs and the target variable *Attrition*. Our study uses the following predictive models:
    a. Decision tree
    b. Regression analysis
    c. Neural networks
2. Choose the best model that yield favorable results based on model assessment comparison.
3. Recommend steps related to employee attrition and points for future research

Data Set-up and Exploration

The IBM dataset is hypothetical and was derived from an open-source website, Kaggle. As mentioned earlier, the database has 1,470 observations, 12 inputs/predictors/independent variables, and 1 target variable. The table below shows the list and description of inputs and target variables used in the study. We identified *attrition* as our target variable, thus we set this to be binary. Moreover, we set the following variables as nominal as they are categorical in nature: *education, education field, marital status, department, environment satisfaction, job satisfaction* and *work-life-balance*. The rest of the variables are interval: *age, distance, monthly income, years at company*, and *number of companies worked*.

Variables	Description	Level
Attrition (Target)	1: Employee resigned (yes) 0: Employee stayed (no)	Binary
Age	Age	Interval
Distance from home	Distance from home from work in kilometers	Interval
Education	1: Below College 2: College 3: Bachelor 4: Master 5: Doctor	Nominal
Education field	Life Sciences Medical Marketing Technical Degree Human Resources Other	Nominal
Marital status	Married Single Divorced	Nominal
Monthly income	Monthly income	Interval
Years at company	Employees' total working years at IBM	Interval
Number of companies worked	Number of companies worked at prior to IBM	Interval
Department	Research and development Sales Human Resources	Nominal
Environment satisfaction	1: Low 2: Medium 3: High	Nominal

	4: Very High	
Job satisfaction	1: Low 2: Medium 3: High 4: Very High	Nominal
Work-life-balance	1: Bad 2: Good 3: Better 4: Best	Nominal

The figure below portrays the data based on its descriptive statistics. The age of the IBM employees ranges from 18 to 60 years old, and the average age is 37 years old. The average distance from home is 9 kilometers. There are more employees who are college graduates, and are in the field of Life Sciences. The average number of years at IBM is 7 years, while the average number of companies worked at before IBM is 3. The average monthly income is USD 6,503. With regards to the survey questions, we got the following results: the average rating for *environment satisfaction* is 2.7 or 3 (high); the average rating for *job satisfaction* is 2.7 or 3 (high); and the average rating for *work life balance* is 2.7 or 3 (better).

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum
Age	INPUT	36.92381	9.135373	1470	0	18	36	60
DistanceFromHome	INPUT	9.192517	8.106864	1470	0	1	7	29
Education	INPUT	2.912925	1.024165	1470	0	1	3	5
EnvironmentSatisfaction	INPUT	2.721769	1.093082	1470	0	1	3	4
JobSatisfaction	INPUT	2.728571	1.102846	1470	0	1	3	4
MonthlyIncome	INPUT	6502.931	4707.957	1470	0	1009	4908	19999
NumCompaniesWorked	INPUT	2.693197	2.498009	1470	0	0	2	9
WorkLifeBalance	INPUT	2.761224	0.706476	1470	0	1	3	4
YearsAtCompany	INPUT	7.008163	6.126525	1470	0	0	5	40

In the figure below, we have the sample statistics for all variables including non-numeric values. Most of the employees (84%) did not resign, are in the *research and development* department (65%), are married (46%), and their education background is in *life sciences* (41%).

Sample Statistics

Obs #	Variable Name	Type	Label	Percen	Minimum	Maximum	Mean	Numbe	Mode	Mode ▼
2	Department	CLASS		0	.	.		3	65.37415	RESEARCH & DEVELOPMENT
1	Attrition	CLASS		0	.	.		2	83.87755	NO
4	MaritalStatus	CLASS		0	.	.		3	45.78231	MARRIED
3	EducationField	CLASS		0	.	.		6	41.22449	LIFE SCIENCES
5	Age	VAR		0	18	60	36.92381	.	.	
6	DistanceFromHome	VAR		0	1	29	9.192517	.	.	
7	Education	VAR		0	1	5	2.912925	.	.	
8	EnvironmentSatisfaction	VAR		0	1	4	2.721769	.	.	
9	JobSatisfaction	VAR		0	1	4	2.728571	.	.	
10	MonthlyIncome	VAR		0	1009	19999	6502.931	.	.	
11	NumCompaniesWorked	VAR		0	0	9	2.693197	.	.	
12	WorkLifeBalance	VAR		0	1	4	2.761224	.	.	
13	YearsAtCompany	VAR		0	0	40	7.008163	.	.	

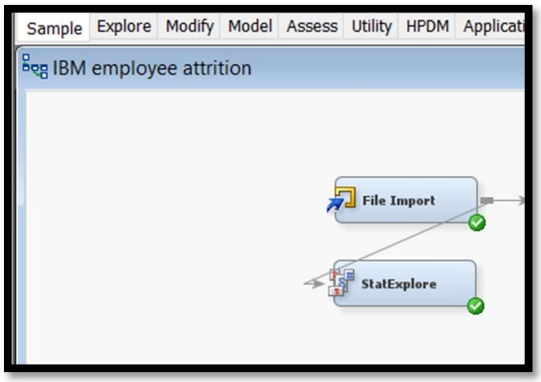
Predictive Models

Decision Tree Model

In constructing a model, our study followed these modelling essentials. First of all, modelling helps in predicting new cases that leads to a decision, rank or estimate. In our case, modelling helps predict the factors that will most likely have an effect on employee attrition, and to what extent. Second, modelling helps us to select useful inputs. Based on our dataset, all variables were deemed to be important in relation to employee attrition. There are no redundant and irrelevant inputs. Third, modelling considers the task of optimizing model complexity. We do not want our model to be insufficiently complex as it might not be flexible enough, which can lead to underfitting and eventually high bias. On the other hand, we do not want it to be too complex and too flexible as it can lead to overfitting, which has a high variance. Our study employed the steps in predictive modelling in order to achieve the objectives.

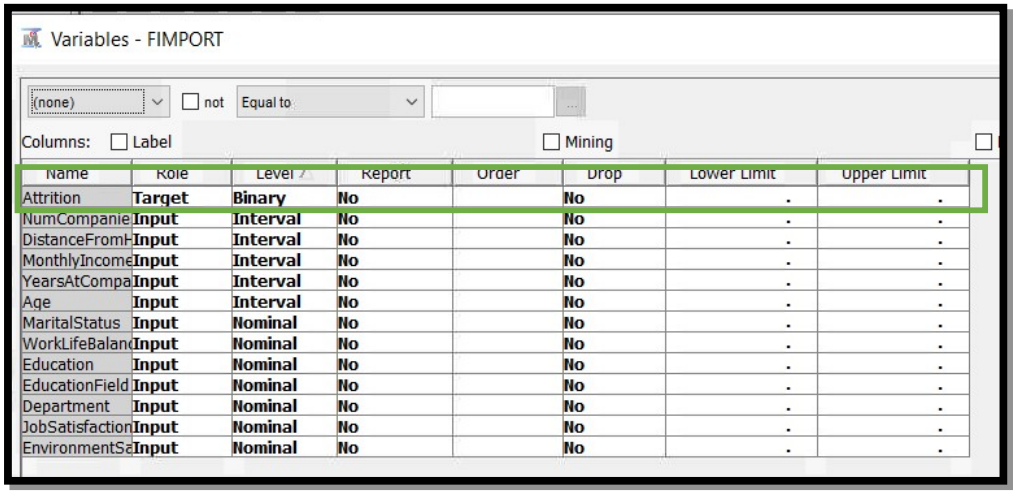
A. Data Preparation for Decision Trees

After downloading the IBM Employee Attrition dataset which was in csv file, we created a new project in SAS Enterprise Miner and uploaded the csv file. We created a new diagram name IBM Employee Attrition and added a Stat Explore to check the variables.



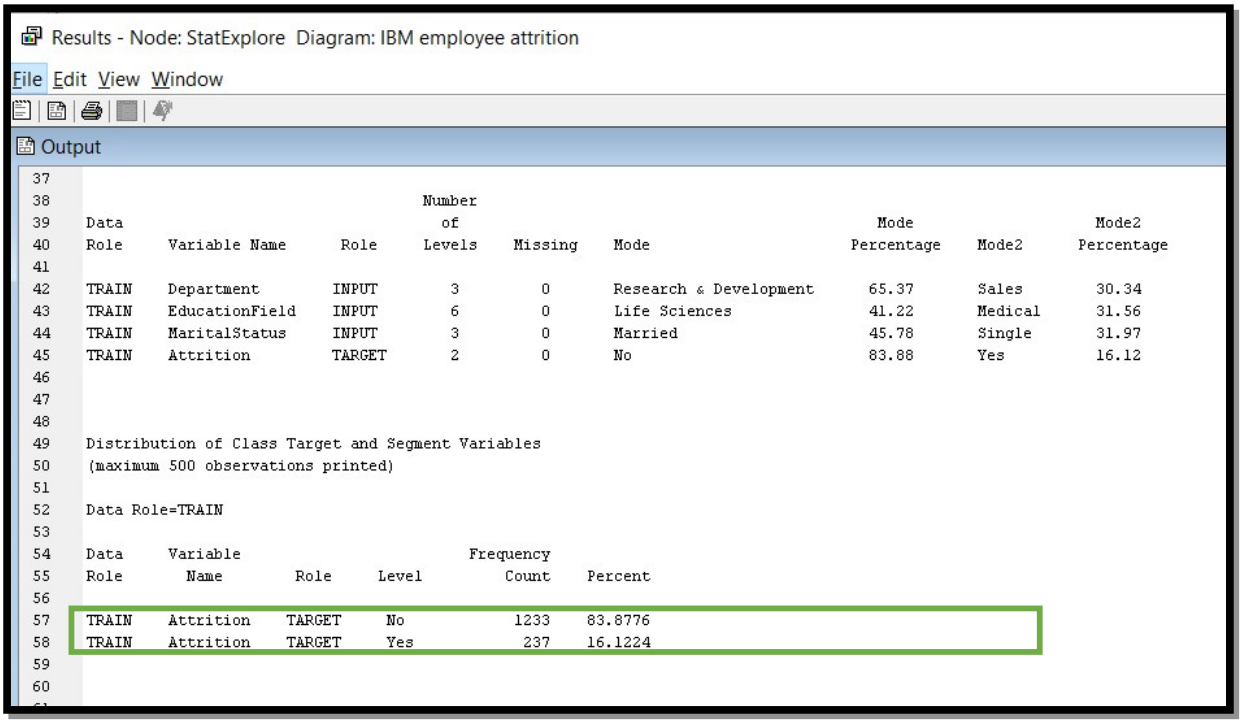
In setting the model roles for the analysis variables, we assigned *attrition* as our Target Variable and set it to a binary level. We changed the variables *education, work life balance, job satisfaction and environment* from interval to nominal level since the aforementioned are

categorical variables. This will help us better understand and interpret our results. After checking all the variables in the dataset, we deemed them as important and included them in the model as they may have a significant impact on our target variable.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Attrition	Target	Binary	No		No	.	.
NumCompanie	Input	Interval	No		No	.	.
DistanceFromH	Input	Interval	No		No	.	.
MonthlyIncome	Input	Interval	No		No	.	.
YearsAtCompa	Input	Interval	No		No	.	.
Age	Input	Interval	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
WorkLifeBalanc	Input	Nominal	No		No	.	.
Education	Input	Nominal	No		No	.	.
EducationField	Input	Nominal	No		No	.	.
Department	Input	Nominal	No		No	.	.
JobSatisfaction	Input	Nominal	No		No	.	.
EnvironmentSa	Input	Nominal	No		No	.	.

Next, we examined the distribution of *Attrition*. Based on the Stat Explore, the percentage of those who resigned from IBM is 16.1%, while the percentage for those who did not resign is 83.9%.



Results - Node: StatExplore Diagram: IBM employee attrition

File Edit View Window

Output

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Department	INPUT	3	0	Research & Development	65.37	Sales	30.34
TRAIN	EducationField	INPUT	6	0	Life Sciences	41.22	Medical	31.56
TRAIN	MaritalStatus	INPUT	3	0	Married	45.78	Single	31.97
TRAIN	Attrition	TARGET	2	0	No	83.88	Yes	16.12

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

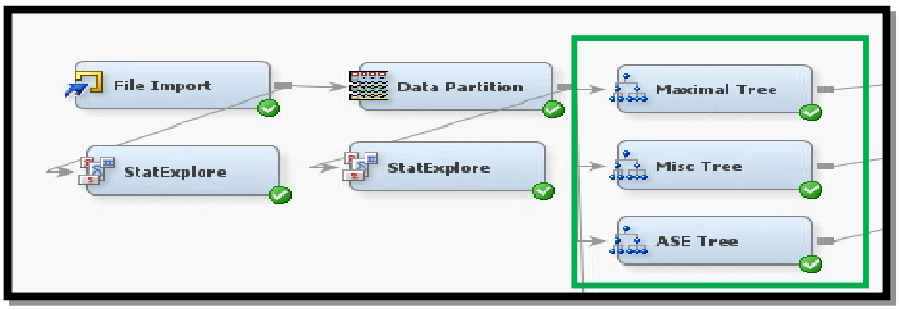
Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Attrition	TARGET	No	1233	83.8776
TRAIN	Attrition	TARGET	Yes	237	16.1224

B. Prediction Formula

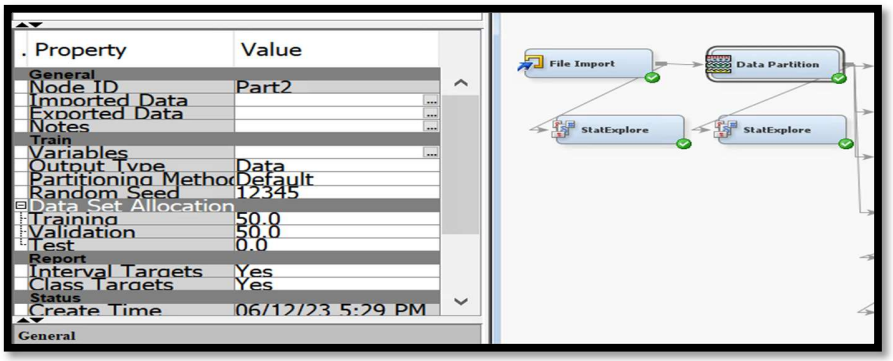
The first predictive model is the Decision Tree. Decision Trees help in identifying significant relationships between input and target variables in a data set. Several types of Trees were called in to observe different insights with the following details below.

Decision Trees			
	Maximal Tree	**Misclassification Tree**	**ASE Tree**
Method	Largest	Assessment	Assessment
Assessment Measure	Decision	Misclassification	Average Score Error

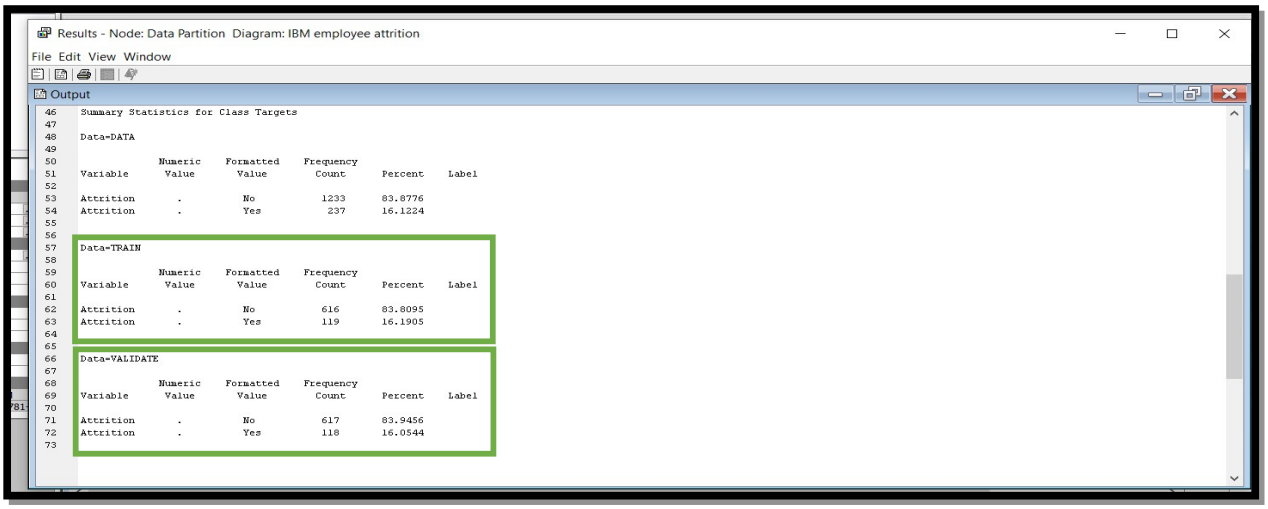


C. Data Partition

We added a data partition node to the File Import/Data Source node. We split the dataset for training and validation allocation to 50% and 50% each. The training data is used 'for fitting the model', while the validation data is used 'for empirical validation' (Parent, 2023). This is very important because we want our model to have the right amount of flexibility (not overfit-too specific or underfit-not complicated enough) to give us the best generalization from the results. With smaller raw data sets, model stability can become an important issue. In this case, increasing the number of cases devoted to the training partition can be a reasonable course of action. But since our model is stable, given its results, we think there is no need to increase the number of cases for the training partition.



After setting the training and validation set to 50-50%, the data was divided accordingly:

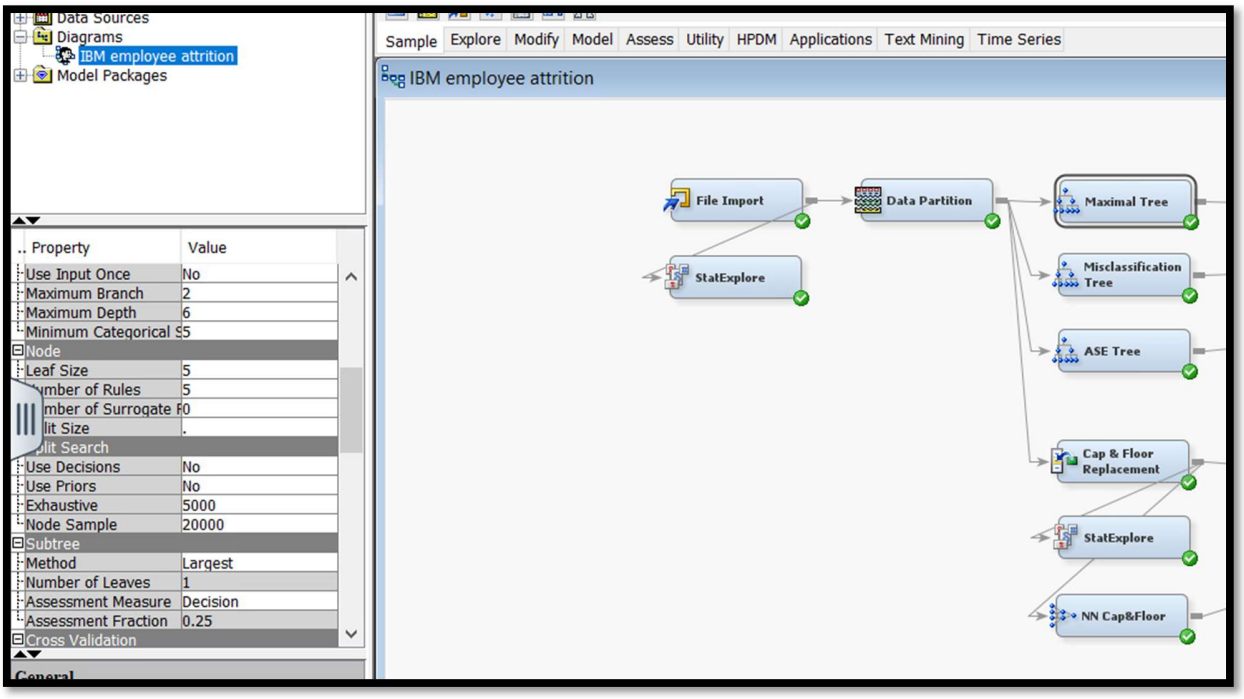


D. Decision Tree Results

1. Maximal Tree

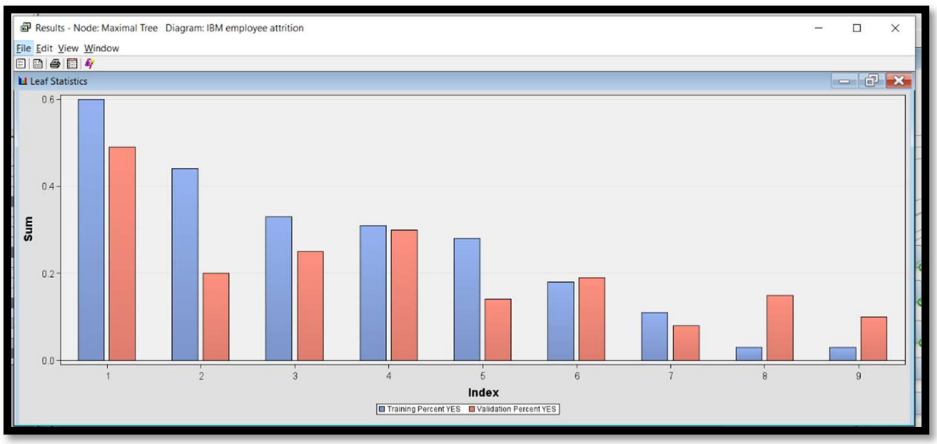
The first step to predictive modeling is Decision Trees. It addresses the modelling essential steps mentioned earlier. Prediction rules were employed, which includes a *split-search* algorithm to identify the significant input variables. If the model is complex, the Decision Tree model addresses this through *pruning*.

We added a Decision Tree node and call it the Maximal Tree with the following properties: Subtree Method – Largest, Assessment Measure - Decision. The Largest option provided an independent way to generate the Maximal Tress.

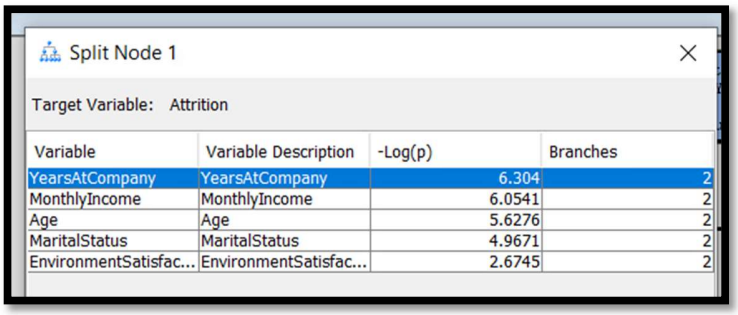


We got the following results:

- Number of leaves: 9

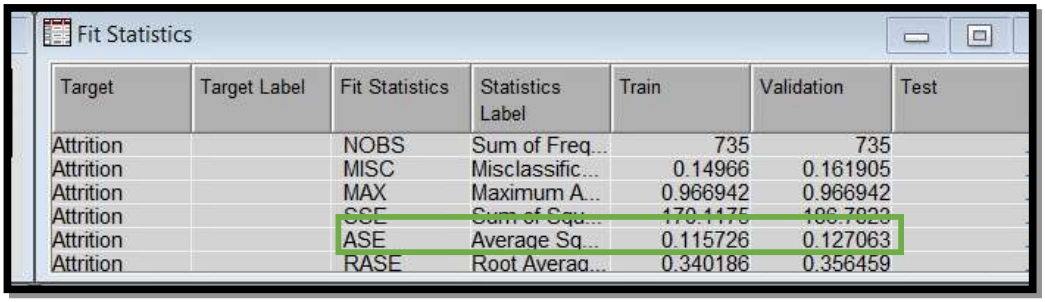


- Variables with the highest Logworth: *years at company, monthly income, age, marital status, environment satisfaction*. The Logworth value determined the split: the highest logworth is the best split. In this case, it is the variable *years at company.*

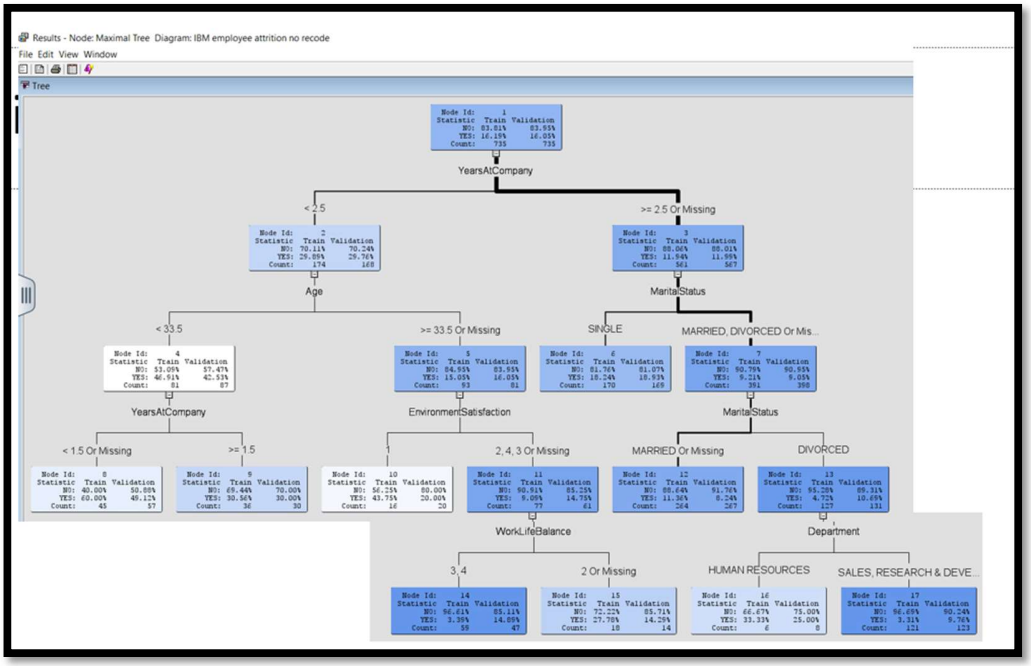


The variable that was used for first split: Years at company (less than 2.5 years and more than, equal to 2.5 years)

- Fit Statistics revealed an Average Squared Error (ASE) value of **0.127063**. The ASE to the average squared difference between the predicted/estimated and actual value. The smaller the value, the better since it signifies that the actual values are very close to the predicted value.

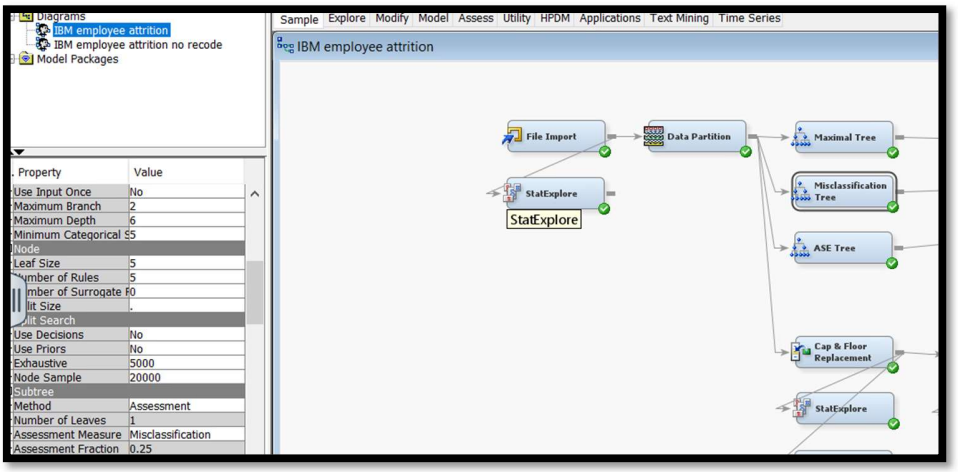


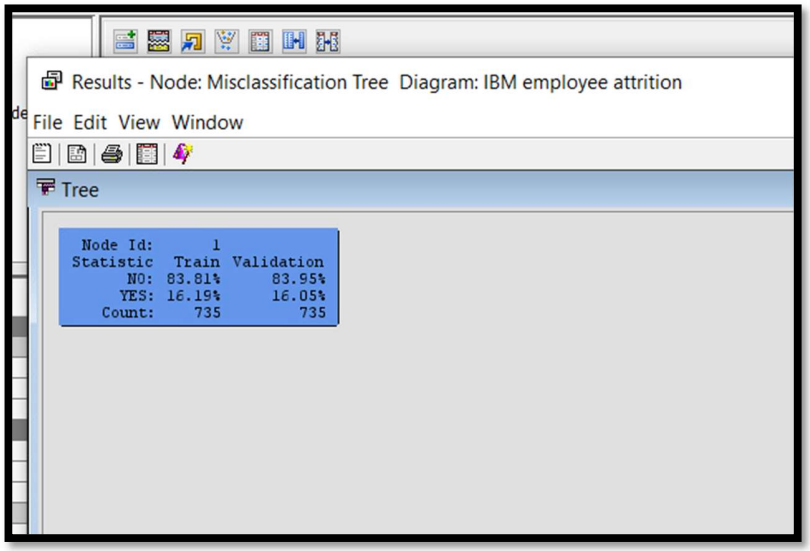
The results from the Maximal Tree reveal the following:



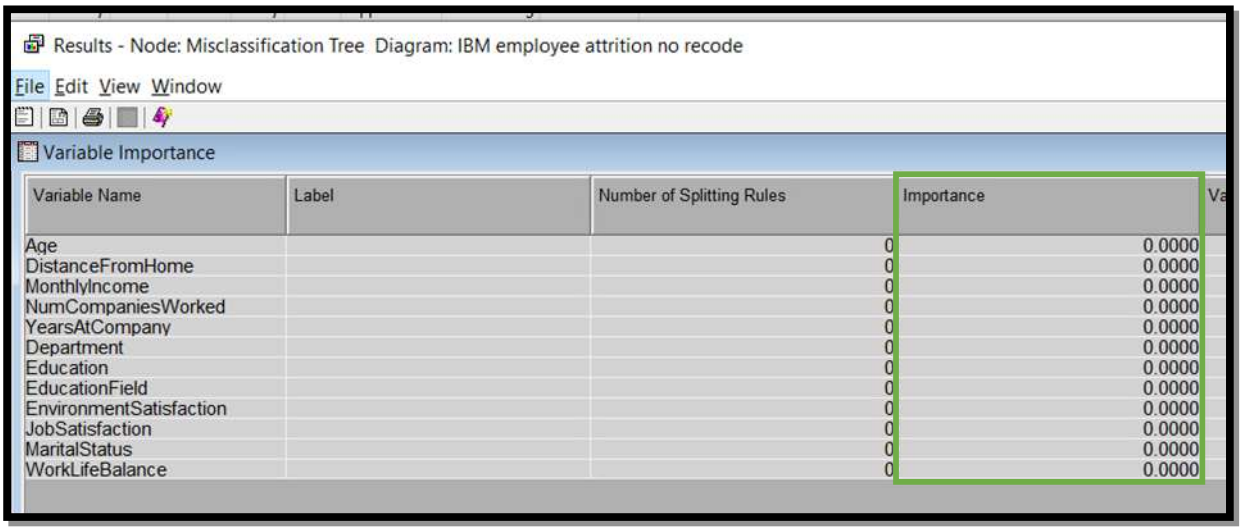
2. **Misclassification Tree**

For the Misclassification Tree model, we changed the subtree property for Method as Assessment and of Assessment Measure as Misclassification. We kept the other properties at default (i.e. splitting rule of 2 maximum branches). We got the following results:



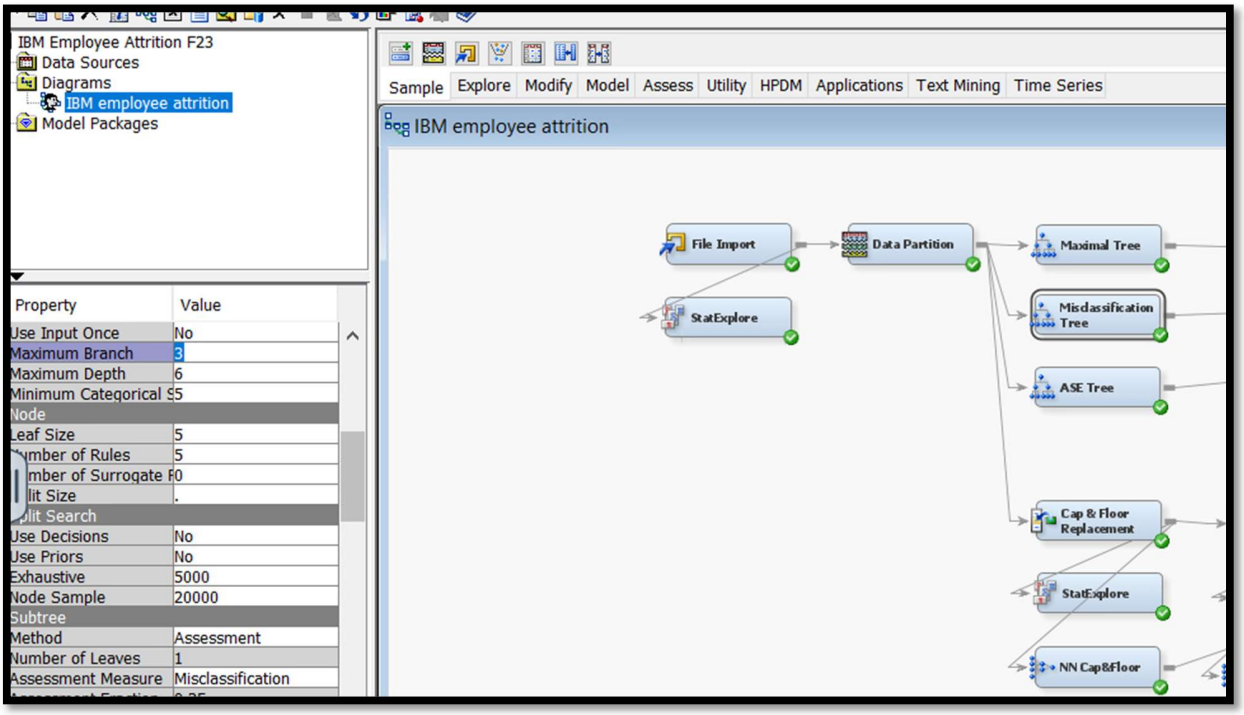


With a 2-way split, it turns out that no leaves were generated. We looked into the importance of the variables and got the following results:

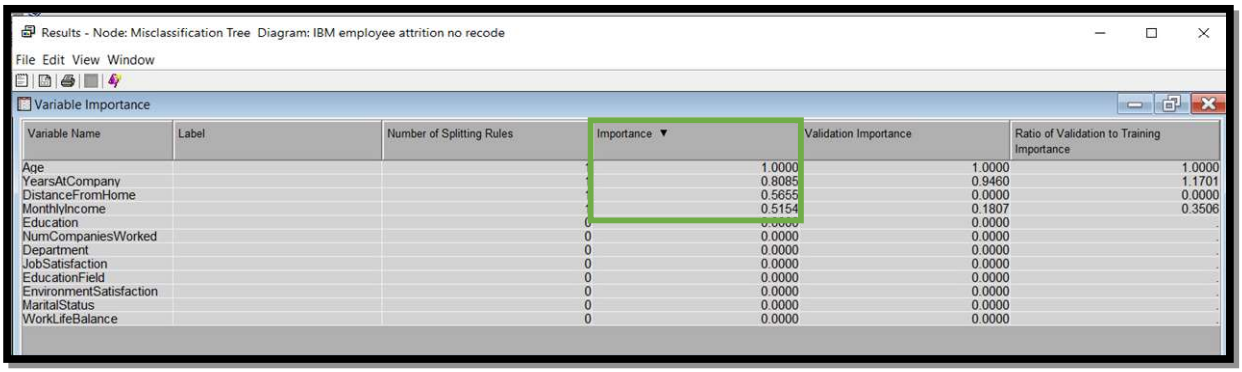


Variable Name	Label	Number of Splitting Rules	Importance	Va
Age		0	0.0000	
DistanceFromHome		0	0.0000	
MonthlyIncome		0	0.0000	
NumCompaniesWorked		0	0.0000	
YearsAtCompany		0	0.0000	
Department		0	0.0000	
Education		0	0.0000	
EducationField		0	0.0000	
EnvironmentSatisfaction		0	0.0000	
JobSatisfaction		0	0.0000	
MaritalStatus		0	0.0000	
WorkLifeBalance		0	0.0000	

It turned out that none of the input variables were important thus no split was generated. To address this, we tried increasing the splitting rule of maximum number of branches from 2 to 3.

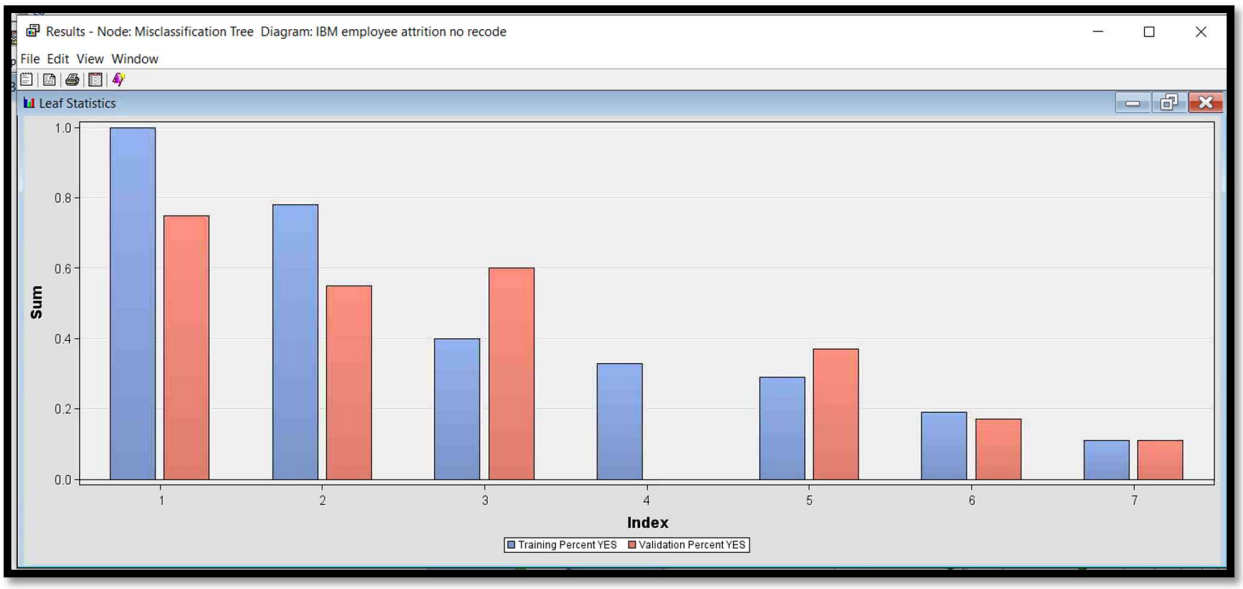


From the 3-way split, several input variables emerged as important: *age, years at company, distance*, and *monthly income.*

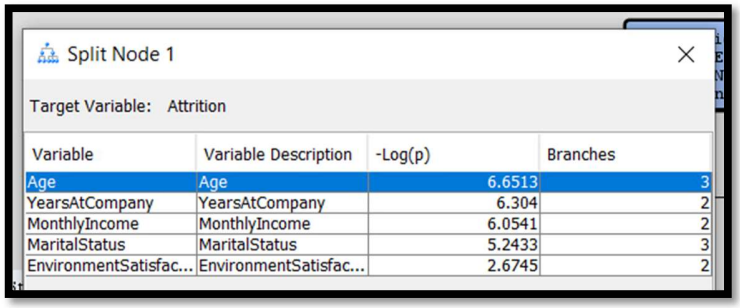


Because of this, the Misclassification Tree was able to generate a tree map. The results are as follows:

- Number of leaves: 7

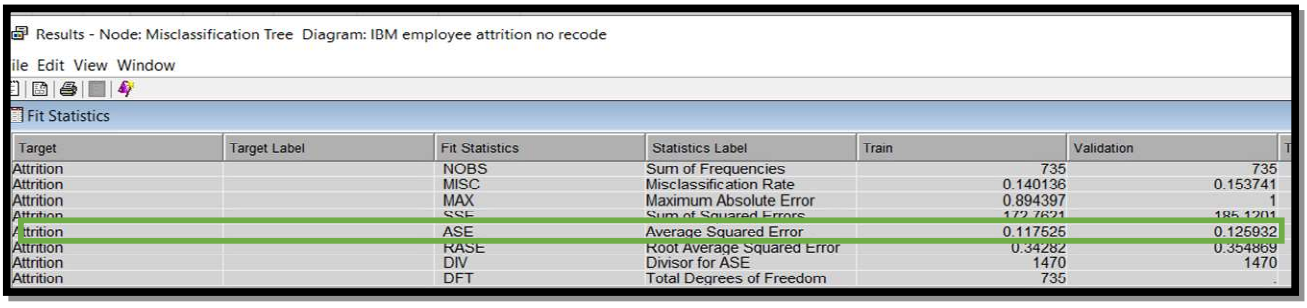


- Variables with the highest Logworth: age. The variable split into 3 branches.



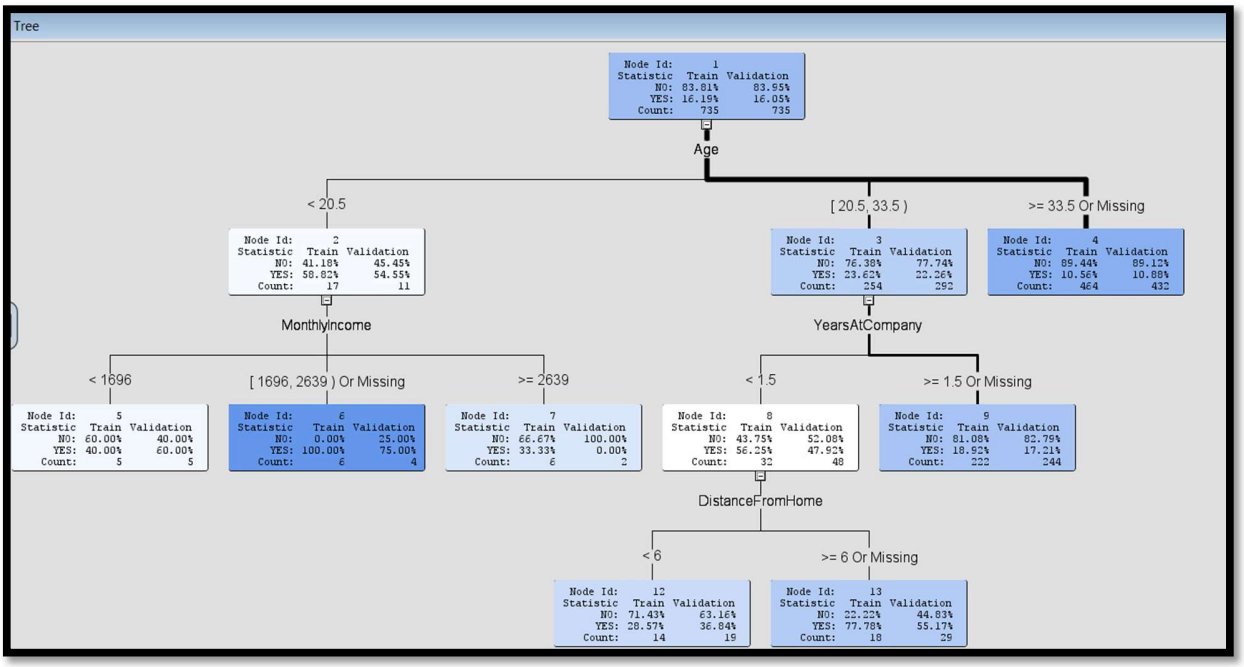
Variable	Variable Description	-Log(p)	Branches
Age	Age	6.6513	3
YearsAtCompany	YearsAtCompany	6.304	2
MonthlyIncome	MonthlyIncome	6.0541	2
MaritalStatus	MaritalStatus	5.2433	3
EnvironmentSatisfac...	EnvironmentSatisfac...	2.6745	2

- Fit Statistics revealed an Average Squared Error (ASE) value of **0.125932**.



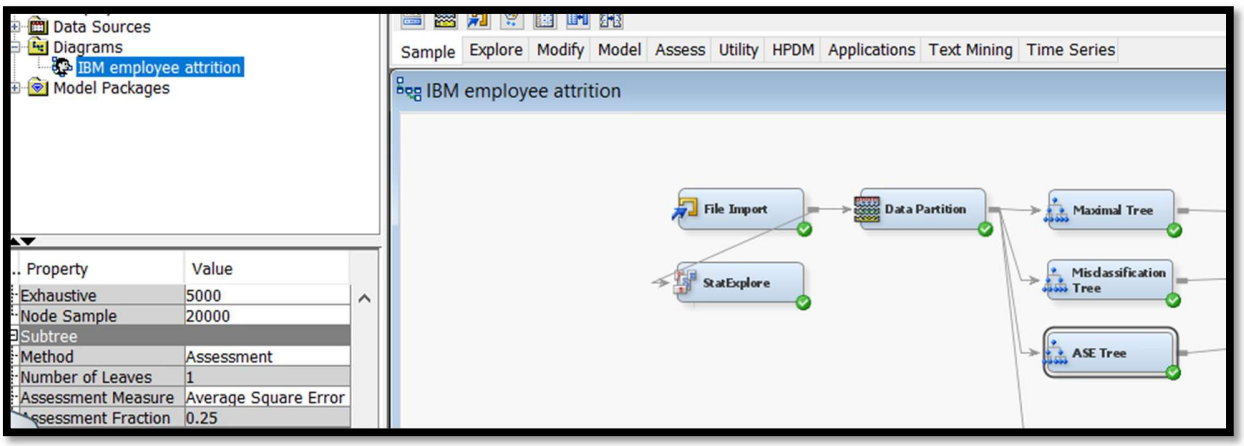
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Attrition		NOBS	Sum of Frequencies	735	735
Attrition		MISC	Misclassification Rate	0.140136	0.153741
Attrition		MAX	Maximum Absolute Error	0.894397	1
Attrition		SSE	Sum of Squared Errors	172.7621	185.1201
Attrition		ASE	Average Squared Error	0.117525	0.125932
Attrition		RASE	Root Average Squared Error	0.34282	0.354869
Attrition		DIV	Divisor for ASE	1470	1470
Attrition		DFT	Total Degrees of Freedom	735	.

- We got the following result for the tree:

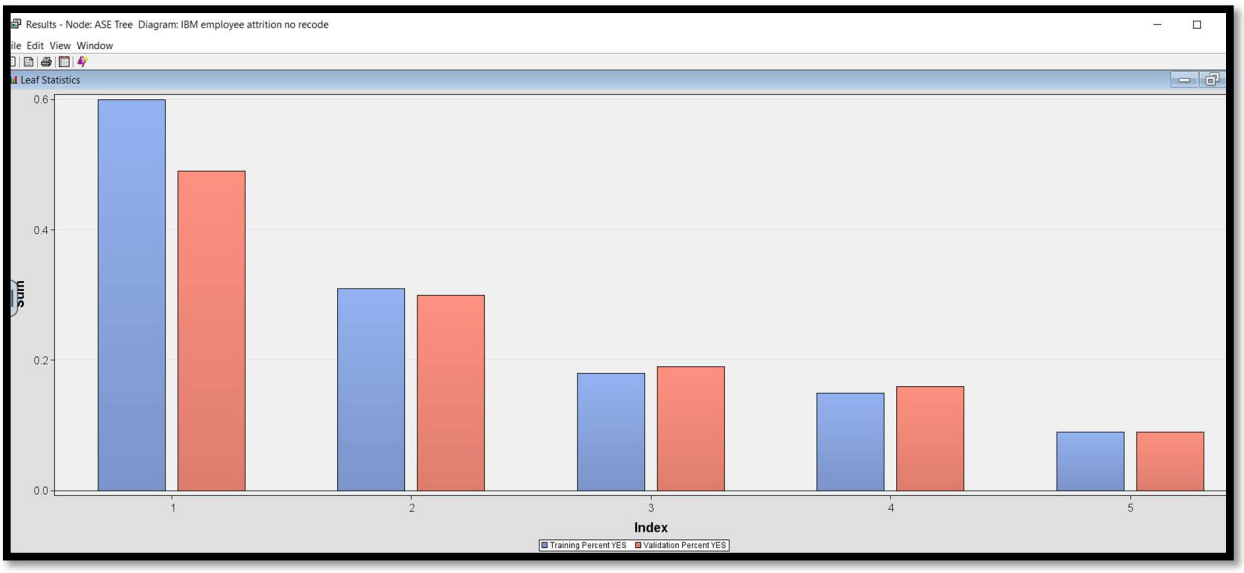


3. ASE Tree

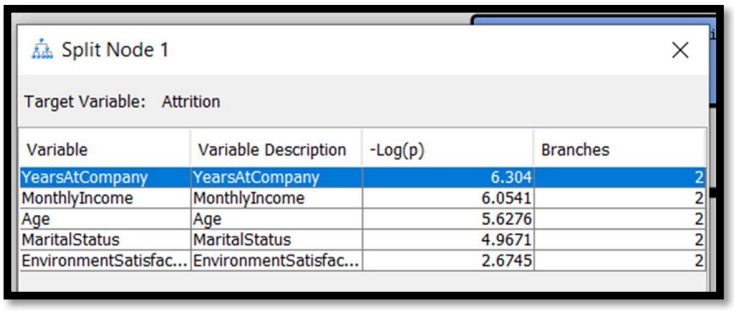
We tried another variation of a Decision Tree wherein we change the Assessment Measure to Average Square Error or ASE. This time we kept the default of 2 for maximum branch splitting rule. We got the following results:



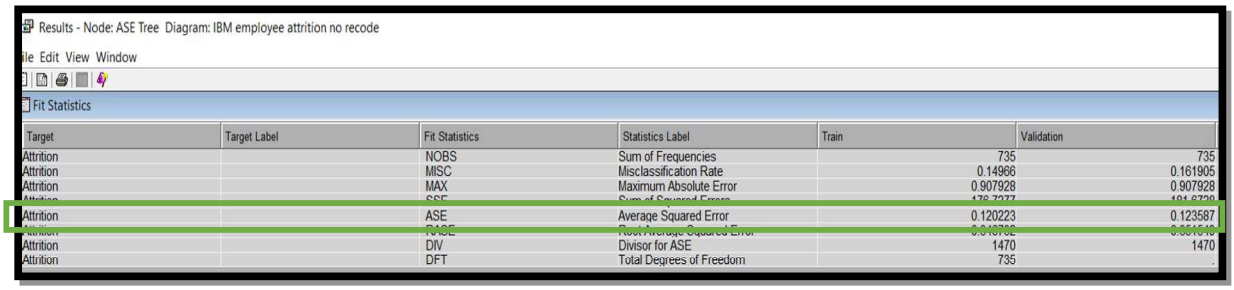
- Number of leaves: 5



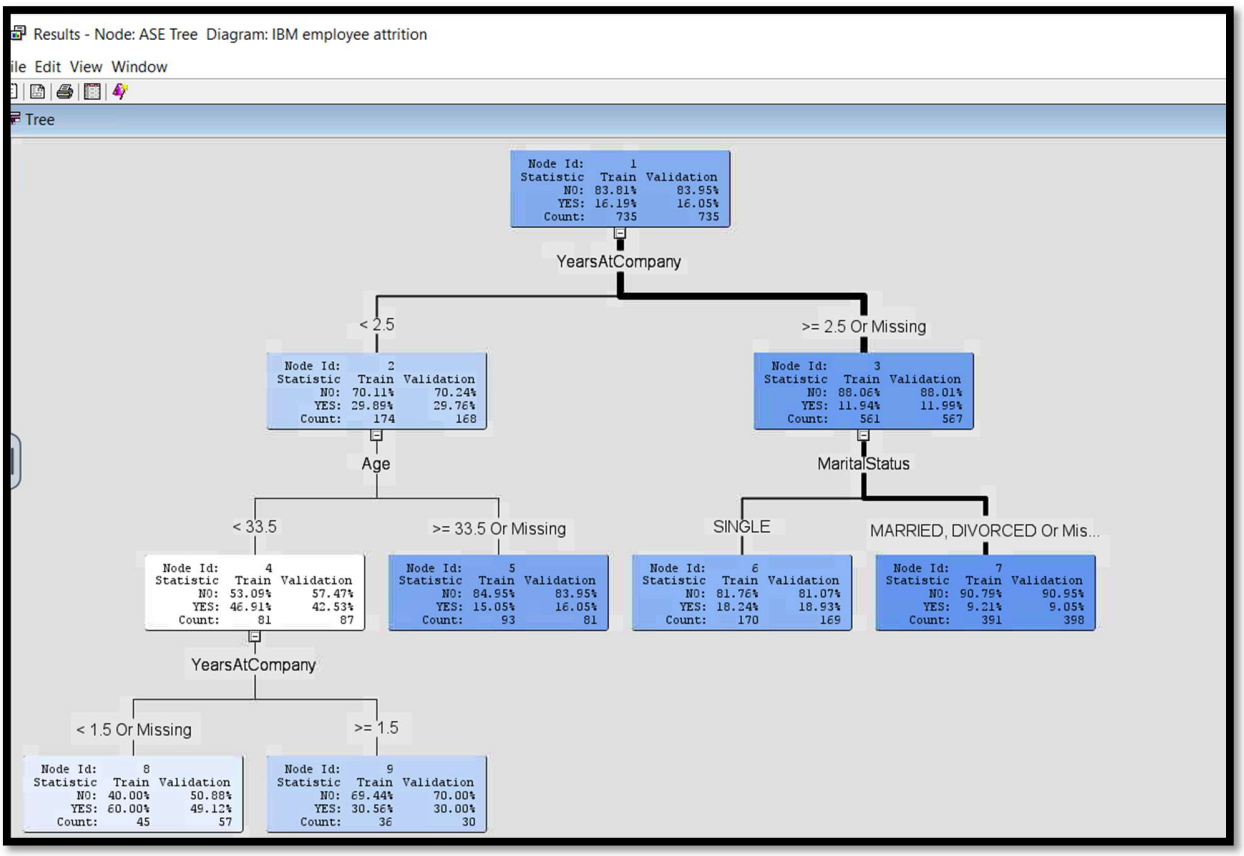
- Variables with the highest Logworth: *Years at company, monthly income, age, marital status, and environmental satisfaction.*



- Fit Statistics revealed an Average Squared Error (ASE) value of **0.123587**.



- We got the following result for the tree:



4. Best Decision Tree Model

The table below summarizes the ASE values for each of the decision tree models generated.

Decision Trees			
	Maximal Tree	**Misclassification Tree**	**ASE Tree**
Method	Largest	Assessment	Assessment
Assessment Measure	Decision	Misclassification	Average Score Error
ASE Value	0.127063	0.125932	0.123587

Based on the ASE values, the best model is the ASE Tree since it has the lowest ASE value of 0.123587.

Based on the results according to log order:

Significant Splits (Split Node)		
Maximal Tree	**Misc Tree**	**ASE Tree**
YearsAtCompany	Age	YearsAtCompany
Monthly Income	YearsAtCompany	Monthly Income
Age	Monthly Income	Age
Marital Status	Marital Status	Marital Status
Environment	Environment	Environment

From this, we used the results from the ASE Tree to make inferences. The following are our observations based on the significant variables in the tree:

- The input variable ***years at the company*** turned out to be the most significant variable. Those who had equal to or more than2.5 years in the company are more likely to stay by 88% compared to those with less than 2.5 years (70%). Employees who stayed longer at IBM are less likely to resign because they have already invested their time, skills and talent in the company.
- From the employees with more than 2.5 years, married and divorced, 91% are more likely to stay in the company vs single (81%).
- Another significant variable is age. Those who are older, 33.5 years old or older, are 84% more likely to stay in the company compared to only 57% for those who are less than 33.5 years old.
- For those who are younger but spent more than 1.5 years in the company, they are more likely to stay 70% compared to new hires 51%.
- In summary, employees who are married or divorced, are working more than 2.5 years at IBM are more likely to stay. Employees who have invested more years in the company are more likely to stay because they have already grown in their professional career. Moreover, employees who are married are more likely to stay since having a secure job is

more important for them because of their family responsibilities. On the other hand, those who are relatively new in the company and are young are more likely to resign. Young graduates have the tendency to move from one job to another as they try to figure out the best career that will work out for them.

- The characteristics of employees who are more likely to resign are: new hires and younger employees.
- The characteristics of employees who are more likely to stay: spent more years in the company, and are married.

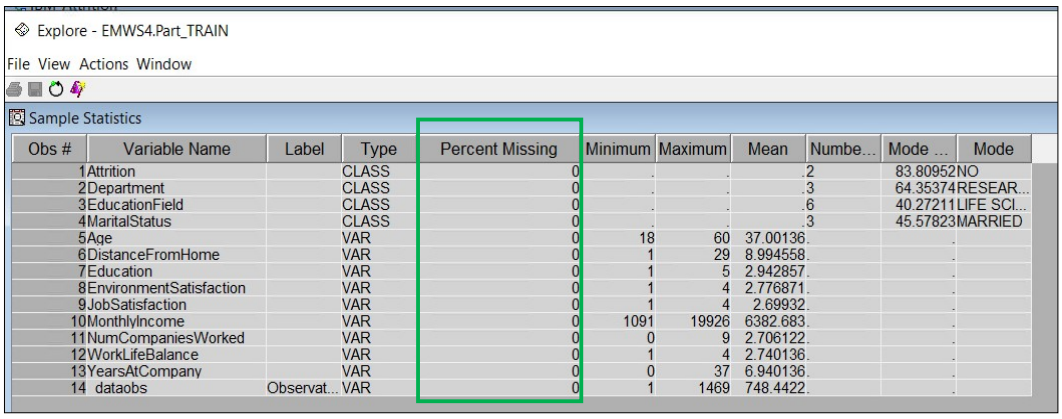
Regression Model

A. Data Preparation

Unlike Decision Tree, Regression model is sensitive to data completeness and distribution, for it uses prediction formula to train the model and score new cases. As preparation for regression modelling, issues such as missing values, skewness, and impact of non-numeric inputs were identified and addressed.

i. Missing Values

To check if there are missing values, train data was exported from the Data Partition node.

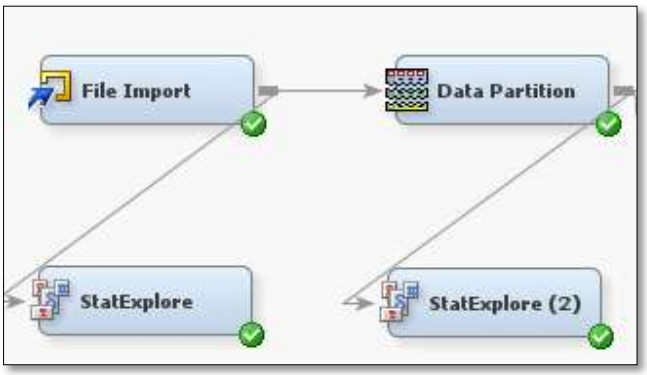


Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	Attrition		CLASS	0	.	.		.2	83.80952	NO
2	Department		CLASS	0	.	.		.3	64.35374	RESEAR...
3	EducationField		CLASS	0	.	.		.6	40.27211	LIFE SCI...
4	MaritalStatus		CLASS	0	.	.		.3	45.57823	MARRIED
5	Age		VAR	0	18	60	37.00136	.	.	
6	DistanceFromHome		VAR	0	1	29	8.994558	.	.	
7	Education		VAR	0	1	5	2.942857	.	.	
8	EnvironmentSatisfaction		VAR	0	1	4	2.776871	.	.	
9	JobSatisfaction		VAR	0	1	4	2.69932	.	.	
10	MonthlyIncome		VAR	0	1091	19926	6382.683	.	.	
11	NumCompaniesWorked		VAR	0	0	9	2.706122	.	.	
12	WorkLifeBalance		VAR	0	1	4	2.740136	.	.	
13	YearsAtCompany		VAR	0	0	37	6.940136	.	.	
14	dataobs	Observat...	VAR	0	1	1469	748.4422	.	.	

Sample Statistics was examined, and since the dataset has no missing values, imputation was not necessary.

ii. Skewness of Input Variables

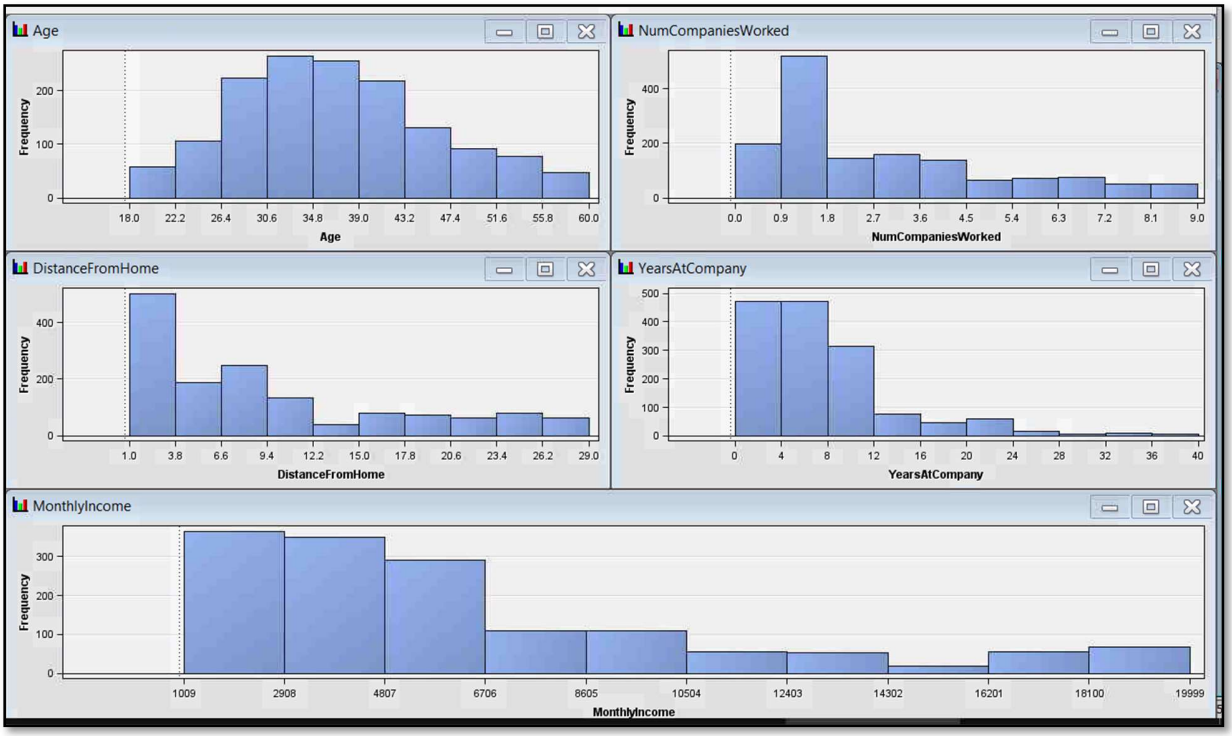
To check for skewness in the distribution of interval variables, Stat Explore node was brought in and connected to Data Partition node. A graph for each of the interval variables is also presented below.



File Edit View Window

Interval Variables

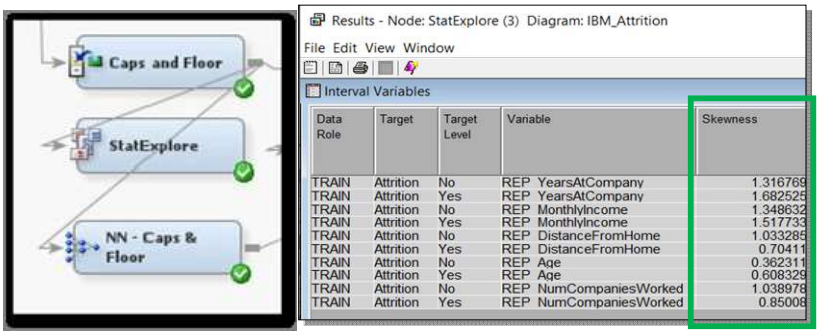
Data Role	Target	Target Level	Variable	Skewness ▼
TRAIN	Attrition	Yes	YearsAtCompany	1.996831
TRAIN	Attrition	No	YearsAtCompany	1.755979
TRAIN	Attrition	Yes	MonthlyIncome	1.517733
TRAIN	Attrition	No	MonthlyIncome	1.348632
TRAIN	Attrition	No	NumCompaniesWorked	1.038978
TRAIN	Attrition	No	DistanceFromHome	1.033285
TRAIN	Attrition	Yes	NumCompaniesWorked	0.85008
TRAIN	Attrition	Yes	DistanceFromHome	0.70411
TRAIN	Attrition	Yes	Age	0.608329
TRAIN	Attrition	No	Age	0.362311



Based on the generated results, the following interval variables have skewness beyond the cut-off of 1: **Years at Company, Monthly Income, Number of Companies Worked, and Distance from Home.** Hence, regularization of skewed data is warranted.

iii. Regularizing Extreme Input Interval Variables

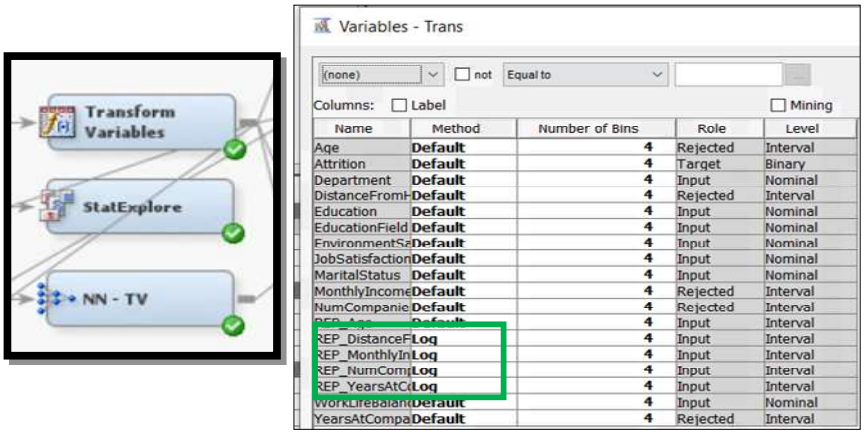
- **Cap & Floor** - To mitigate the impact of extreme variables, the first approach performed was Cap and Floor which used a standard deviation of 3 as a cut-off basis for replacement values.



Data Role	Target	Target Level	Variable	Skewness
TRAIN	Attrition	No	REP_YearsAtCompany	1.316769
TRAIN	Attrition	Yes	REP_YearsAtCompany	1.682525
TRAIN	Attrition	No	REP_MonthlyIncome	1.348632
TRAIN	Attrition	Yes	REP_MonthlyIncome	1.517733
TRAIN	Attrition	No	REP_DistanceFromHome	1.033285
TRAIN	Attrition	Yes	REP_DistanceFromHome	0.70411
TRAIN	Attrition	No	REP_Age	0.362311
TRAIN	Attrition	Yes	REP_Age	0.608329
TRAIN	Attrition	No	REP_NumCompaniesWorked	1.038978
TRAIN	Attrition	Yes	REP_NumCompaniesWorked	0.85008

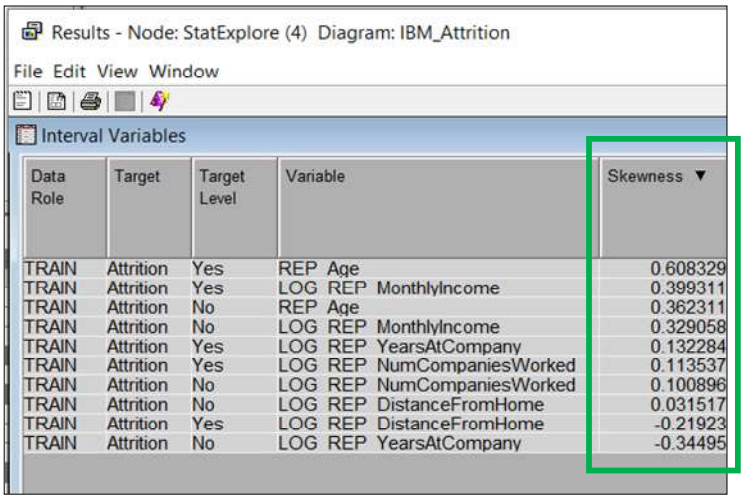
Although Cap & Floor reduced the skewness of identified variables, they are still beyond the cut-off of 1.

- **Transformation** – the next approach performed to address the remaining skewed variables was transformation, and log was used as basis for replacement values.



Name	Method	Number of Bins	Role	Level
Age	Default	4	Rejected	Interval
Attrition	Default	4	Target	Binary
Department	Default	4	Input	Nominal
DistanceFromH	Default	4	Rejected	Interval
Education	Default	4	Input	Nominal
EducationField	Default	4	Input	Nominal
EnvironmentSa	Default	4	Input	Nominal
JobSatisfaction	Default	4	Input	Nominal
MaritalStatus	Default	4	Input	Nominal
MonthlyIncome	Default	4	Rejected	Interval
NumCompanie	Default	4	Rejected	Interval
[illegible]	[illegible]	4	Input	Interval
REP_DistanceF	Log	4	Input	Interval
REP_MonthlyIn	Log	4	Input	Interval
REP_NumComp	Log	4	Input	Interval
REP_YearsAtCo	Log	4	Input	Interval
WorkLifeBalan	Default	4	Input	Nominal
YearsAtCompa	Default	4	Rejected	Interval

Another Stat Explore node was added and connected to Transform Variables to check the results of Transformation.



Data Role	Target	Target Level	Variable	Skewness ▼
TRAIN	Attrition	Yes	REP Age	0.608329
TRAIN	Attrition	Yes	LOG REP MonthlyIncome	0.399311
TRAIN	Attrition	No	REP Age	0.362311
TRAIN	Attrition	No	LOG REP MonthlyIncome	0.329058
TRAIN	Attrition	Yes	LOG REP YearsAtCompany	0.132284
TRAIN	Attrition	Yes	LOG REP NumCompaniesWorked	0.113537
TRAIN	Attrition	No	LOG REP NumCompaniesWorked	0.100896
TRAIN	Attrition	No	LOG REP DistanceFromHome	0.031517
TRAIN	Attrition	Yes	LOG REP DistanceFromHome	-0.21923
TRAIN	Attrition	No	LOG REP YearsAtCompany	-0.34495

After the log transformation, all remaining skewed variables were fixed to below 1.

iv. Recoding Non-numeric Inputs

To examine the impact of reduced levels of non-numeric or categorical inputs, recode simulation was performed.

Replacement Values for Class Variables

Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value	Label
Education	3	N		3	23	
Education	4	N		4	45	
Education	2	N		2	23	
Education	1	N		1	1	
Education	5	N		5	45	
EducationField	Life Sciences	C	Life Sciences	.	LM	
EducationField	Medical	C	Medical	.	LM	
EducationField	Marketing	C	Marketing	.	MKTG	
EducationField	Technical Degree	C	Technical Degree	.	TECH	
EducationField	Other	C	Other	.	OHR	
EducationField	Human Resources	C	Human Resources	.	OHR	
MaritalStatus	Married	C	Married	.	M	
MaritalStatus	Single	C	Single	.	SD	
MaritalStatus	Divorced	C	Divorced	.	SD	

Replacement Counts

Obs	Variable	Role	Label	Train	Validation
1	Education	INPUT		663	637
2	EducationField	INPUT		735	735
3	MaritalStatus	INPUT		735	735

Education, Education Field, and Marital Status were recoded to reduce the levels of non-numeric inputs. Overall, a total of 2,107 records were affected.

With Recode

Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.682	0.123587
Backward Regression	0.777	0.109754
Forward Regression	0.74	0.115813
Full Regression	0.771	0.111461
Maximal Tree	0.671	0.127063
Misclassification Tree	0.646	0.125932
NN BReg 2H	0.775	0.111829
NN BReg 3H	0.783	0.110521
NN BReg 4H	0.773	0.112225
NN BReg 5H	0.762	0.113858
NN BReg 6H	0.777	0.109461
NN BReg 7H	0.767	0.111574
NN BReg 8H	0.767	0.110065
NN Cap&Floor	0.755	0.115545
NN Recode	0.759	0.114041
NN Transform	0.777	0.110954
Stepwise Regression	0.724	0.118873

Without Recode

Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.682	0.123587
Backward Regression	0.776	0.10897
Forward Regression	0.751	0.113473
Full Regression	0.767	0.110665
Maximal Tree	0.671	0.127063
Misclassification Tree	0.646	0.125932
NN BReg 2H	0.777	0.10975
NN BReg 3H	0.766	0.114322
NN BReg 4H	0.762	0.108916
NN BReg 5H	0.779	0.106623
NN BReg 6H	0.784	0.107588
NN BReg 7H	0.779	0.108853
NN BReg 8H	0.775	0.110949
NN Cap&Floor	0.755	0.115545
NN Transform	0.777	0.110954
Stepwise Regression	0.751	0.113473

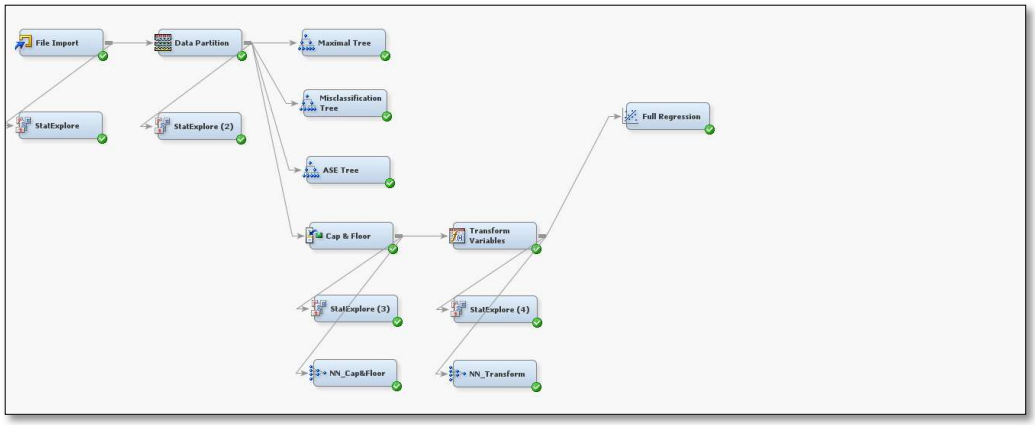
The simulation results show that the model is better off without recoding the categorical levels. Thus, the succeeding models were based on non-recoded inputs.

B. Regression Prediction Formula

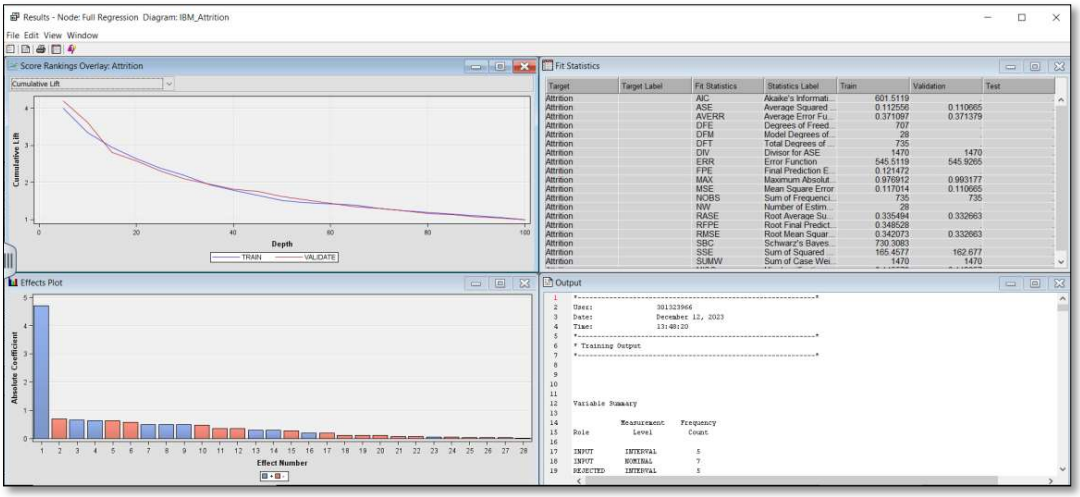
Since the target outcome is binary in nature, which is to predict the occurrence of employees' attrition expressed as Yes and No, the prediction formulas used were *logit link and logistic function*. In this way, primary outcomes are expressed in the form of odds or probabilities.

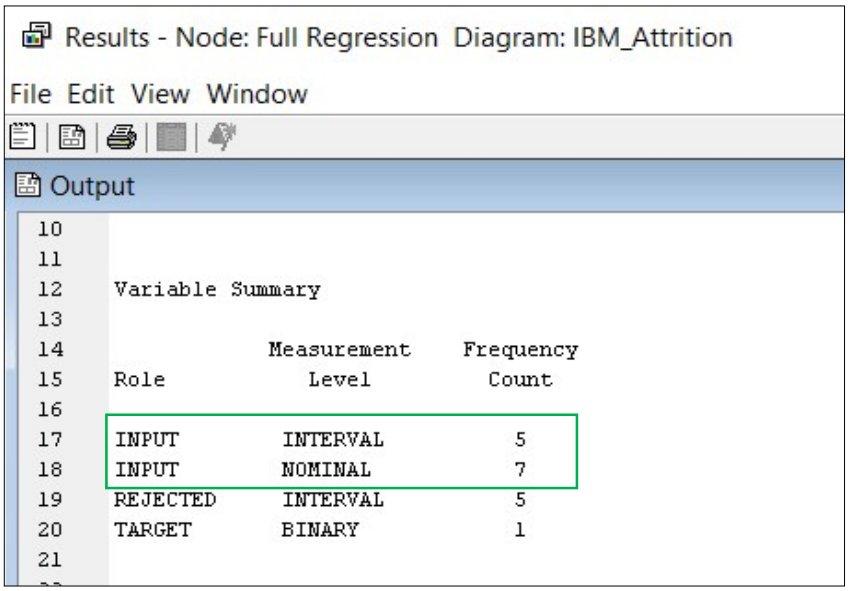
C. Initializing Full Regression Model

Based on the simulation exercise performed for recoding non-numeric inputs, results suggest that it would worsen the model fit. Thus, regression models were run based on set of non-recoded variables.

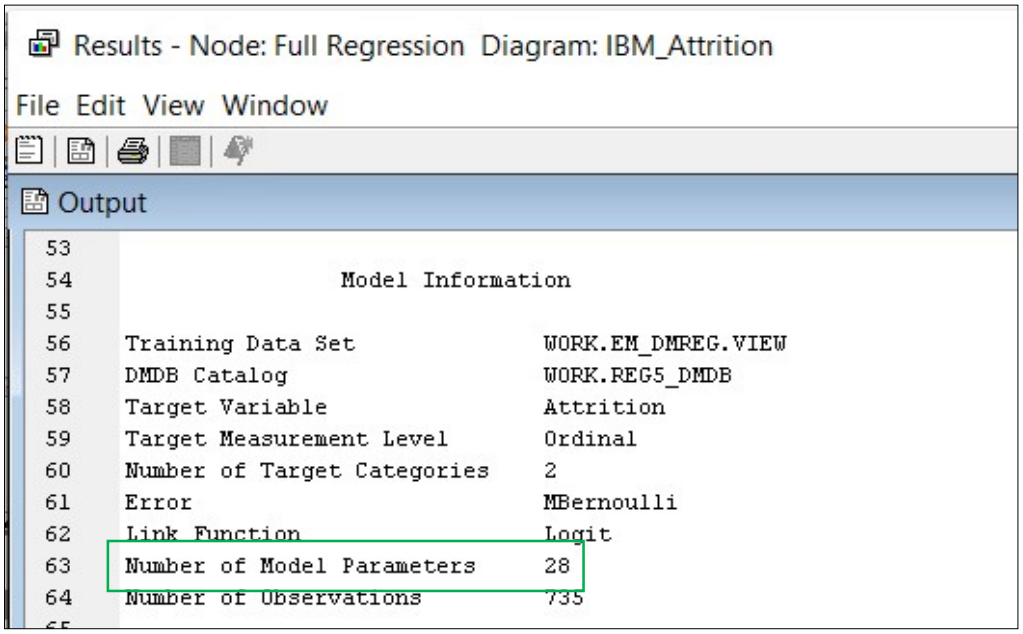


A Full Regression model was integrated into the diagram, connected to transform variables node, run, and generated the related results.

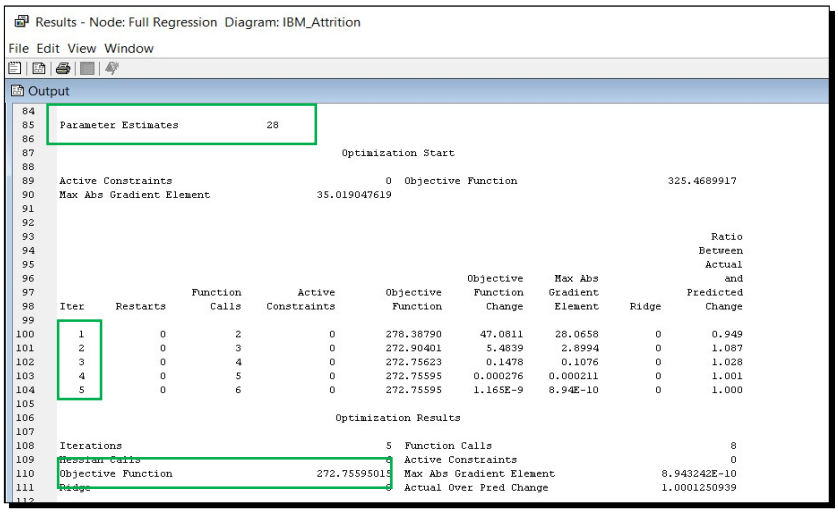




Referring to the variable summary, the model used**12 input variables** comprising 5intervals and 7 categorical.



According to the model information, Full Regression has **28parameter estimates (**including intercept), which apparently not equal to the number of input variables mentioned earlier.



Results - Node: Full Regression Diagram: IBM_Attrition

File Edit View Window

Output

Parameter Estimates 28

Optimization Start

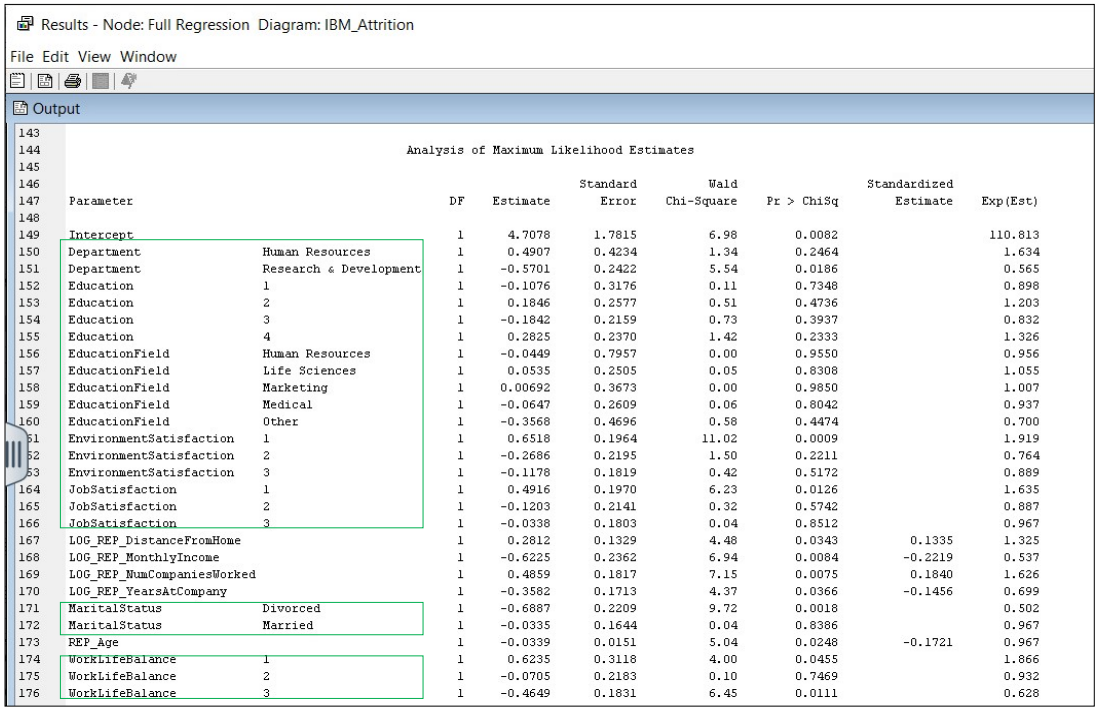
Active Constraints			0	Objective Function				325.4689917
Max Abs Gradient Element			35.019047619					

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Ridge	Ratio Between Actual and Predicted Change
1	0	2	0	278.38790	47.0811	28.0658	0	0.949
2	0	3	0	272.90401	5.4839	2.8994	0	1.087
3	0	4	0	272.75623	0.1478	0.1076	0	1.028
4	0	5	0	272.75595	0.000276	0.000211	0	1.001
5	0	6	0	272.75595	1.165E-9	8.94E-10	0	1.000

Optimization Results

Iterations	5	Function Calls	8
Hessian Calls	6	Active Constraints	0
Objective Function	272.75595015	Max Abs Gradient Element	8.943242E-10
Ridge	0	Actual Over Pred Change	1.0001250939

The model underwent five iterations to come up with optimal estimates of these 28 parameters. The iterations stopped when decline in objective function is negligible.



Results - Node: Full Regression Diagram: IBM_Attrition

File Edit View Window

Output

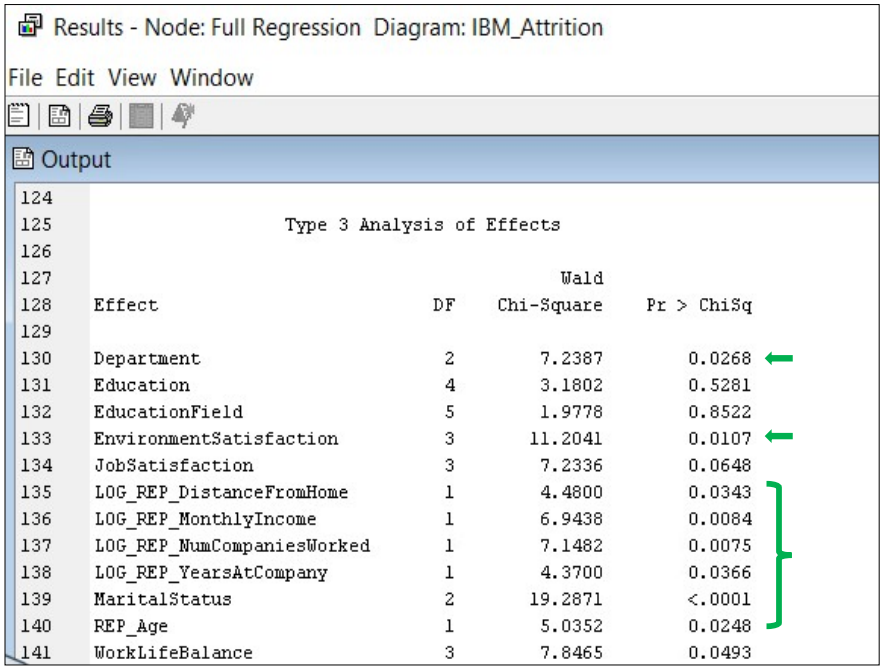
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	4.7078	1.7815	6.98	0.0082		110.813
Department	Human Resources	1	0.4907	0.4234	1.34	0.2464		1.634
Department	Research & Development	1	-0.5701	0.2422	5.54	0.0186		0.565
Education	1	1	-0.1076	0.3176	0.11	0.7348		0.898
Education	2	1	0.1846	0.2577	0.51	0.4736		1.203
Education	3	1	-0.1842	0.2159	0.73	0.3937		0.832
Education	4	1	0.2825	0.2370	1.42	0.2333		1.326
EducationField	Human Resources	1	-0.0449	0.7957	0.00	0.9550		0.956
EducationField	Life Sciences	1	0.0535	0.2505	0.05	0.8308		1.055
EducationField	Marketing	1	0.00692	0.3673	0.00	0.9850		1.007
EducationField	Medical	1	-0.0647	0.2609	0.06	0.8042		0.937
EducationField	Other	1	-0.3568	0.4696	0.58	0.4474		0.700
EnvironmentSatisfaction	1	1	0.6518	0.1964	11.02	0.0009		1.919
EnvironmentSatisfaction	2	1	-0.2686	0.2195	1.50	0.2211		0.764
EnvironmentSatisfaction	3	1	-0.1178	0.1819	0.42	0.5172		0.889
JobSatisfaction	1	1	0.4916	0.1970	6.23	0.0126		1.635
JobSatisfaction	2	1	-0.1203	0.2141	0.32	0.5742		0.887
JobSatisfaction	3	1	-0.0338	0.1803	0.04	0.8512		0.967
LOG_REP_DistanceFromHome		1	0.2812	0.1329	4.48	0.0343	0.1335	1.325
LOG_REP_MonthlyIncome		1	-0.6225	0.2362	6.94	0.0084	-0.2219	0.537
LOG_REP_NumCompaniesWorked		1	0.4859	0.1817	7.15	0.0075	0.1840	1.626
LOG_REP_YearsAtCompany		1	-0.3582	0.1713	4.37	0.0366	-0.1456	0.699
MaritalStatus	Divorced	1	-0.6887	0.2209	9.72	0.0018		0.502
MaritalStatus	Married	1	-0.0335	0.1644	0.04	0.8386		0.967
REP_Age		1	-0.0339	0.0151	5.04	0.0248	-0.1721	0.967
WorkLifeBalance	1	1	0.6235	0.3118	4.00	0.0455		1.866
WorkLifeBalance	2	1	-0.0705	0.2183	0.10	0.7469		0.932
WorkLifeBalance	3	1	-0.4649	0.1831	6.45	0.0111		0.628

If we examine the Analysis of Maximum Likelihood, aside from the intercept and interval variables, categorical levels of **Department, Education, Educational Field, Environment Satisfaction, Job Satisfaction, Marital Status, and Work Life Balance** also produced model parameters. Thus, the primary reason for the unequal number of

model parameters versus input variables was when categorical levels further generated model parameters.

However, it is worthy to note that not all categorical levels were considered as model parameters like the **Sales** department, educational degree of **5**, **Technical** education field, satisfaction rating of **4**, and **Single** marital status. We could assume that these levels are not statistically significant in making the observed data most likely.



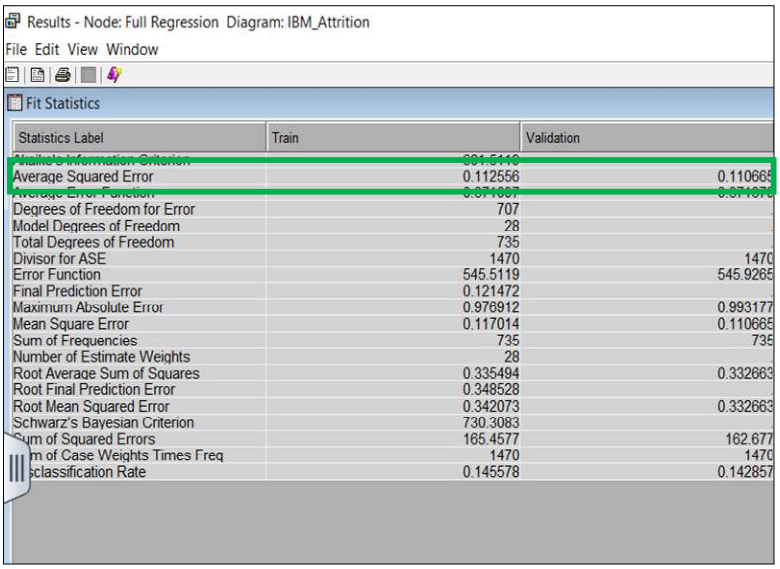
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Department	2	7.2387	0.0268
Education	4	3.1802	0.5281
EducationField	5	1.9778	0.8522
EnvironmentSatisfaction	3	11.2041	0.0107
JobSatisfaction	3	7.2336	0.0648
LOG_REP_DistanceFromHome	1	4.4800	0.0343
LOG_REP_MonthlyIncome	1	6.9438	0.0084
LOG_REP_NumCompaniesWorked	1	7.1482	0.0075
LOG_REP_YearsAtCompany	1	4.3700	0.0366
MaritalStatus	2	19.2871	<.0001
REP_Age	1	5.0352	0.0248
WorkLifeBalance	3	7.8465	0.0493

If we further explore the Type 3 Analysis of Effects, **Marital Status** has the highest significance for having the lowest p-value of **<.0001**.

Other variables identified as significant with p-values below .05 are: **Department, Environment Satisfaction**, **Distance from home, Monthly Income, Number of Companies Worked, YearsatCompany, and Age.**

On the other hand, **Education, Educational Field, Job Satisfaction, and WorkLifeBalance** have p-values closer to 1 and appeared as extraneous. These can be possibly eliminated to improve the model performance.

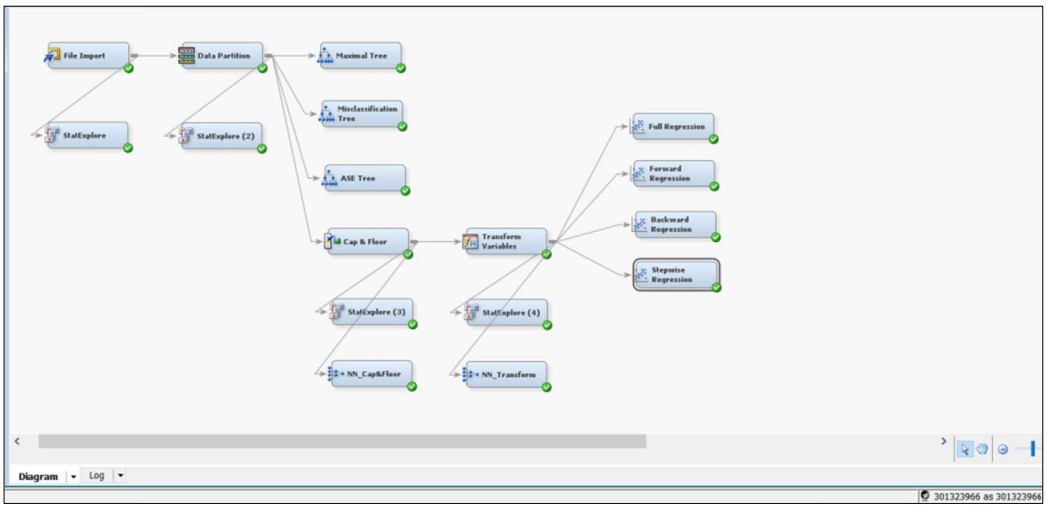


Statistics Label	Train	Validation
Average Squared Error	0.112556	0.110665
Degrees of Freedom for Error	707	
Model Degrees of Freedom	28	
Total Degrees of Freedom	735	
Divisor for ASE	1470	1470
Error Function	545.5119	545.9265
Final Prediction Error	0.121472	
Maximum Absolute Error	0.976912	0.993177
Mean Square Error	0.117014	0.110665
Sum of Frequencies	735	735
Number of Estimate Weights	28	
Root Average Sum of Squares	0.335494	0.332663
Root Final Prediction Error	0.348528	
Root Mean Squared Error	0.342073	0.332663
Schwarz's Bayesian Criterion	730.3083	
Sum of Squared Errors	165.4577	162.677
m of Case Weights Times Freq	1470	1470
sclassification Rate	0.145578	0.142857

The Validation ASE of Full Regression at **0.110665** was used as baseline to determine whether elimination of extraneous variables will improve or worsen the model.

D. Input Selection and Model Optimization

Sequential selection methods such as *Forward, Backward, and Stepwise* were explored to identify the best set of input variables and optimize the model complexity of Full Regression.

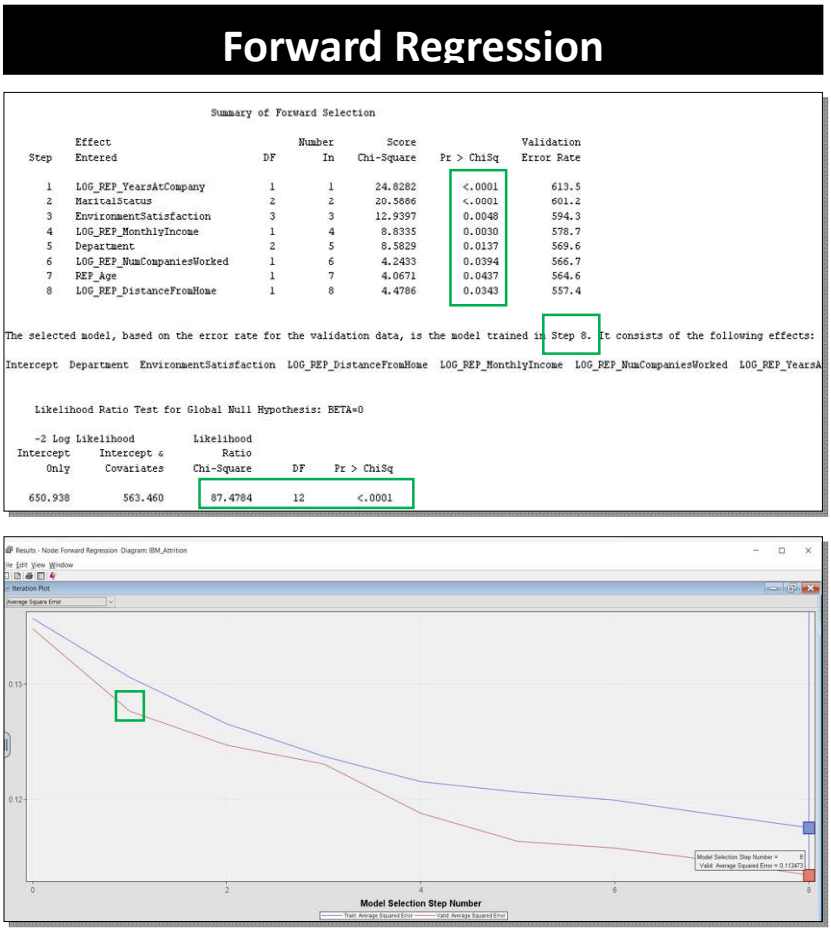


Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error

In the properties of each sequential node, the selection model was set accordingly, and the selection criterion used was *validation error.*

In the following sections, we will review the results of the sequential regression models.

- **Forward Selection**



Forward selection reached until **Step 8**, which was the selected model. **Eight input variables** were qualified to be added since their p-values are below the

entry cut-off of <.05. The final input combination of Forward regression is comprised of **12 degrees of freedom** or parameter estimates resulting to overall p-value of <.0001.

The iteration plot shows the continuous improvement in model performance as the model progress from Step 0 until it stopped at Step 8 with validation ASE of **0. 113473.** The performance of Forward Regression is worse compared to Full Regression for having higher ASE.

- **Backward Selection**

Backward Regression

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	EducationField	5	11	1.9778	0.8522	548.6
2	Education	4	10	2.9928	0.5590	539.2
3	JobSatisfaction	3	9	6.6466	0.0841	550.4
4	WorkLifeBalance	3	8	6.7427	0.0806	557.4
5	LOG_REP_YearsAtCompany	1	7	3.7179	0.0538	561.3

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

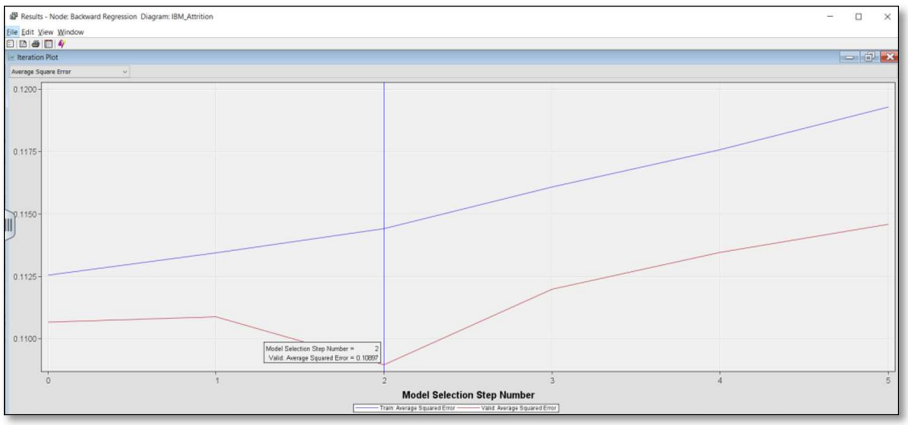
Intercept Department EnvironmentSatisfaction JobSatisfaction LOG_REP_DistanceFromHome LOG_REP_MonthlyIncome LOG_REP_NumCompaniesWorked

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
650.938	550.487	100.4509	18	<.0001

The Summary of Backward Elimination indicates that the model sequentially removed the input variables with p-values more than the stay cut-off of <.05.

Five variables were eliminated producing a seven-input model. However, the selected model is the one trained during **Step 2, a ten-input model** when Education Field and Education were eliminated. The other three variables were not removed despite of having p-values more than the stay cut-off, for it would drive higher validation error and worsen model performance.



If we further examine the iteration plot, it shows that initial validation ASE at Step 0-1 is high; however, it significantly declined in **Step 2 at 0.10897.** Then, it continuously increased in the succeeding steps until it peaked at Step 5.

*** Notice that model fit improved when the two categorical variables were removed. However, it worsened when the interval variable was eliminated.*

- **Stepwise Selection**

Stepwise Regression

Summary of Stepwise Selection

Step	Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	LOG_REP_YearsAtCompany		1	1	24.8282		<.0001	613.5
2	MaritalStatus		2	2	20.5886		<.0001	601.2
3	EnvironmentSatisfaction		3	3	12.9397		0.0048	594.3
4	LOG_REP_MonthlyIncome		1	4	8.8335		0.0030	578.7
5	Department		2	5	8.5829		0.0137	569.6
6	LOG_REP_NumCompaniesWorked		1	6	4.2433		0.0394	566.7
7	REP_Age		1	7	4.0671		0.0437	564.6
8	LOG_REP_DistanceFromHome		1	8	4.4786		0.0343	557.4
9		LOG_REP_YearsAtCompany	1	7		3.7179	0.0538	561.3

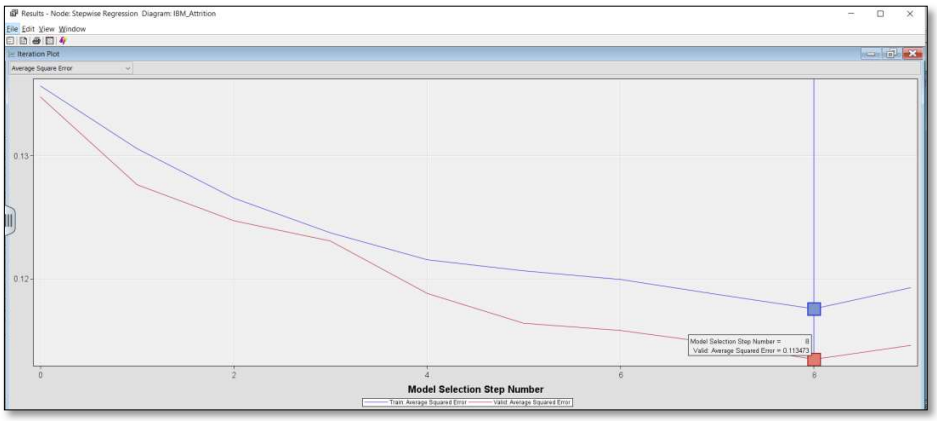
The selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:

Intercept Department EnvironmentSatisfaction LOG_REP_DistanceFromHome LOG_REP_MonthlyIncome LOG_REP_NumCompaniesWorked LOG_REP_YearsAtComp

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
650.938	563.460	87.4784	12	<.0001

Stepwise selection underwent a total of nine steps. The sequence of input entry is the same with forward selection; however, the final step removed **YearsAtCompany** after the model re-evaluated that its p-value went beyond the stay cut-off of <.05. Thus, it selected the model trained in Step 8, **a seven-input model**.



The iteration plot shows the continuous improvement in model performance as from Step 0 until it stopped at Step 8 with validation ASE of **0.113473.** When the model proceeded to Step 9, validation ASE got worse.

Stepwise regression resulted to the same validation ASE as with Forward regression.

E. Selection of the Best Model

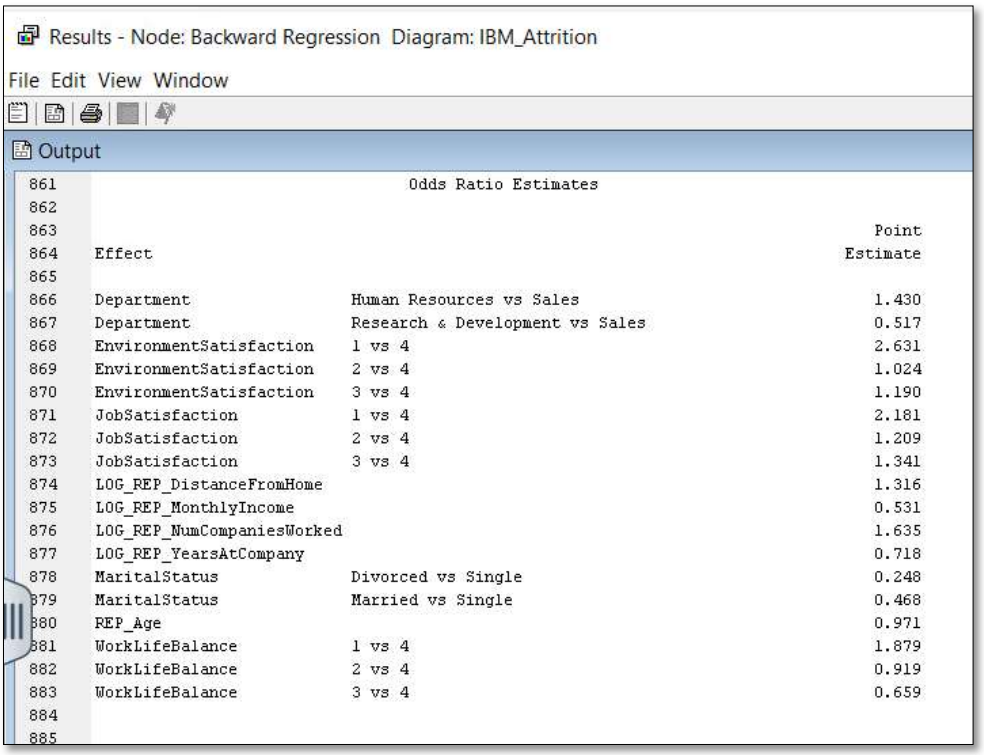
With the aim of optimizing model complexity, sequential-based models were explored to configure the best configuration of input variables. Below is the comparative summary of all regression models:

Summary of Regression Models				
Comparison Points	**Full Regression**	**Forward Regression**	**Backward Regression**	**Stepwise Regression**
Number of Steps		8	5	9
Model Selected trained at:		Step 8	Step 2	Step 8
Input Variables	**Total : 12** Interval: 5 Categorical: 7	**Total : 8** Interval: 5 Categorical: 3	**Total : 10** Interval: 5 Categorical: 5	**Total : 7** Interval: 4 Categorical: 3
Parameter Estimates (excluding intercept)	**Total : 27** Interval: 5 Categorical: 22	**Total : 12** Interval: 5 Categorical: 7	**Total : 18** Interval: 5 Categorical: 13	**Total : 11** Interval: 4 Categorical: 7
Validation ASE	**0.110665**	**0.113473**	**0.10897**	**0.113473**

Compared to Full Regression, Forward and Stepwise worsen the model for generating higher validation ASE. However, **Backward** came off as the model that optimized the complexity of Full Regression for it decreased the Validation ASE.

Therefore, Backward regression was selected as the best model and became the basis of succeeding analyses.

F. Interpretation of Backward Regression Model



Odds Ratio Estimates

Effect		Point Estimate
Department	Human Resources vs Sales	1.430
Department	Research & Development vs Sales	0.517
EnvironmentSatisfaction	1 vs 4	2.631
EnvironmentSatisfaction	2 vs 4	1.024
EnvironmentSatisfaction	3 vs 4	1.190
JobSatisfaction	1 vs 4	2.181
JobSatisfaction	2 vs 4	1.209
JobSatisfaction	3 vs 4	1.341
LOG_REP_DistanceFromHome		1.316
LOG_REP_MonthlyIncome		0.531
LOG_REP_NumCompaniesWorked		1.635
LOG_REP_YearsAtCompany		0.718
MaritalStatus	Divorced vs Single	0.248
MaritalStatus	Married vs Single	0.468
REP_Age		0.971
WorkLifeBalance	1 vs 4	1.879
WorkLifeBalance	2 vs 4	0.919
WorkLifeBalance	3 vs 4	0.659

The interpretations of odds ratio estimates are as follows:

- Employees in Sales Department are more likely to resign by **48%** compared to Research and Development**;** however, they are more likely to stay by **43%**compared to Human Resources
- With respect to satisfaction surveys, low degree of satisfaction in terms of work environment and job satisfaction increases the probability of attrition. Below is the extent of high attrition likelihood versus **4** being the highest rating.

Rating	Environment Satisfaction	Job Satisfaction
1	2.63 times	2.18 times
2	2%	21%
3	19%	34%

- An increase in the kilometer distance between employees' residence and workplace, multiplied to a factor of 2.74, increases the chance of resignation by **32%.**
- An increase in employees' monthly income, multiplied to a factor of 2.74, decreases the probability of resignation by **47%**
- The higher the number of companies previously worked in by employees, multiplied to a factor of 2.74, the higher the chance of resignation by **64%.**
- The longer the years of employees' tenure multiplied to a factor of 2.74, the lesser likelihood of resignation by **28%**
- Single employees are more likely to resign by **75%** and **53%** vs. Divorced and Married employees, respectively.
- Younger employees are more likely to resign by **3%**
- In the aspect of Work Life Balance, employees who rate the company at 1(lowest), are the most likely to resign by **88%**. However, employees that rate the company with 2 & 3(better& good) are less likely to leave the company by **8%** and **35%** respectively compared to those with ratings of 4(best).

Neural Network Model

A. Preparation

Neural network model requires a complete record for estimation and scoring. Imputation is done to resolve this complication; however, for this study no imputation has been done as there are no missing values in the dataset.

In addition, in neural network model, transformations and replacements are not necessarily needed, but it takes advantage of these two. In this study, Cap and Floor and Transformation of variables have been used.

Though neural networks in general have no selection of inputs, the models used all those input variables selected by the Backward Regression model, which is the optimal regression model, in preparation for the neural network model that uses hidden units.

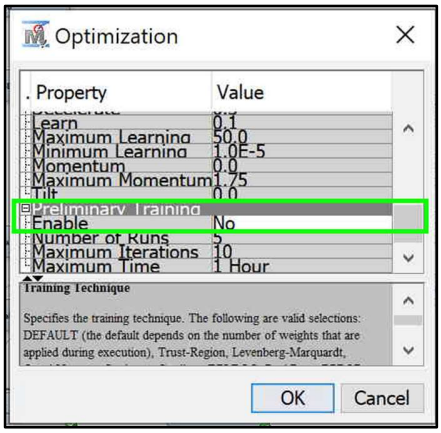
B. Prediction Formula

Neural network has different components such as hidden units, which includes an activation function called hyperbolic tangent and weight estimates.

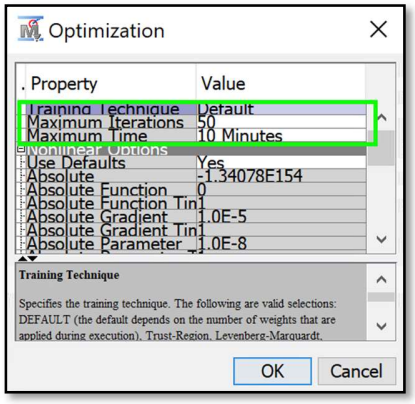
Given that the target variable is binary, the primary equation for the neural network regression utilizes the same logit link function as in logistic regression. Like logistic regression, the process of estimating weights shifts from least squares to maximum likelihood.

C. Complexity Optimization

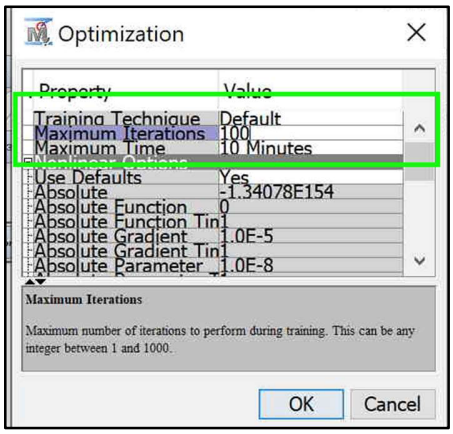
Complexity optimization is an integral part of neural network modelling. In this study, neural network models, we disabled the “preliminary training” as not to use randomly initialized weights.



Initially, we ran the model using the default maximum of 50 iterations, and maximum time set to 10 minutes.



We have tried increasing the maximum number of iterations to 100; however, fit statistics did not change at all.



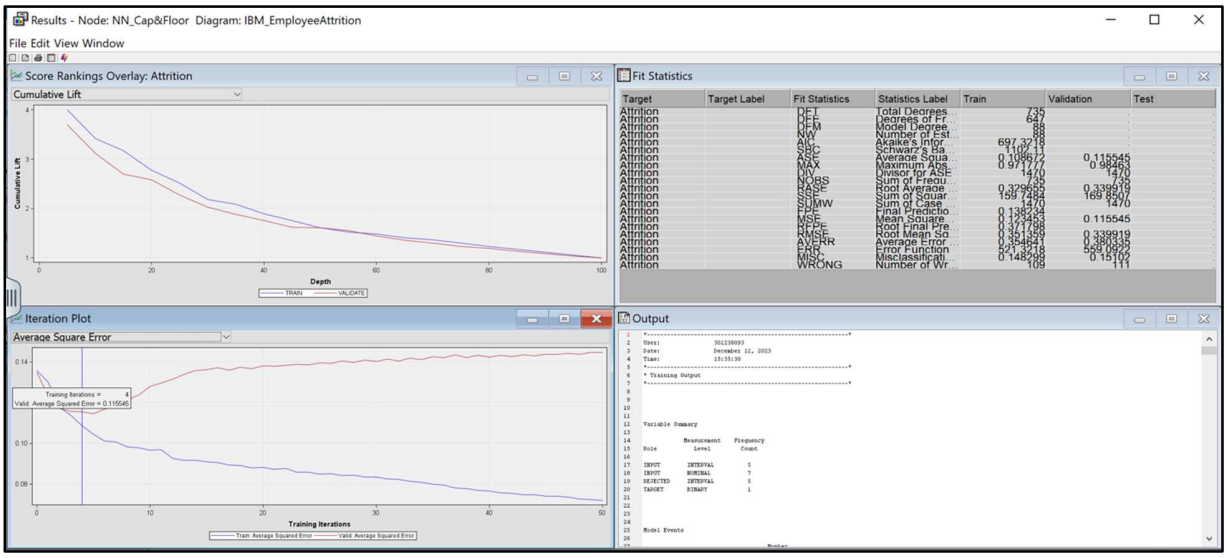
Below is the summary of the model optimization (maximizing the likelihood estimates of the model weights).

		Optimization Results			
	Iterations	50	Function Calls	58	
	Jacobian Calls	52	Active Constraints	0	
	Objective Function	0.2695346348	Max Abs Gradient Element	0.0013944801	
	Lambda	0.0069955836	Actual Over Pred Change	0.9367593001	
	Radius	0.4762285655			
	LEVMAR needs more than 50 iterations or 2147483647 function calls.				
	WARNING: LEVMAR Optimization cannot be completed.				

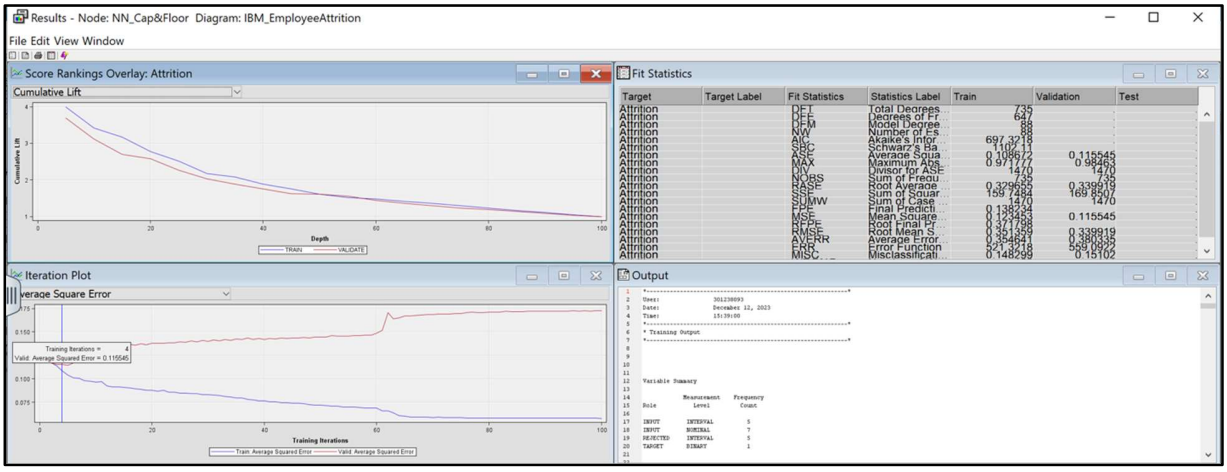
		Optimization Results			
	Iterations	100	Function Calls	120	
	Jacobian Calls	103	Active Constraints	0	
	Objective Function	0.2475165827	Max Abs Gradient Element	0.0003490251	
	Lambda	1.3735004359	Actual Over Pred Change	0.4586639168	
	Radius	0.0038606831			
	LEVMAR needs more than 100 iterations or 2147483647 function calls.				
	WARNING: LEVMAR Optimization cannot be completed.				

The model-fitting did not converge in 50 iterations or 100 iterations. Sampling NN Cap & Floor for both iterations, (50 and 100), it yielded same results.

NN_Cap&Floor (50)



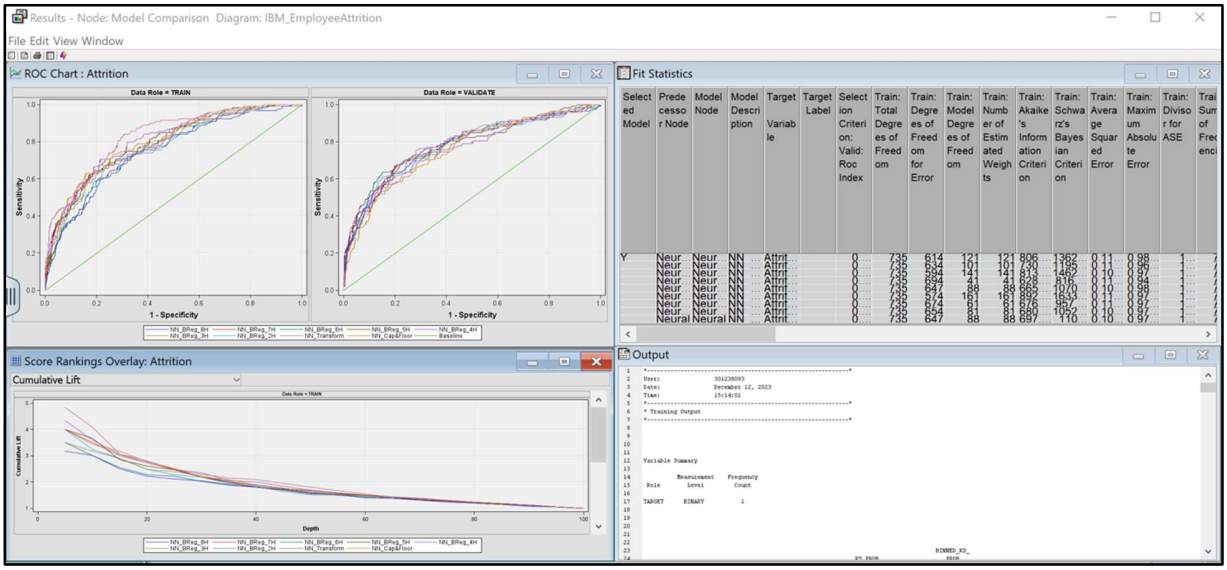
NN_Cap&Floor (100)



*** Note changing iterations have been done in all other neural network models and does not change the fit statistics at all.*

D. Neural Network Model Comparison

Performance of neural network models can be improved by changing the selected inputs and increasing the number of hidden units. We had manually changed the hidden units to see the performance starting from two (2) to eight (8) as guessing the best number involved “trial and error”.



Selected Model	Model Node	Model Description	Valid: Roc Index	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural7	NN_BReg_6H	0.784	0.11786	0.15646	0.10759	0.14286
	Neural6	NN_BReg_5H	0.779	0.11066	0.15102	0.10662	0.13605
	Neural8	NN_BReg_7H	0.779	0.10992	0.14422	0.10885	0.14014
	Neural3	NN_BReg_2H	0.777	0.11357	0.14830	0.10975	0.14150
	Neural2	NN_Transform	0.777	0.10119	0.12381	0.11095	0.14694
	Neural9	NN_BReg_8H	0.775	0.11937	0.15782	0.11095	0.14694
	Neural4	NN_BReg_3H	0.766	0.11621	0.16190	0.11432	0.16054
	Neural5	NN_BReg_4H	0.762	0.10789	0.14422	0.10892	0.13469
	Neural	NN_Cap&Floor	0.755	0.10867	0.14830	0.11554	0.15102

Fit Statistics
Model Selection based on Valid: Roc Index (_VAUR_)

Data Role=Valid

Statistics	Neural7	Neural6	Neural8	Neural3	Neural2	Neural9	Neural4	Neural5	Neural
Valid: Kolmogorov-Smirnov Statistic	0.48	0.47	0.47	0.46	0.47	0.45	0.42	0.42	0.38
Valid: Average Squared Error	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.12
Valid: Roc Index	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.76	0.76
Valid: Average Error Function	0.36	0.36	0.37	0.36	0.37	0.37	0.38	0.37	0.38
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.23	0.20	0.20	0.23	0.23	0.14	0.23	0.15	0.24
Valid: Cumulative Percent Captured Response	34.75	37.29	35.59	34.75	33.05	33.90	33.05	38.98	31.36
Valid: Percent Captured Response	12.71	14.41	13.56	13.56	11.86	11.86	11.02	15.25	12.71
Valid: Divisor for VASE	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
Valid: Error Function	534.03	526.66	536.95	535.80	547.74	542.02	551.68	541.69	559.09
Valid: Gain	245.11	270.36	253.53	245.11	228.28	236.69	228.28	287.20	211.44
Valid: Gini Coefficient	0.57	0.56	0.56	0.55	0.55	0.55	0.53	0.52	0.51
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.46	0.45	0.46	0.44	0.46	0.43	0.41	0.41	0.38
Valid: Kolmogorov-Smirnov Probability Cutoff	0.22	0.19	0.20	0.15	0.17	0.14	0.20	0.24	0.09
Valid: Cumulative Lift	3.45	3.70	3.54	3.45	3.28	3.37	3.28	3.87	3.11
Valid: Lift	2.53	2.86	2.69	2.69	2.36	2.36	2.19	3.03	2.53
Valid: Maximum Absolute Error	0.99	0.98	0.99	0.96	0.99	0.99	0.96	0.98	0.98
Valid: Misclassification Rate	0.14	0.14	0.14	0.14	0.15	0.15	0.16	0.13	0.15
Valid: Mean Squared Error	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.12
Valid: Sum of Frequencies	735.00	735.00	735.00	735.00	735.00	735.00	735.00	735.00	735.00
Valid: Root Average Squared Error	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.33	0.34
Valid: Cumulative Percent Response	55.41	59.46	56.76	55.41	52.70	54.05	52.70	62.16	50.00
Valid: Percent Response	40.54	45.95	43.24	43.24	37.84	37.84	35.14	48.65	40.54
Valid: Root Mean Squared Error	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.33	0.34
Valid: Sum of Squared Errors	158.16	156.74	160.01	161.33	163.10	163.09	168.05	160.11	169.85
Valid: Sum of Case Weights Times Freq	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
Valid: Number of Wrong Classifications	105.00	100.00	103.00	104.00	108.00	108.00	118.00	99.00	111.00

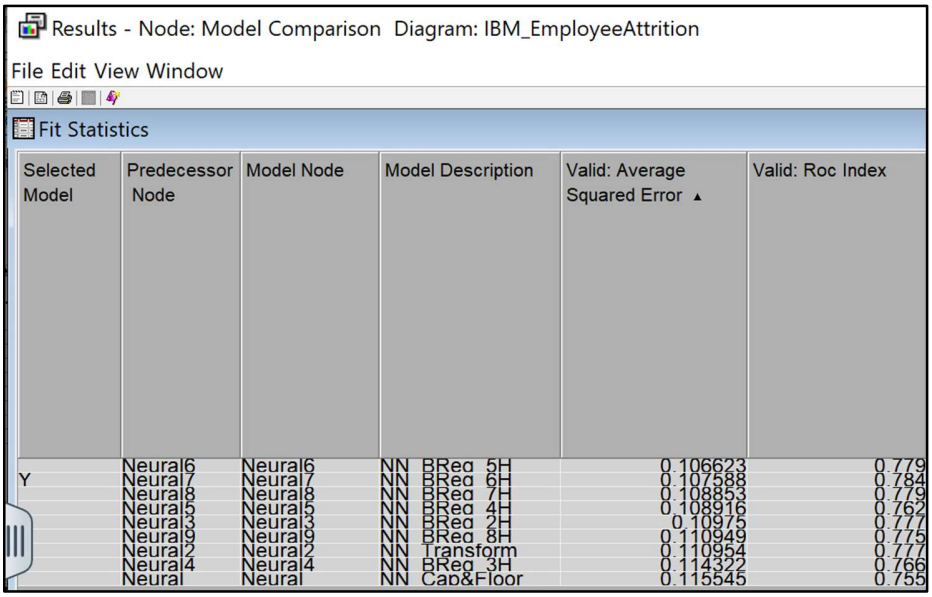
Event Classification Table

Model Selection based on Valid: Roc Index (_VAUR_)

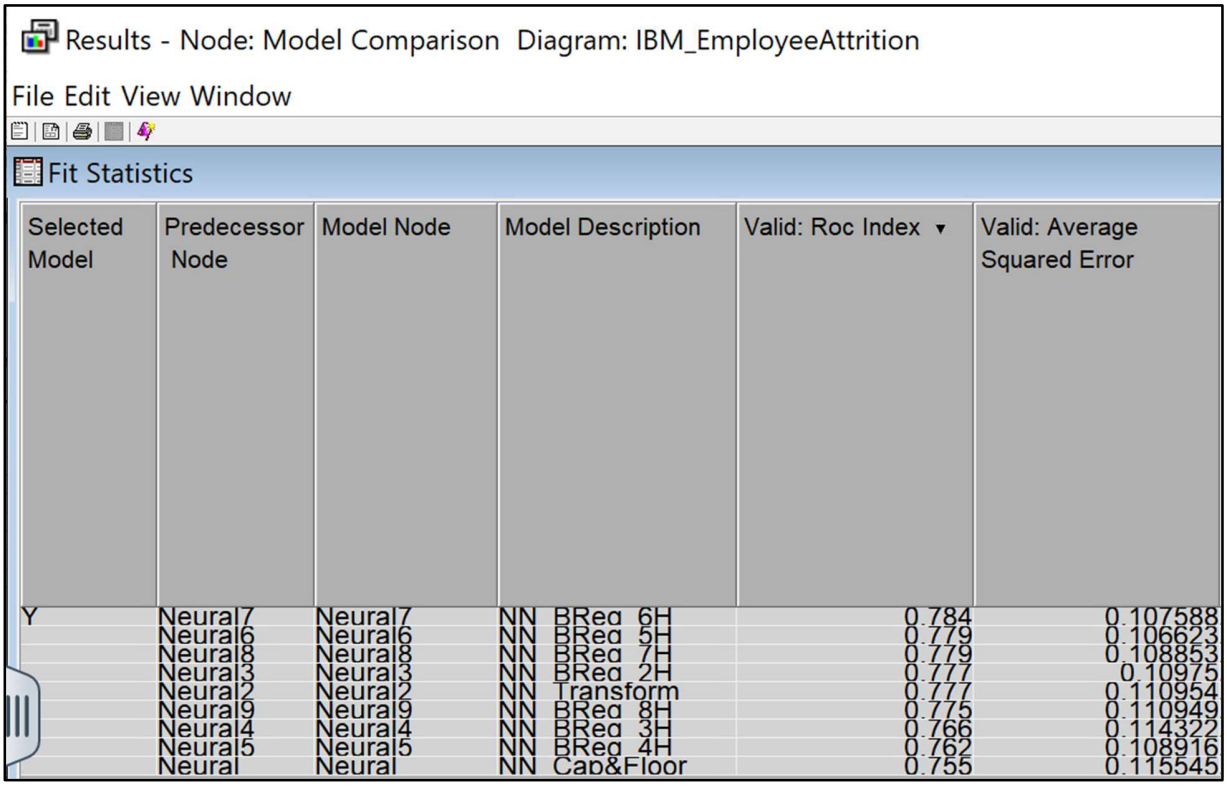
Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Neural	NN_Cap&Floor	TRAIN	Attrition		99	606	10	20
Neural	NN_Cap&Floor	VALIDATE	Attrition		93	599	18	25
Neural2	NN_Transform	TRAIN	Attrition		81	606	10	38
Neural2	NN_Transform	VALIDATE	Attrition		86	595	22	32
Neural9	NN_BReg_8H	TRAIN	Attrition		111	611	5	8
Neural9	NN_BReg_8H	VALIDATE	Attrition		105	614	3	13
Neural8	NN_BReg_7H	TRAIN	Attrition		88	598	18	31
Neural8	NN_BReg_7H	VALIDATE	Attrition		81	595	22	37
Neural7	NN_BReg_6H	TRAIN	Attrition		95	596	20	24
Neural7	NN_BReg_6H	VALIDATE	Attrition		89	601	16	29
Neural6	NN_BReg_5H	TRAIN	Attrition		99	604	12	20
Neural6	NN_BReg_5H	VALIDATE	Attrition		88	605	12	30
Neural5	NN_BReg_4H	TRAIN	Attrition		97	607	9	22
Neural5	NN_BReg_4H	VALIDATE	Attrition		91	609	8	27
Neural4	NN_BReg_3H	TRAIN	Attrition		119	616	0	0
Neural4	NN_BReg_3H	VALIDATE	Attrition		118	617	0	0
Neural3	NN_BReg_2H	TRAIN	Attrition		96	603	13	23
Neural3	NN_BReg_2H	VALIDATE	Attrition		86	599	18	32

Based on the result, Neural Network with 5 hidden units (NN_BReg_5H) is the best model in terms of ASE (0.106623), but Neural Network with 6 hidden units (NN_BReg_6H) is the best model in terms of ROC Index (0.784)

ASE:



ROC Index:



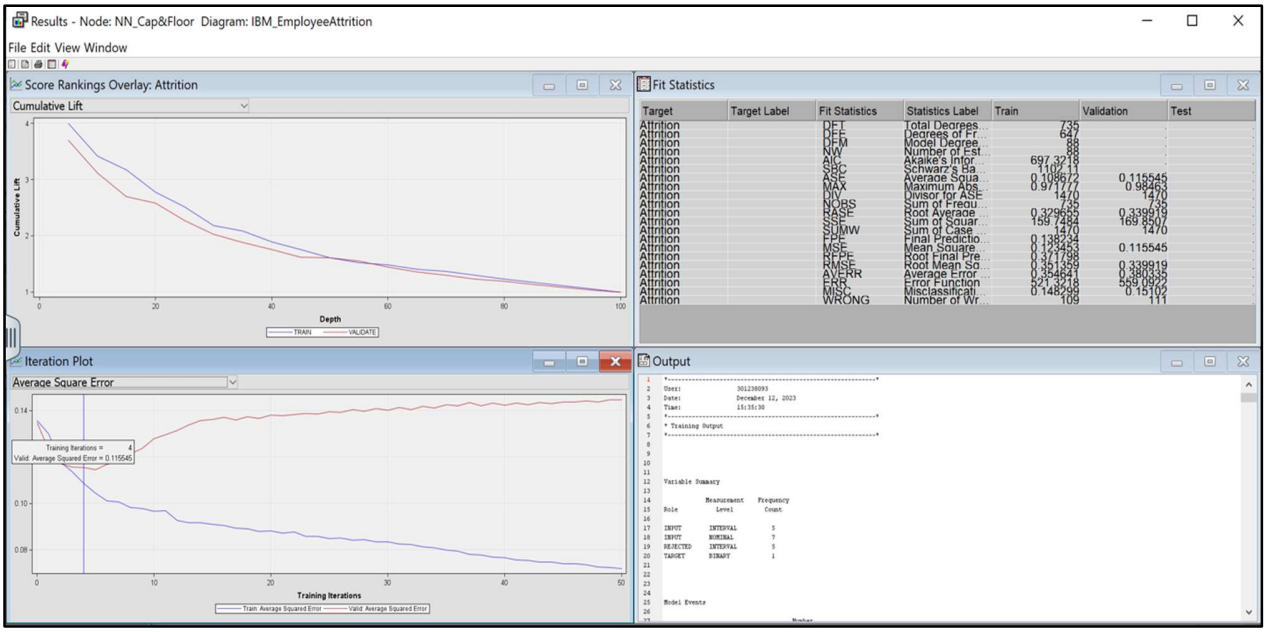
Selected Model	Predecessor Node	Model Node	Model Description	Valid: Roc Index	Valid: Average Squared Error
Y	Neural7	Neural7	NN_BReg_6H	0.784	0.107588
	Neural6	Neural6	NN_BReg_5H	0.779	0.106623
	Neural8	Neural8	NN_BReg_7H	0.779	0.108853
	Neural3	Neural3	NN_BReg_2H	0.777	0.10975
	Neural2	Neural2	NN_Transform	0.777	0.110954
	Neural9	Neural9	NN_BReg_8H	0.775	0.110949
	Neural4	Neural4	NN_BReg_3H	0.766	0.114322
	Neural5	Neural5	NN_BReg_4H	0.762	0.108916
	Neural	Neural	NN_Cap&Floor	0.755	0.115545

The iteration plot shows optimal validation average squared error occurring on iteration number for each model and each corresponding number of weights.

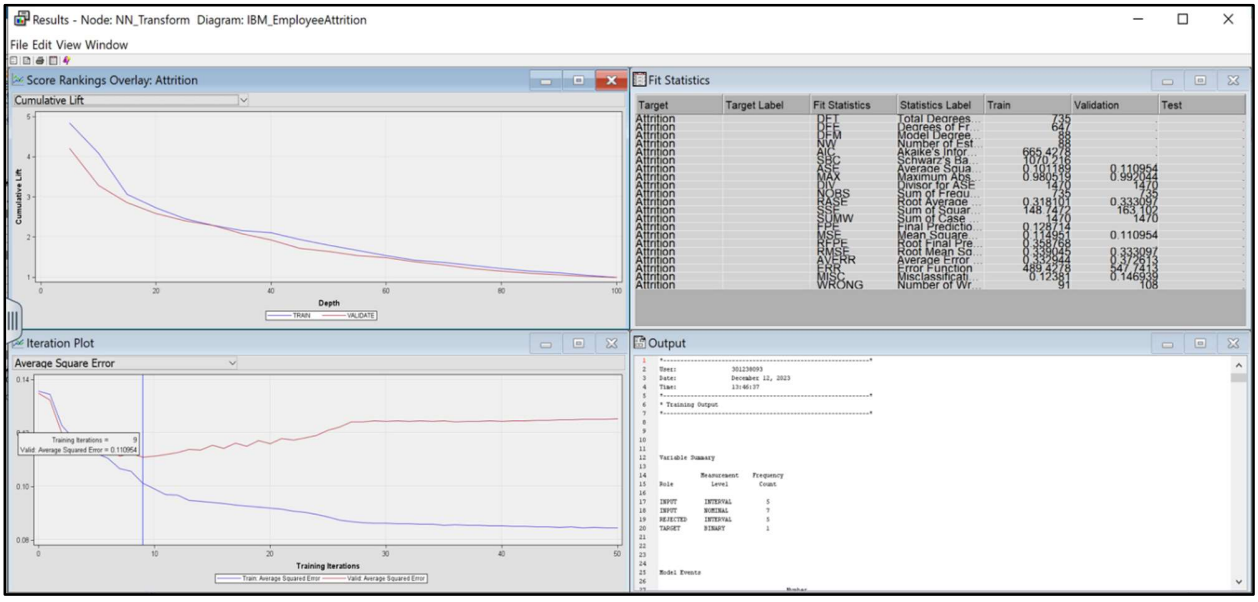
Neural Network Model	Iteration Number	Valid: Average Squared Error	ROC Index	Number of Estimated Weights
NN_Cap&Floor	4	0.115545	0.755	88
NN_Transform	9	0.110954	0.777	88
NN_BReg_2H	4	0.10975	0.777	41
NN_BReg_3H	3	0.114322	0.755	61
NN_BReg_4H	4	0.108916	0.762	81
NN_BReg_5H	7	0.106623	0.779	101
NN_BReg_6H	4	0.107588	0.784	121
NN_BReg_7H	5	0.108853	0.779	141
NN_BReg_8H	4	0.110954	0.775	161

Please see below output and iteration plot for all the neural network models:

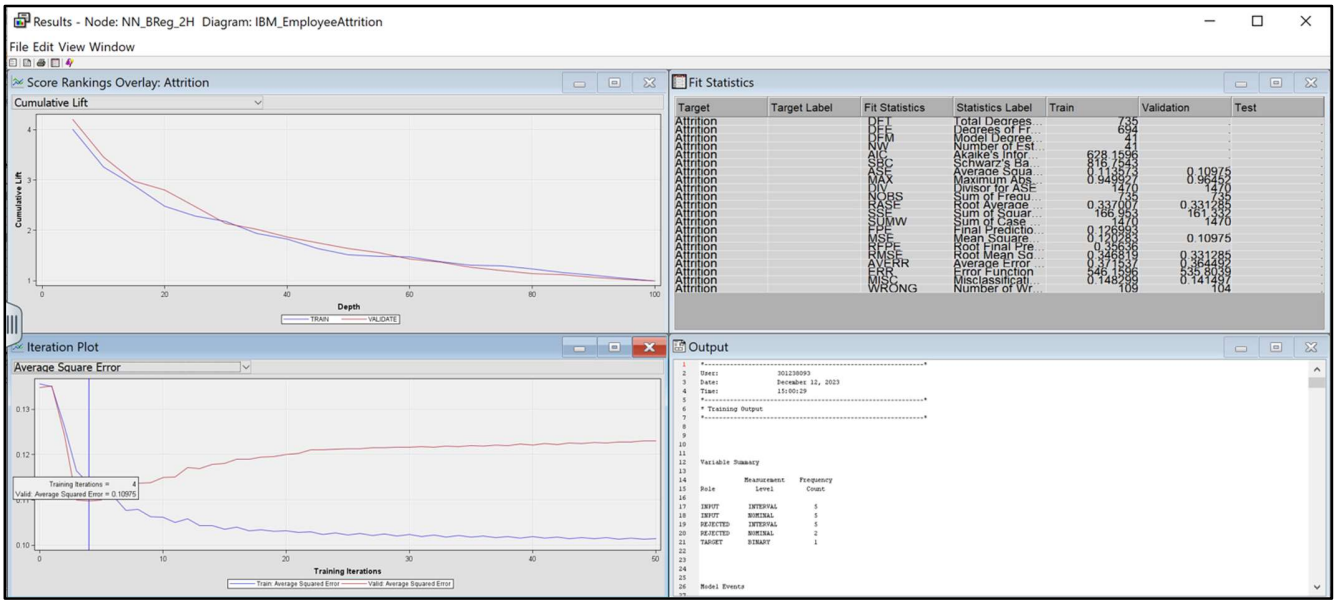
NN_Cap&Floor



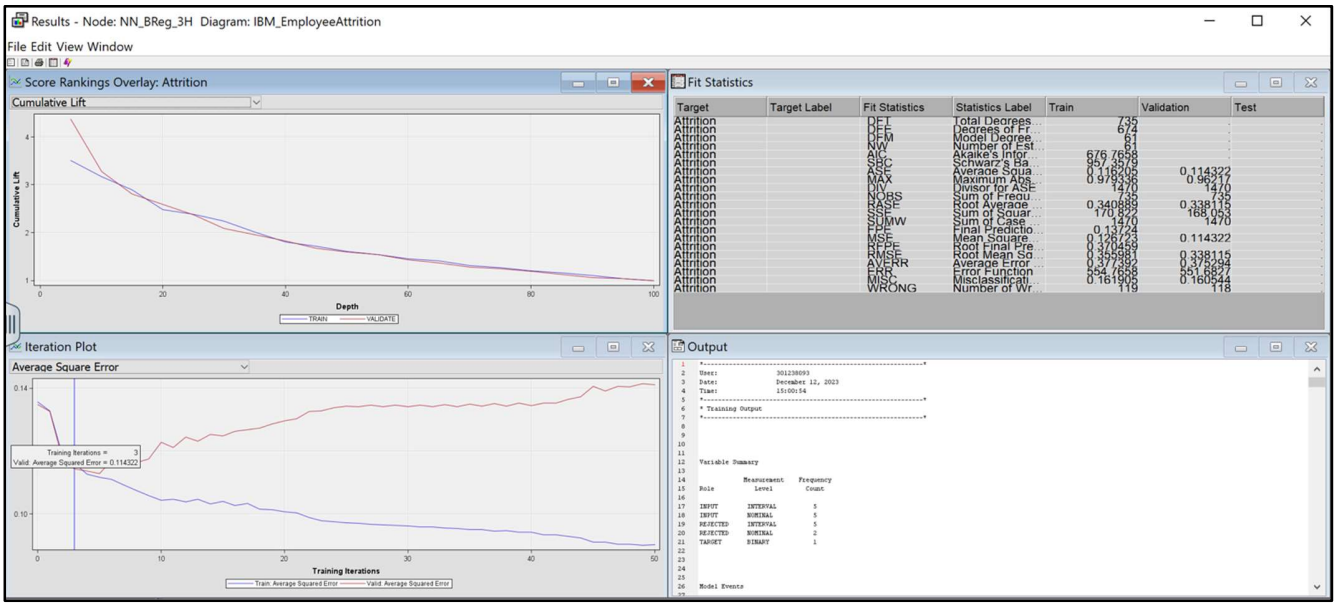
NN_Transform



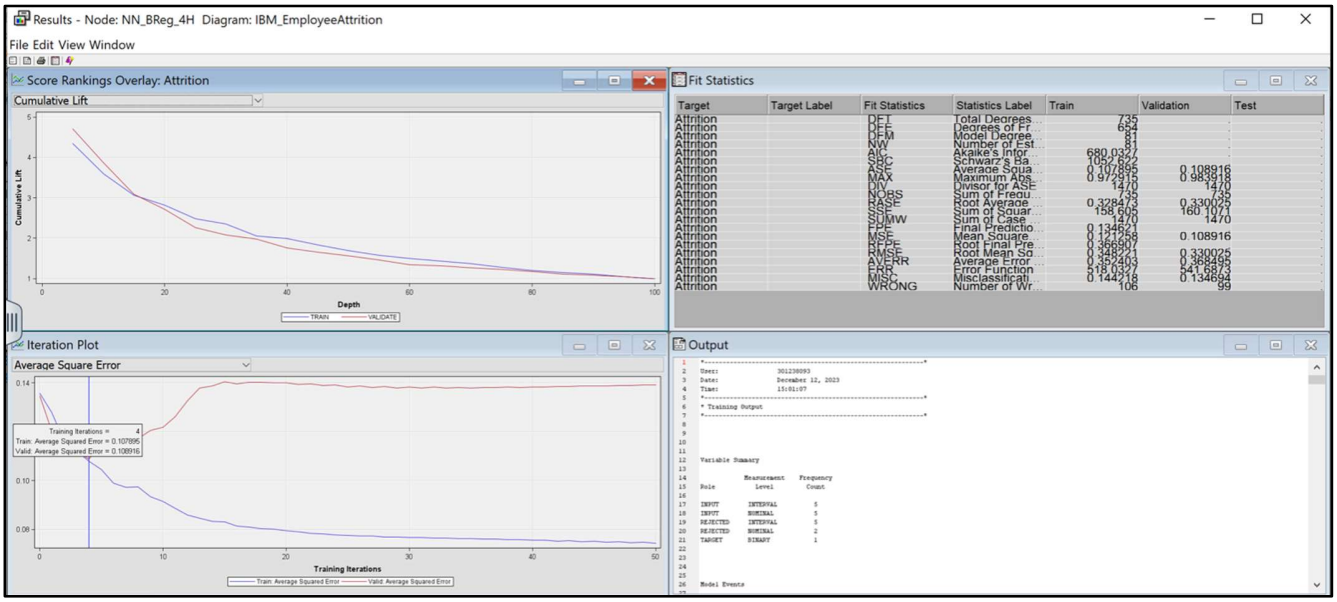
NN_BReg_2H



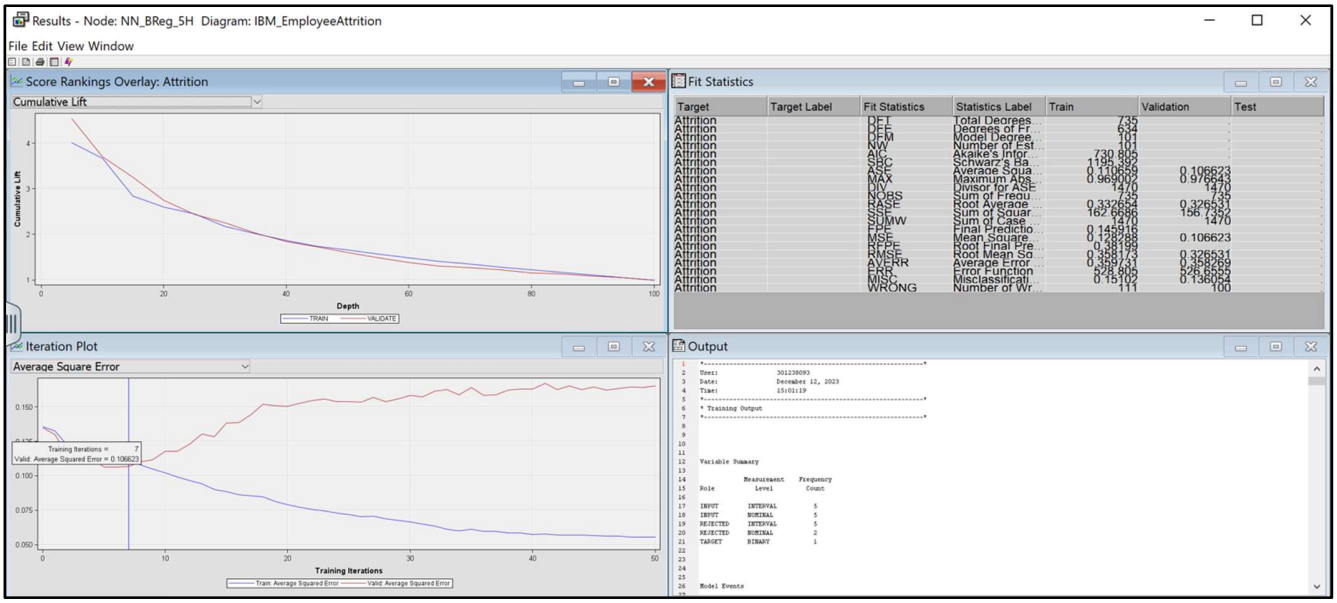
NN_BReg_3H



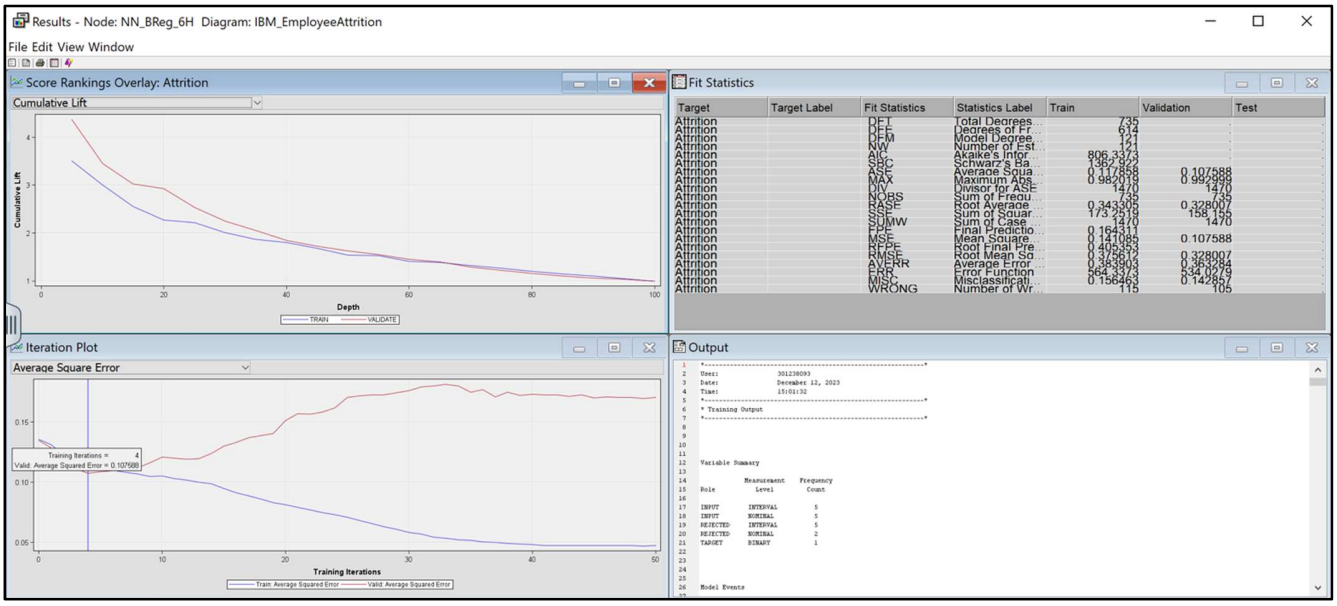
NN_BReg_4H



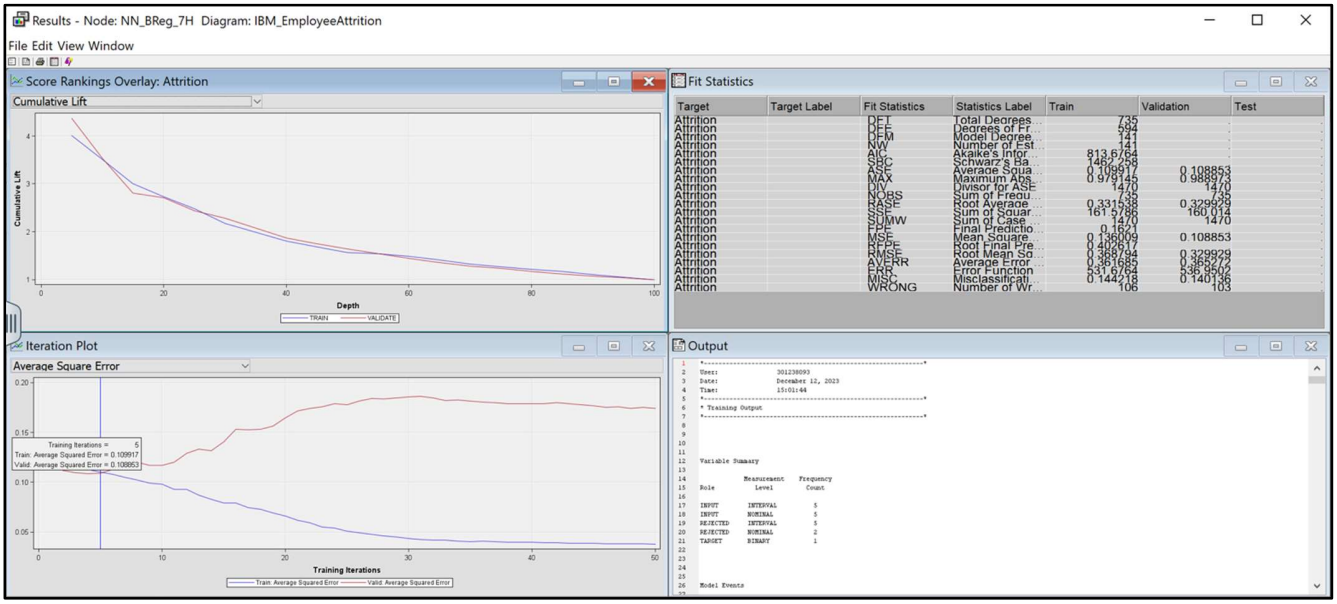
NN_BReg_5H



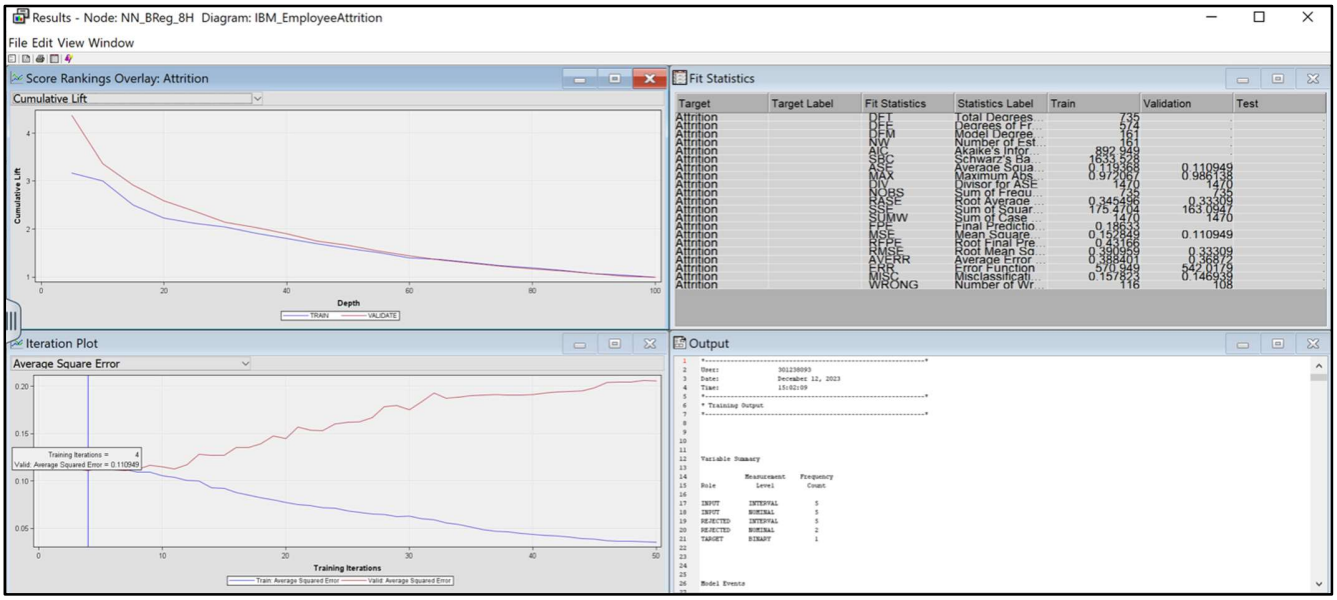
NN_BReg_6H



NN_BReg_7H



NN_BReg_8H

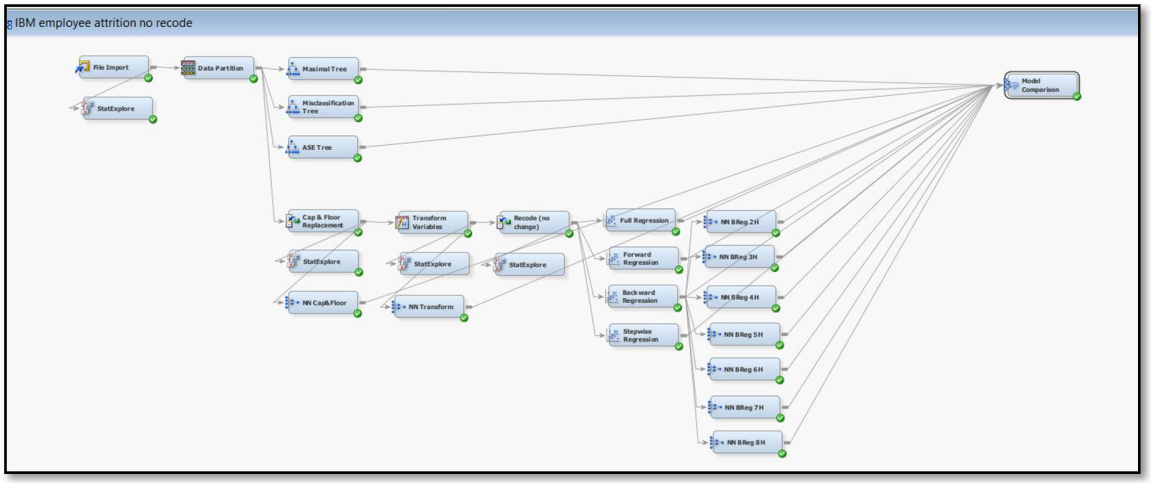

Results - Node: NN_BReg_8H Diagram: IBM_EmployeeAttrition
File Edit View Window
Score Rankings Overlay: Attrition
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
Iteration Plot
Average Square Error
Training Iterations
Output

Model Assessment

Comparison and Assessment of Models

The predictive modelling study allowed us to work on two models: the first model is with recode. For this model, we grouped the categorical variables into dummy variables to lessen the number of dimensions. The second model is without recode. In this model we left the categorical variable as is.

We connected all models to a Model Comparison node:



- After running the Model Comparison node, we got the results based on the ROC and Validation:

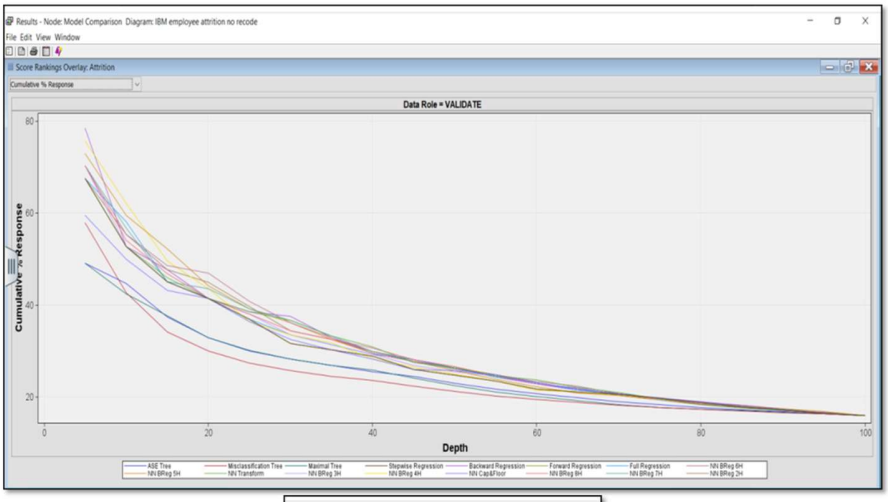
With Recode

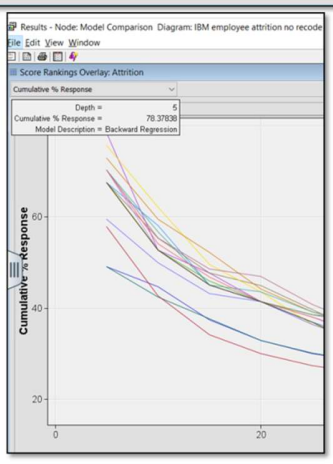
Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.682	0.123587
Backward Regression	0.777	0.109754
Forward Regression	0.74	0.115813
Full Regression	0.771	0.111461
Maximal Tree	0.671	0.127063
Misclassification Tree	0.646	0.125932
NN BReg 2H	0.775	0.111829
NN BReg 3H	0.783	0.110521
NN BReg 4H	0.773	0.112225
NN BReg 5H	0.762	0.113858
NN BReg 6H	0.777	0.109461
NN BReg 7H	0.767	0.111574
NN BReg 8H	0.767	0.110065
NN Cap&Floor	0.755	0.115545
NN Recode	0.759	0.114041
NN Transform	0.777	0.110954
Stepwise Regression	0.724	0.118873

Without Recode

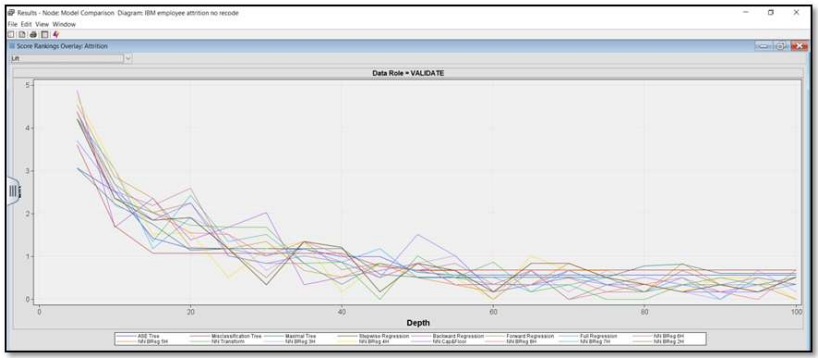
Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.682	0.123587
Backward Regression	0.776	0.10897
Forward Regression	0.751	0.113473
Full Regression	0.767	0.110665
Maximal Tree	0.671	0.127063
Misclassification Tree	0.646	0.125932
NN BReg 2H	0.777	0.10975
NN BReg 3H	0.766	0.114322
NN BReg 4H	0.762	0.108916
NN BReg 5H	0.779	0.106623
NN BReg 6H	0.784	0.107588
NN BReg 7H	0.779	0.108853
NN BReg 8H	0.775	0.110949
NN Cap&Floor	0.755	0.115545
NN Transform	0.777	0.110954
Stepwise Regression	0.751	0.113473

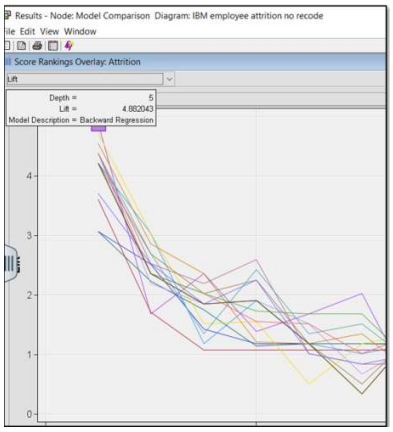
- The model without recode yielded better results with higher ROC value and lower validation ASE.
- With ROC index chosen as selection criterion, the best model from the without recode was the **Neural Network 6 hidden units** with an **ROC value of 0.784**.
- The model that has the best **cumulative response rate** for the best 5% (depth of 5) of scores is the **Backward Regression model with 78.38%**.





- Based on the figures below, the model with the best lift for the best 5% of the scores is still the Backward Regression model with a **lift of 4.882043**. It means that with depth of 5, we get 24.4% response in the future.





Prediction Type	Decision	Ranking	Estimates
Model Description ▲	Valid: Misclassification Rate	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.161905	0.682	0.123587
Backward Regression	0.131973	0.776	0.10897
Forward Regression	0.144218	0.751	0.113473
Full Regression	0.142857	0.767	0.110665
Maximal Tree	0.161905	0.671	0.127063
Misclassification Tree	0.153741	0.646	0.125932
NN BReg 2H	0.141497	0.777	0.10975
NN BReg 3H	0.160544	0.766	0.114322
NN BReg 4H	0.134694	0.762	0.108916
NN BReg 5H	0.136054	0.779	0.106623
NN BReg 6H	0.142857	0.784	0.107588
NN BReg 7H	0.140136	0.779	0.108853
NN BReg 8H	0.146939	0.775	0.110949
NN Cap&Floor	0.15102	0.755	0.115545
NN Transform	0.146939	0.777	0.110954
Stepwise Regression	0.144218	0.751	0.113473

Based on the assessment measure per prediction type, the following models were identified as best:

- **Decision**- Backward Regression for having the lowest Misclassification rate. It is the most accurate in matching decision with outcome.
- **Ranking**- Neural Network 6H for having the highest ROC index. It is the most accurate in ordering primary and secondary outcomes.
- **Estimates** - Neural Network 5H for having the lowest ASE. It has the lowest variance between the target and estimate.

Conclusion

In order to answer this study's research question on the factors affecting employee attrition at IBM, we employed a predictive modelling approach. We used the three predictive models namely the decision tree, regression and neural networks. In the course of running our models, we worked on two simulations: the first one involved consolidating levels of non-numeric inputs using the Replacement Editor for the categorical variables **Education, Education Field,** and **Marital Status** in order to reduce the dimensionality of our dataset. In the second simulation, we did not consolidate the categorical variables. After running the model assessment, it turned out that the second simulation, the model without any recode, was the best model based on the ROC value. Despite the high degree of dimensionality due to non-recode of categorical levels, it turned out to be the best option, for it resulted in a better model fit.

The best model that yielded the highest ROC was Neural Network with 6 hidden units. While neural network models are a natural extension of a regression model, it faces interpretability challenges. Thus, for the purpose of interpreting the results in relation to the factors affecting employee attrition, we focused our attention to the next best non-NN which is the Backward Regression. Based on the odds-ratio estimates, we conclude that the following demographic, experience, tenure, department and perception based on work environment characteristics affect employee attrition:

- **Single**- Single employees are more likely to resign by **75%** and **53%** vs. Divorced and Married employees, respectively.
- **Young** - Younger employees are more likely to resign by **3%.**
- **Distance** - An increase in the kilometer distance between employees' residence and workplace, multiplied to a factor of 2.74, increases the chance of resignation by **32%.**
- **Income** - An increase in employees' monthly income, multiplied to a factor of 2.74, decreases the probability of resignation by **47%**
- **Experience** - The higher the number of companies previously worked in by employees, multiplied to a factor of 2.74, the higher the chance of resignation by **64%.**
- **Tenure** - The longer the years of employees' tenure multiplied to a factor of 2.74, the lesser likelihood of resignation by **28%.**

- **Human Resources Department** - Employees in Sales Department are more likely to resign by **48%** compared to Research and Development**;** however, they are more likely to stay by **43%**compared to Human Resources.
- **Perception**:
    - With respect to satisfaction surveys, low degree of satisfaction in terms of work environment and job satisfaction increases the probability of attrition.
    - In the aspect of Work Life Balance, employees who rate the company at 1(lowest), are the most likely to resign by **88%**. However, employees that rate the company with 2 & 3(better& good) are less likely to leave the company by **8%** and **35%** respectively compared to those with ratings of 4(best).

This study gave us significant insights about why employees at IBM leave their jobs. Organizations similar to IBM may utilize these data to implement targeted retention strategies, enhance employee satisfaction, and lower attrition.

Recommendations

Given the insights gleaned from the predictive modelling exercise that we conducted, we recommend the following for IBM:

- Since most of those who resign are new hires and young employees, it could be worthwhile for IBM to invest in professional development programs that will enhance the skills of their employees.
- Investigate the drivers of high attrition rate within Human Resource department. Address the identified drivers and implement effective strategies to increase employee retention.
- Extensively review the candidate's retention history upon hiring; especially young and single applicants, and those with robust experience in the industry
- It is also equally important for the IBM management to ensure that the company has conducive working conditions, effective job structure, and promotes a culture that values work life balance.
- Another recommendation is for IBM to recognize the efforts of their employees through giving incentives, whether monetary or other ways, that could boost their morale and consider staying in the company.

References

Conchada, M., Doña, M. and Francisco, K. (2023). Analysis on factors affecting IBM employee attrition. Analytic Lifecycle Management Final Paper.

Kaggle. (2023). IBM Attrition Dataset. https://www.kaggle.com/datasets/yasserh/ibm-attrition-dataset/code