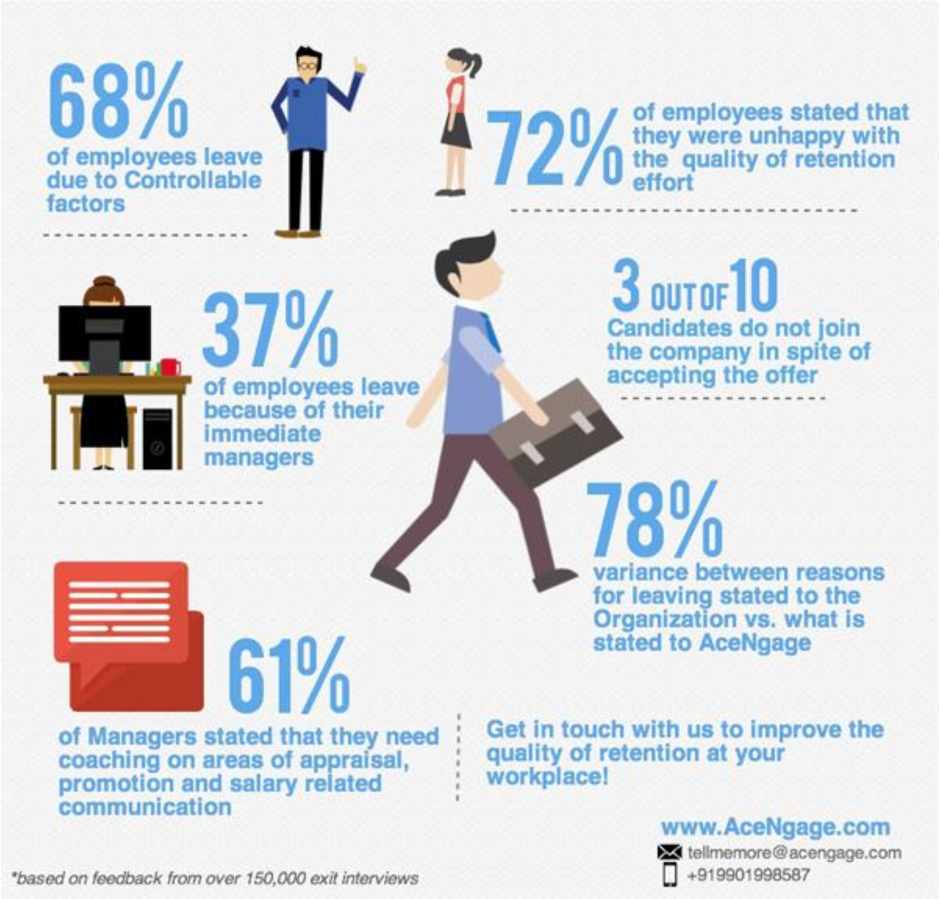
Business Analytics with SAS Project on Finding Attrition rate in a company



68% of employees leave due to Controllable factors

72% of employees stated that they were unhappy with the quality of retention effort

37% of employees leave because of their immediate managers

3 OUT OF 10 Candidates do not join the company in spite of accepting the offer

78% variance between reasons for leaving stated to the Organization vs. what is stated to AceNgage

61% of Managers stated that they need coaching on areas of appraisal, promotion and salary related communication

Get in touch with us to improve the quality of retention at your workplace!

www.AceNgage.com

tellmemore@acengage.com

+919901998587

*based on feedback from over 150,000 exit interviews

Image Source: https://www.talentlyft.com/en/resources/what-is-attrition

Instructor: Dr. James Zhang

<u>Project Members (Group 3)</u>

Venkata Sarath Chandra Muktevi

Vinay Manideep Cheviti

Vamshi Enishetty

Harshavardhan Yarlagadda

Contents

1. Data Mining Objective and Motivation:

1.1 Objective:

To predict the likelihood of a person leaving a company (Attrition rate), to analyze the elements which are contributing towards this factor, and to analyze the descriptive pattern among data set.

1.2 Motivation:

Employee attrition means the reduction of workforce in a company through normal process, like resignation or retirement. In recent years this has become a serious problem in the Industries. In one of the studies conducted by 'FurstPerson' the attrition rate has a financial toll on companies. It states that if they lose an employee, they have to suffer a loss anywhere between $1500 to $16500. To avoid it, companies are analyzing what are the key factors and circumstances that are leading to this cause. Differences in pay scale, Level of job satisfaction, involvement in job, Total working hours, Distance from home, work life balance, years with current manager, Education field, Total working years, work Environment are some of the factors that are leading to the Employee attrition rate. If companies can find out effective reasons why an employee likes to leave the industry they can avoid it by taking necessary actions which eventually decrease their financial burden in Employee replacement.

1.3 Executive Summary:

We took a Third-party dataset and found out the important factors contributing to Employee attrition. We also measured the best model or classifier which helps in predicting the attrition rate in a company. There are total of 1470 rows and 35 columns in our data set.

1.4 Data set:

In this project, we will be working on the second-hand dataset named 'IBM HR Analytics Employee Attrition and Performance' obtained from -

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

The data consists of Attributes of employees, there are 35 attributes related to an employee.

- TIME: Yearsatcompany, Yearsincurrentrole, YearsSinceLastPromotion, YearsWithCurrManager
- EMPLOYEE DETAILS: Age, Education, EducationField, Gender, WorkLifeBalance, DistanceFromHome, JobSatisfaction, MaritalStatus, NumCompaniesWorked
- INCOME: MonthlyIncome, MonthlyRate, DailyRate, HourlyRate, StockOptionLevel, PercentSalaryHike
- JOB RELATED: PerformanceRating, JobInvolvement, JobLevel, JobRole, Department, EnvironmentSatisfaction, TotalWorkingYears, TrainingTimesLastYear

Age	Age of employee	INTERVAL
Attrition	Weather employee quits or not	BINARY
BusinessTravel	Frequency of Travel	NOMINAL
DailyRate	The amount of money employees is paid per day	INTERVAL
Department	Name of the department employee work	NOMINAL
DistanceFromHome	Commute Distance	INTERVAL
Education	Education level	INTERVAL
EducationField	Field of education	NOMINAL
EmployeeNumber	Actual Departure Time (local time: hh mm)	NOMINAL
EnvironmentSatisfaction	Environment Satisfaction	NOMINAL
Gender	Gender	NOMINAL
HourlyRate	the amount of money employees are paid per hour	INTERVAL
JobInvolvement	Employee involvement in assigned task	NOMINAL
JobLevel	Job Level	NOMINAL
JobRole	Job Role	NOMINAL
JobSatisfaction	Job Satisfaction	NOMINAL

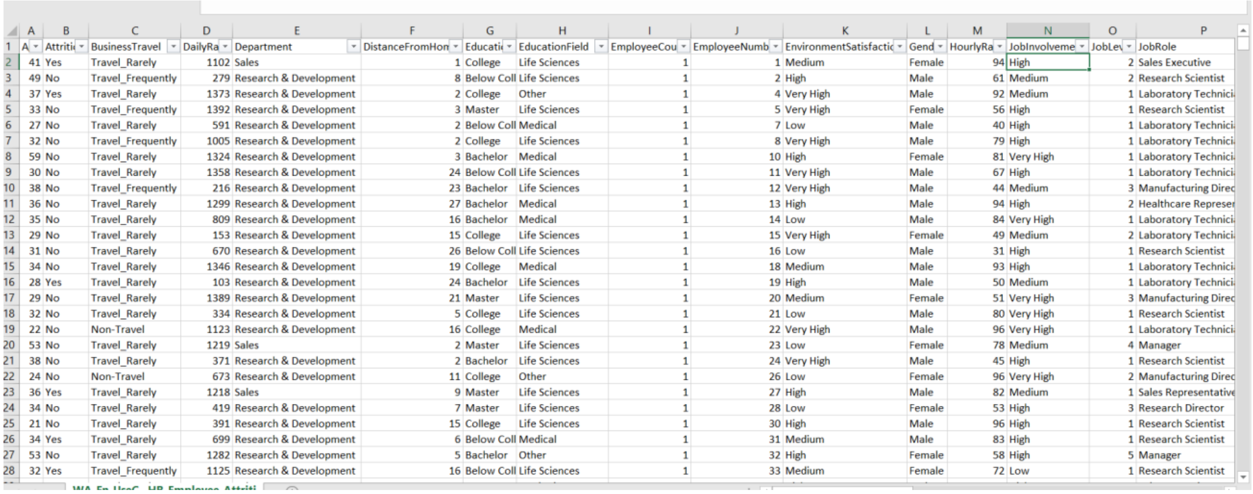
MaritalStatus	Marital Status	NOMINAL
MonthlyIncome	the amount of money employees are paid per month	INTERVAL
NumCompaniesWorked	Number of companies previously worked	INTERVAL
Over Time	Over Time	NOMINAL
PercentSalaryHike	Salary hike	INTERVAL
PerformanceRating	Performance Rating	INTERVAL
StandardHours	Standard Hours	INTERVAL
StockOptionLevel	Stock Option Level	INTERVAL
TotalWorkingYears	Number of year employee worked in his total career	INTERVAL
TrainingTimesLastYear	Times a particular employee trained	INTERVAL
WorkLifeBalance	Work Life Balance	INTERVAL
YearsAtCompany	Year worked in the company	INTERVAL
YearsInCurrentRole	Years worked in current role	INTERVAL
YearsSinceLastPromotion	Years since last promotion	INTERVAL
YearsWithCurrManager	Team worked with current manager	INTERVAL

1.5 Data Preprocessing:

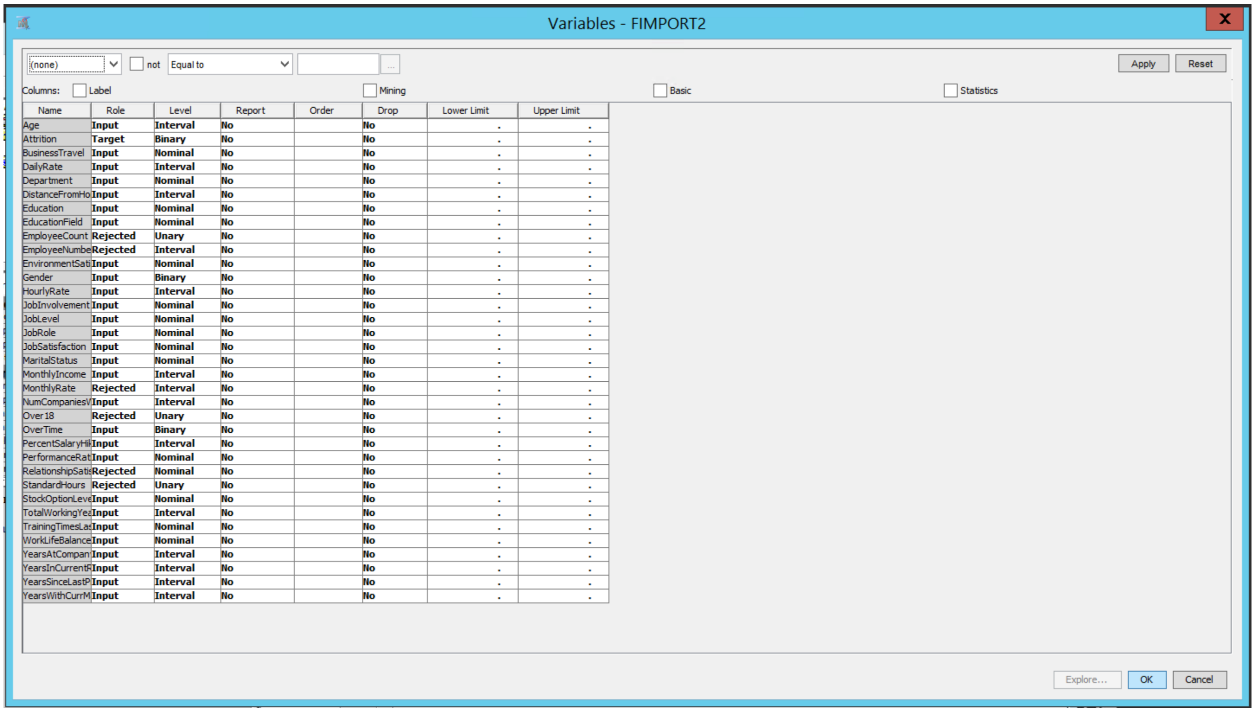
Following steps are performed for Data Preprocessing –

1. We changed the data format of Education Column from 1,2,3,4,5 to 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
2. We changed the data format of Environment Satisfaction from 1,2,3,4,5 to 1 'Low' 2 'Medium' 3 'High' 4 'Very High'.
3. We changed the data format of Job Involvement from 1,2,3,4,5 to 1 'Low' 2 'Medium' 3 'High' 4 'Very High'.

4. Removed Columns: Employee Number (Not relevant), Standard Hours (unary data), over 18 (Unary data), Employee Count (Unary Data), Monthly Rate (Field Context is Not Discussed), Relationship Satisfaction (Field Context is Not Discussed), Daily Rate (Field Context is Not Discussed) which seems to be out of context in finding Attrition Rate.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	A	Attriti	BusinessTravel	DailyRa	Department	DistanceFromHon	Educati	EducationField	EmployeeCou	EmployeeNumb	EnvironmentSatisfactic	Gend	HourlyRa	JobInvolveme	JobLev	JobRole
2	41	Yes	Travel_Rarely	1102	Sales	1	College	Life Sciences	1	1	Medium	Female	94	High	2	Sales Executive
3	49	No	Travel_Frequently	279	Research & Development	8	Below Coll	Life Sciences	1	2	High	Male	61	Medium	2	Research Scientist
4	37	Yes	Travel_Rarely	1373	Research & Development	2	College	Other	1	4	Very High	Male	92	Medium	1	Laboratory Technici
5	33	No	Travel_Frequently	1392	Research & Development	3	Master	Life Sciences	1	5	Very High	Female	56	High	1	Research Scientist
6	27	No	Travel_Rarely	591	Research & Development	2	Below Coll	Medical	1	7	Low	Male	40	High	1	Laboratory Technici
7	32	No	Travel_Frequently	1005	Research & Development	2	College	Life Sciences	1	8	Very High	Male	79	High	1	Laboratory Technici
8	59	No	Travel_Rarely	1324	Research & Development	3	Bachelor	Medical	1	10	High	Female	81	Very High	1	Laboratory Technici
9	30	No	Travel_Rarely	1358	Research & Development	24	Below Coll	Life Sciences	1	11	Very High	Male	67	High	1	Laboratory Technici
10	38	No	Travel_Frequently	216	Research & Development	23	Bachelor	Life Sciences	1	12	Very High	Male	44	Medium	3	Manufacturing Direc
11	36	No	Travel_Rarely	1299	Research & Development	27	Bachelor	Medical	1	13	High	Male	94	High	2	Healthcare Represer
12	35	No	Travel_Rarely	809	Research & Development	16	Bachelor	Medical	1	14	Low	Male	84	Very High	1	Laboratory Technici
13	29	No	Travel_Rarely	153	Research & Development	15	College	Life Sciences	1	15	Very High	Female	49	Medium	2	Laboratory Technici
14	31	No	Travel_Rarely	670	Research & Development	26	Below Coll	Life Sciences	1	16	Low	Male	31	High	1	Research Scientist
15	34	No	Travel_Rarely	1346	Research & Development	19	College	Medical	1	18	Medium	Male	93	High	1	Laboratory Technici
16	28	Yes	Travel_Rarely	103	Research & Development	24	Bachelor	Life Sciences	1	19	High	Male	50	Medium	1	Laboratory Technici
17	29	No	Travel_Rarely	1389	Research & Development	21	Master	Life Sciences	1	20	Medium	Female	51	Very High	3	Manufacturing Direc
18	32	No	Travel_Rarely	334	Research & Development	5	College	Life Sciences	1	21	Low	Male	80	Very High	1	Research Scientist
19	22	No	Non-Travel	1123	Research & Development	16	College	Medical	1	22	Very High	Male	96	Very High	1	Laboratory Technici
20	53	No	Travel_Rarely	1219	Sales	2	Master	Life Sciences	1	23	Low	Female	78	Medium	4	Manager
21	38	No	Travel_Rarely	371	Research & Development	2	Bachelor	Life Sciences	1	24	Very High	Male	45	High	1	Research Scientist
22	24	No	Non-Travel	673	Research & Development	11	College	Other	1	26	Low	Female	96	Very High	2	Manufacturing Direc
23	36	Yes	Travel_Rarely	1218	Sales	9	Master	Life Sciences	1	27	High	Male	82	Medium	1	Sales Representative
24	34	No	Travel_Rarely	419	Research & Development	7	Master	Life Sciences	1	28	Low	Female	53	High	3	Research Director
25	21	No	Travel_Rarely	391	Research & Development	15	College	Life Sciences	1	30	High	Male	96	High	1	Research Scientist
26	34	Yes	Travel_Rarely	699	Research & Development	6	Below Coll	Medical	1	31	Medium	Male	83	High	1	Research Scientist
27	53	No	Travel_Rarely	1282	Research & Development	5	Bachelor	Other	1	32	High	Female	58	High	5	Manager
28	32	Yes	Travel_Frequently	1125	Research & Development	16	Below Coll	Life Sciences	1	33	Medium	Female	72	Low	1	Research Scientist



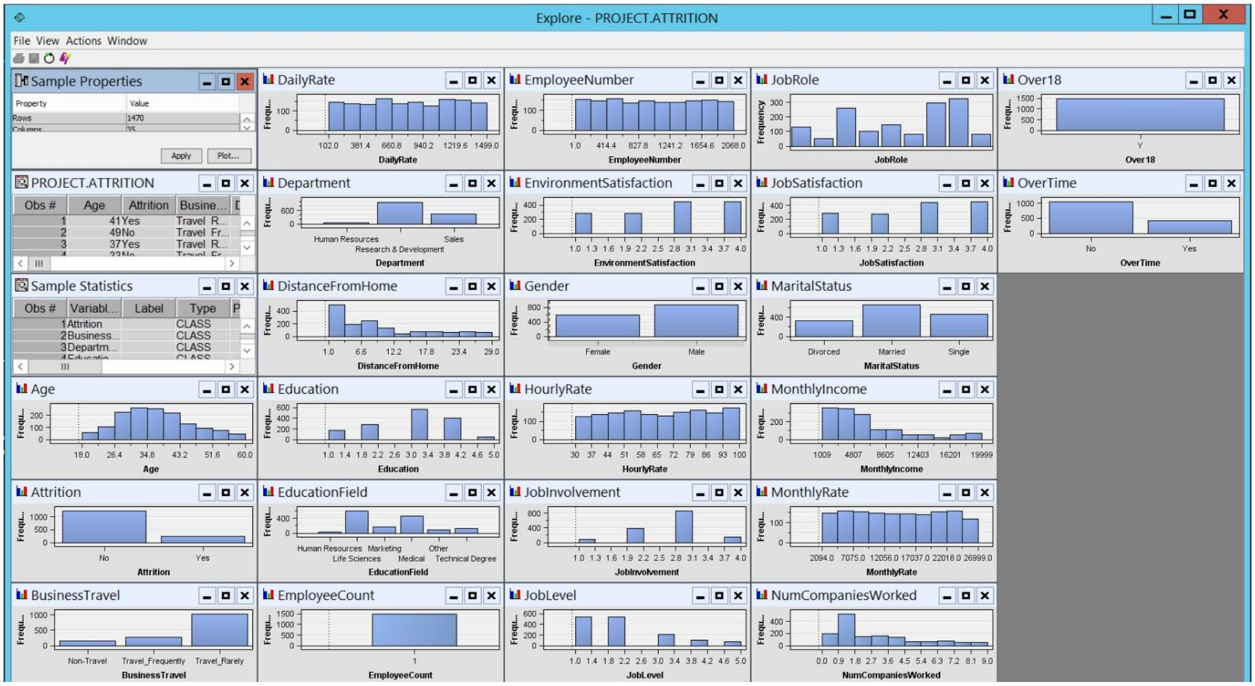
Variables - FIMPORT2

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Attrition	Target	Binary	No		No	.	.
BusinessTravel	Input	Nominal	No		No	.	.
DailyRate	Input	Interval	No		No	.	.
Department	Input	Nominal	No		No	.	.
DistanceFromHo	Input	Interval	No		No	.	.
Education	Input	Nominal	No		No	.	.
EducationField	Input	Nominal	No		No	.	.
EmployeeCount	Rejected	Unary	No		No	.	.
EmployeeNumbe	Rejected	Interval	No		No	.	.
EnvironmentSati	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
HourlyRate	Input	Interval	No		No	.	.
JobInvolvement	Input	Nominal	No		No	.	.
JobLevel	Input	Nominal	No		No	.	.
JobRole	Input	Nominal	No		No	.	.
JobSatisfaction	Input	Nominal	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
MonthlyIncome	Input	Interval	No		No	.	.
MonthlyRate	Rejected	Interval	No		No	.	.
NumCompanies\	Input	Interval	No		No	.	.
Over18	Rejected	Unary	No		No	.	.
OverTime	Input	Binary	No		No	.	.
PercentSalaryHi	Input	Interval	No		No	.	.
PerformanceRat	Input	Nominal	No		No	.	.
RelationshipSati	Rejected	Nominal	No		No	.	.
StandardHours	Rejected	Unary	No		No	.	.
StockOptionLeve	Input	Nominal	No		No	.	.
TotalWorkingYea	Input	Interval	No		No	.	.
TrainingTimesLa	Input	Nominal	No		No	.	.
WorkLifeBalance	Input	Nominal	No		No	.	.
YearsAtCompan	Input	Interval	No		No	.	.
YearsInCurrentR	Input	Interval	No		No	.	.
YearsSinceLastP	Input	Interval	No		No	.	.
YearsWithCurrM	Input	Interval	No		No	.	.

Before Preprocessing



Explore - PROJECT.ATTRITION

Sample Properties

Property	Value
Rows	1470
Columns	35

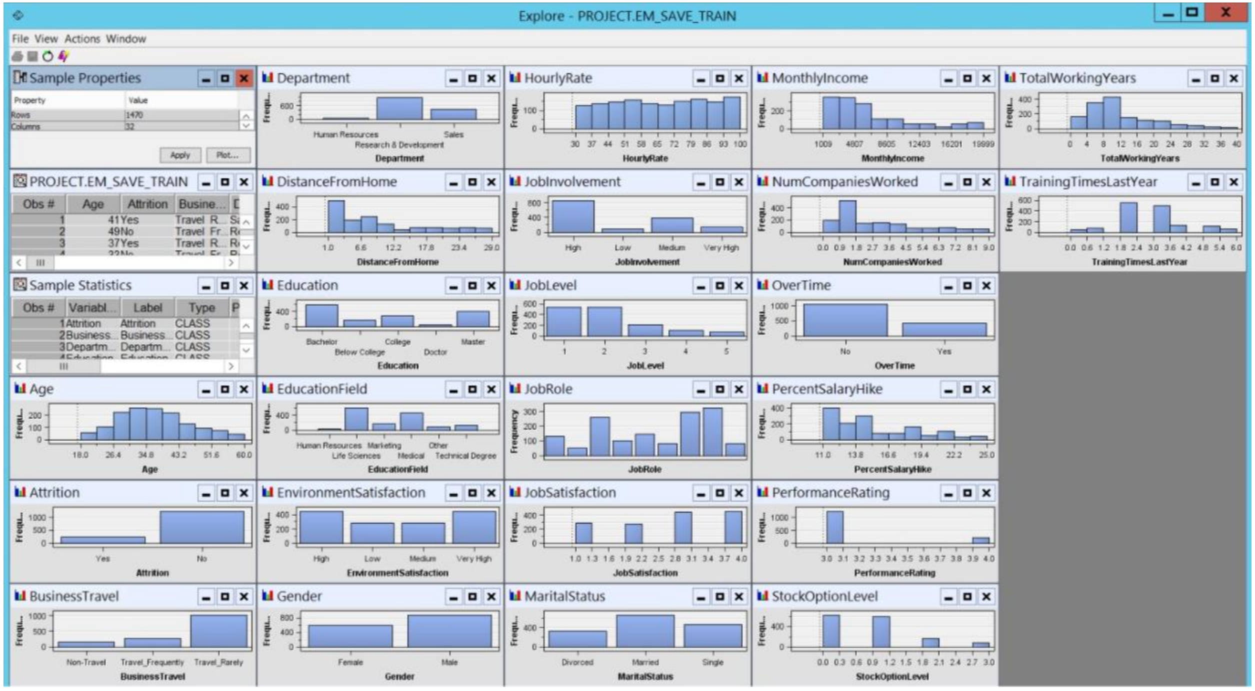
PROJECT.ATTRITION

Obs #	Age	Attrition	Busine...
1	41	Yes	Travel R...
2	49	No	Travel Fr...
3	37	Yes	Travel R...

Sample Statistics

Obs #	Variabl...	Label	Type
1	Attrition		CLASS
2	Business...		CLASS
3	Departm...		CLASS

After Preprocessing



Explore - PROJECT.EM_SAVE_TRAIN

Sample Properties

Property	Value
Rows	1470
Columns	32

PROJECT.EM_SAVE_TRAIN

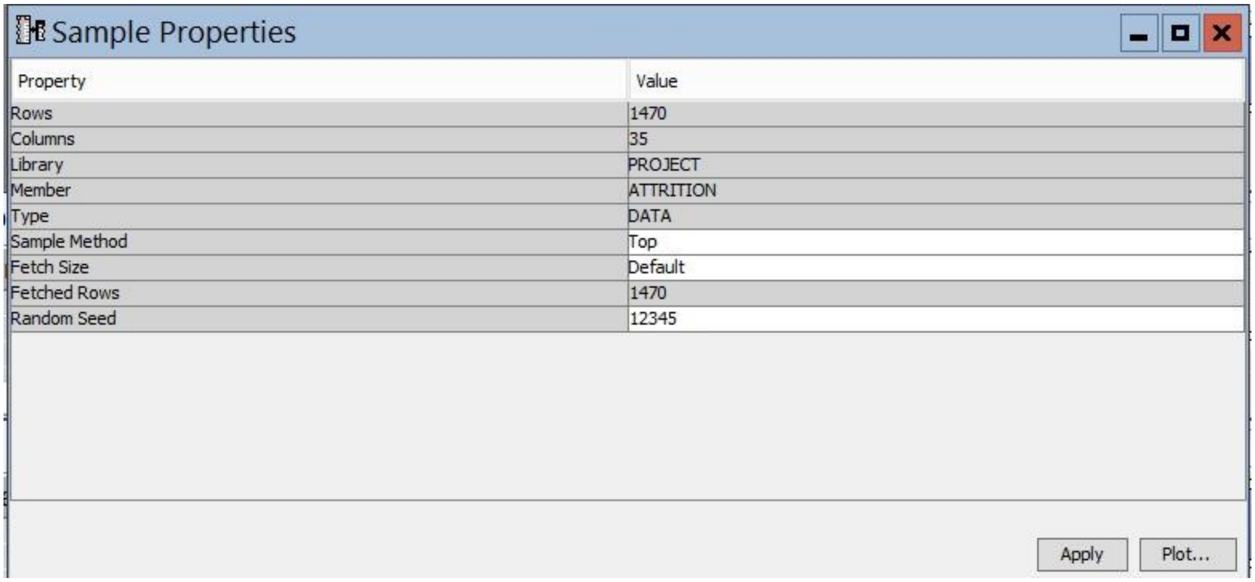
Obs #	Age	Attrition	Busine...
1	41	Yes	Travel R...
2	49	No	Travel Fr...
3	37	Yes	Travel R...

Sample Statistics

Obs #	Variabl...	Label	Type
1	Attrition	Attrition	CLASS
2	Business...	Business...	CLASS
3	Departm...	Departm...	CLASS

5. Data summary

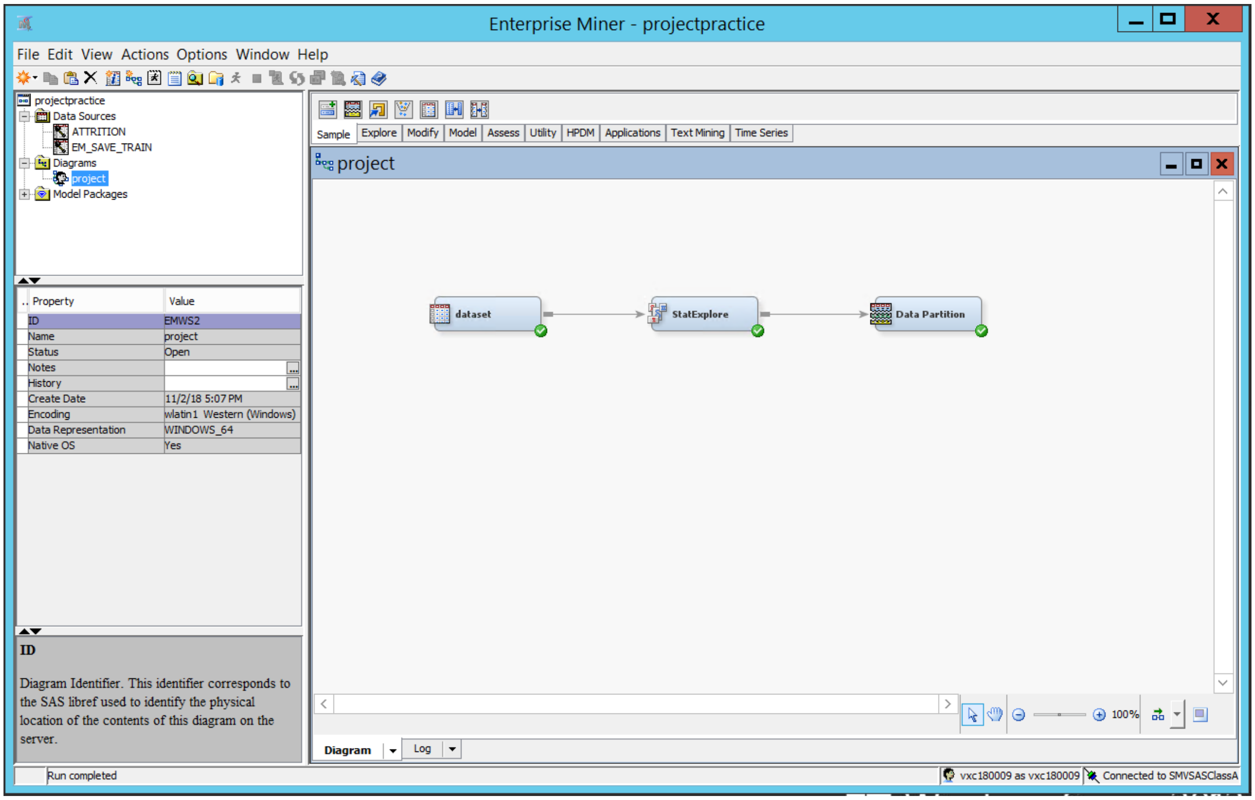
Summarization of the dataset. We can see that there are total of 1470 rows and 35 columns in our data set.



Property	Value
Rows	1470
Columns	35
Library	PROJECT
Member	ATTRITION
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	1470
Random Seed	12345

6. Final Data Preprocessing:

Firstly, we imported the file. As there were no missing values in our dataset, we didn't perform data impute. Also, there was no need of using the data replacement node as there were no requirement of replacing the data. So we directly connected the File import node to StatExplore node.



Property	Value
ID	EMWS2
Name	project
Status	Open
Notes	
History	
Create Date	11/2/18 5:07 PM
Encoding	wlatin1 Western (Windows)
Data Representation	WINDOWS_64
Native OS	Yes

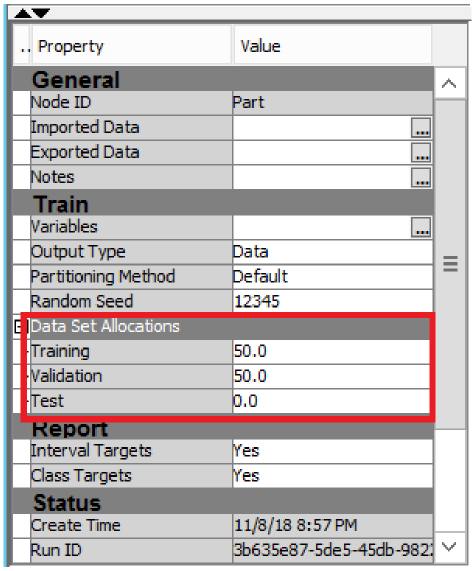
ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Run completed

7. Data partition

In data mining, the quality of Model generalization is assessed by partitioning the data source. A portion of the data, from the project called the *training data set*, is used for initial model fitting. The remaining is reserved for empirical validation of the dataset and is often split into two parts: validation data and test data. The *validation data set* is mainly used to prevent a modeling node from overfitting the training data. The final *test data set* is used for assessment of the model. We partitioned the data as, 50 % for training and 50% for validation.



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
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	11/8/18 8:57 PM
Run ID	3b635e87-5de5-45db-982

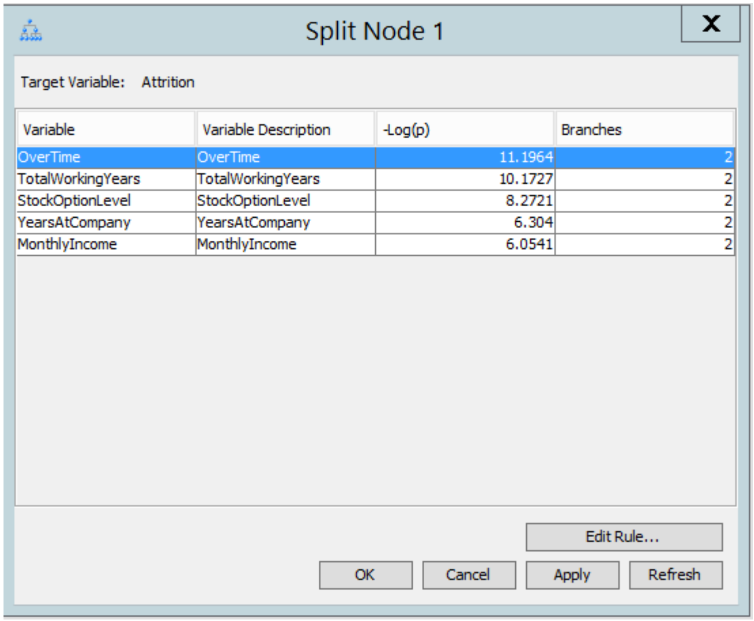
2. Predictive Analysis:

Attrition rate is predicted based on different input variables. Since our target variable – attrition is binary Yes/No, we have used four different models for predictive analysis:

1. **Decision Tree**
2. **Logistic Regression**
3. **Neural Network**
4. **Gradient Boosting**

2.1 Decision Tree:

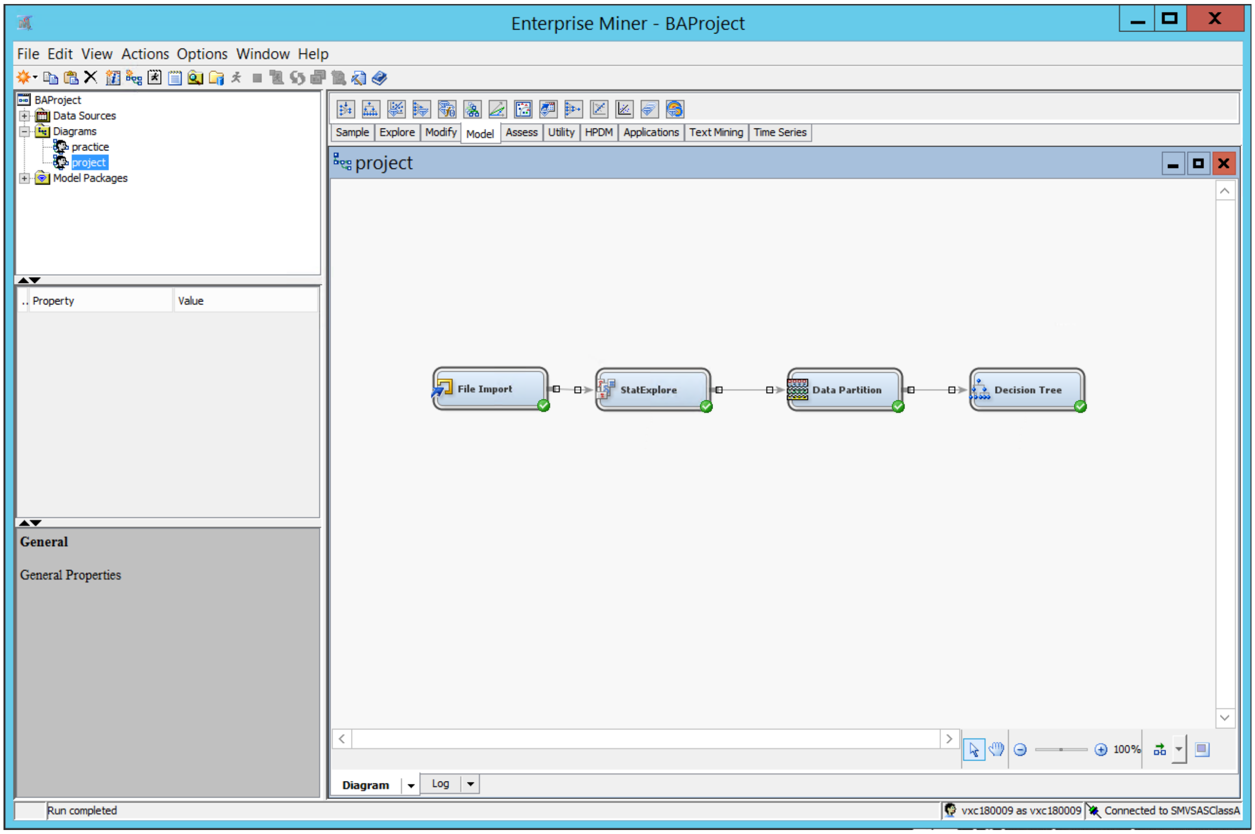
To find the important variables in the Dataset we ran decision tree on processed data. We selected the variables based on the 'Variable Importance' table in the output which are Overetime, Totalworkingyears, stockoptionlevel, Yearsatcompany, MonthlyIncome
the results are as follow:



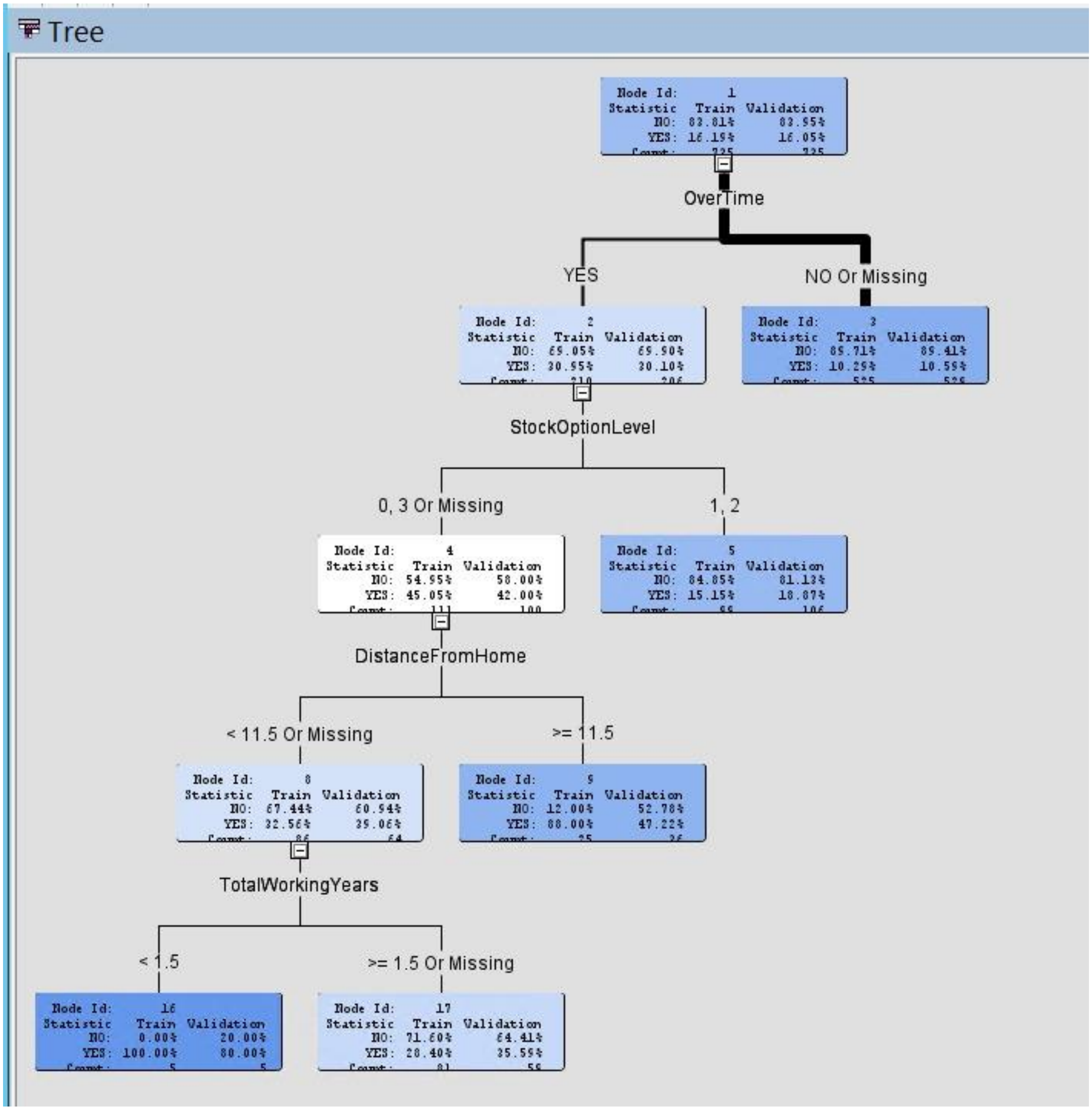
Split Node 1

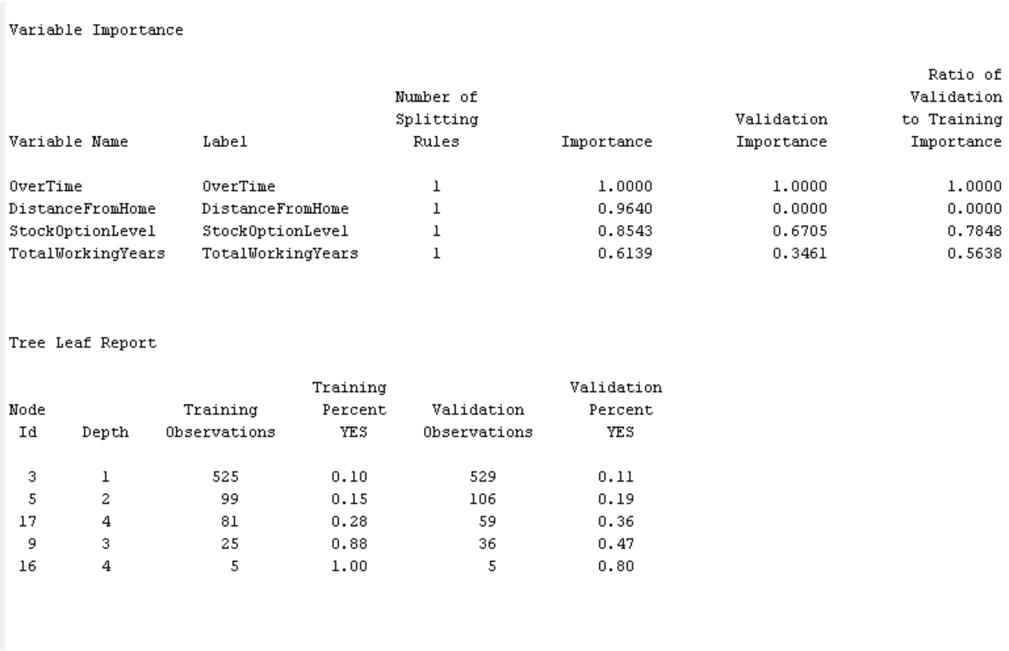
Target Variable: Attrition

Variable	Variable Description	-Log(p)	Branches
OverTime	OverTime	11.1964	2
TotalWorkingYears	TotalWorkingYears	10.1727	2
StockOptionLevel	StockOptionLevel	8.2721	2
YearsAtCompany	YearsAtCompany	6.304	2
MonthlyIncome	MonthlyIncome	6.0541	2



Enterprise Miner - BAProject

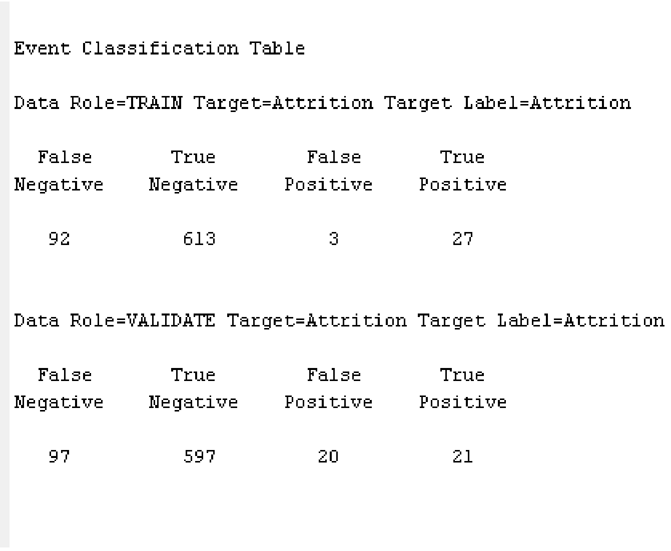

Tree
Node Id: 1
Statistic Train Validation
NO: 83.81% 83.95%
YES: 16.19% 16.05%
Count: 735 735
OverTime
YES
NO Or Missing
Node Id: 2
Statistic Train Validation
NO: 69.05% 69.90%
YES: 30.95% 30.10%
Count: 210 206
Node Id: 3
Statistic Train Validation
NO: 89.71% 89.41%
YES: 10.29% 10.59%
Count: 525 529
StockOptionLevel
0, 3 Or Missing
1, 2
Node Id: 4
Statistic Train Validation
NO: 54.95% 58.00%
YES: 45.05% 42.00%
Count: 111 100
Node Id: 5
Statistic Train Validation
NO: 84.85% 81.13%
YES: 15.15% 18.87%
Count: 99 106
DistanceFromHome
< 11.5 Or Missing
>= 11.5
Node Id: 8
Statistic Train Validation
NO: 67.44% 60.94%
YES: 32.56% 39.06%
Count: 86 64
Node Id: 9
Statistic Train Validation
NO: 12.00% 52.78%
YES: 88.00% 47.22%
Count: 25 36
TotalWorkingYears
< 1.5
>= 1.5 Or Missing
Node Id: 16
Statistic Train Validation
NO: 0.00% 20.00%
YES: 100.00% 80.00%
Count: 5 5
Node Id: 17
Statistic Train Validation
NO: 71.60% 64.41%
YES: 28.40% 35.59%
Count: 81 59



```
Variable Importance

                                                                                      Ratio of
                                           Number of                                 Validation
                                           Splitting                     Validation   to Training
Variable Name       Label                    Rules         Importance    Importance    Importance

OverTime            OverTime                   1              1.0000        1.0000        1.0000
DistanceFromHome    DistanceFromHome           1              0.9640        0.0000        0.0000
StockOptionLevel    StockOptionLevel           1              0.8543        0.6705        0.7848
TotalWorkingYears   TotalWorkingYears          1              0.6139        0.3461        0.5638


Tree Leaf Report

                                Training                  Validation
Node                Training     Percent    Validation     Percent
 Id     Depth     Observations     YES      Observations      YES

  3       1           525         0.10          529           0.11
  5       2            99         0.15          106           0.19
 17       4            81         0.28           59           0.36
  9       3            25         0.88           36           0.47
 16       4             5         1.00            5           0.80
```

Confusion Matrix:

Confusion matrix is calculated from Classification table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False       True       False       True
Negative    Negative    Positive    Positive

   92         613          3          27


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False       True       False       True
Negative    Negative    Positive    Positive

   97         597         20          21
```

From above we can plot confusion matrix as below -

		Predicted	
		Positive	**Negative**
Actual	**Positive**	21	20
	Negative	97	597

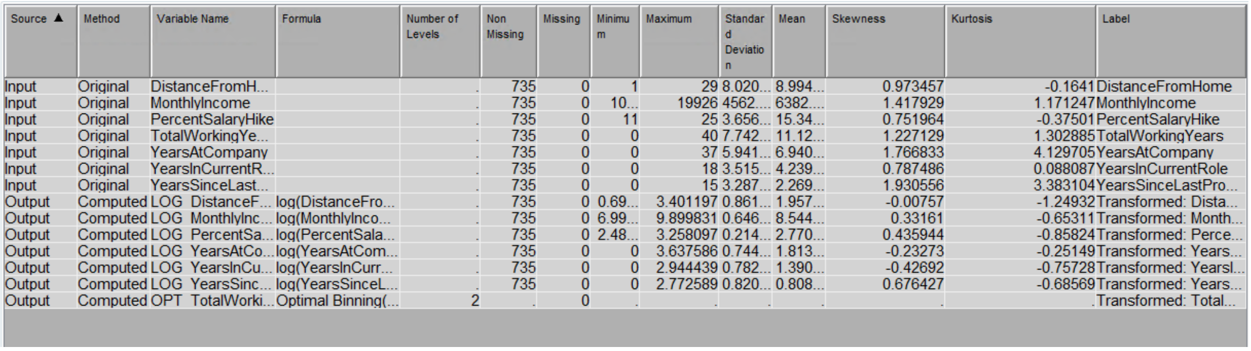
Using Confusion Matrix, the Accuracy of the model is calculated using the following formula -

$$Accuracy = \frac{(597 + 21)}{(21 + 20 + 97 + 597)}$$

Accuracy of the model is 0.8408 i.e. 84.08%

2.2 Logistic Regression:

The skewness in the data set will affect the overall prediction of Attrition rate. To avoid it, we performed some data transformations on the variables which affect the target variable to reduce the skewness of the data. For performing the transformation, we used the available Transformation node in the SAS Enterprise Miner. Below are the results of transformations performed which led us to the best results.

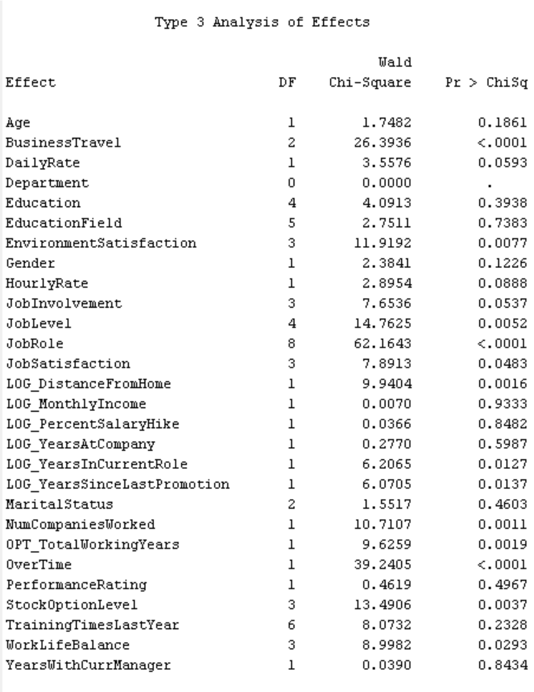


Source ▲	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Standard Deviation	Mean	Skewness	Kurtosis	Label
Input	Original	DistanceFromH...		.	735	0	1	29	8.020...	8.994...	0.973457	-0.1641	DistanceFromHome
Input	Original	MonthlyIncome		.	735	0	10...	19926	4562....	6382....	1.417929	1.171247	MonthlyIncome
Input	Original	PercentSalaryHike		.	735	0	11	25	3.656...	15.34...	0.751964	-0.37501	PercentSalaryHike
Input	Original	TotalWorkingYe...		.	735	0	0	40	7.742...	11.12...	1.227129	1.302885	TotalWorkingYears
Input	Original	YearsAtCompany		.	735	0	0	37	5.941...	6.940...	1.766833	4.129705	YearsAtCompany
Input	Original	YearsInCurrentR...		.	735	0	0	18	3.515...	4.239...	0.787486	0.088087	YearsInCurrentRole
Input	Original	YearsSinceLast...		.	735	0	0	15	3.287...	2.269...	1.930556	3.383104	YearsSinceLastPro...
Output	Computed	LOG_DistanceF...	log(DistanceFro...	.	735	0	0.69...	3.401197	0.861...	1.957...	-0.00757	-1.24932	Transformed: Dista...
Output	Computed	LOG_MonthlyInc...	log(MonthlyInco...	.	735	0	6.99...	9.899831	0.646...	8.544...	0.33161	-0.65311	Transformed: Month...
Output	Computed	LOG_PercentSa...	log(PercentSala...	.	735	0	2.48...	3.258097	0.214...	2.770...	0.435944	-0.85824	Transformed: Perce...
Output	Computed	LOG_YearsAtCo...	log(YearsAtCom...	.	735	0	0	3.637586	0.744...	1.813...	-0.23273	-0.25149	Transformed: Years...
Output	Computed	LOG_YearsInCu...	log(YearsInCurr...	.	735	0	0	2.944439	0.782...	1.390...	-0.42692	-0.75728	Transformed: YearsI...
Output	Computed	LOG_YearsSinc...	log(YearsSinceL...	.	735	0	0	2.772589	0.820...	0.808...	0.676427	-0.68569	Transformed: Years...
Output	Computed	OPT_TotalWorki...	Optimal Binning(...	2	.	0	.	.	.	.	.	.	Transformed: Total...

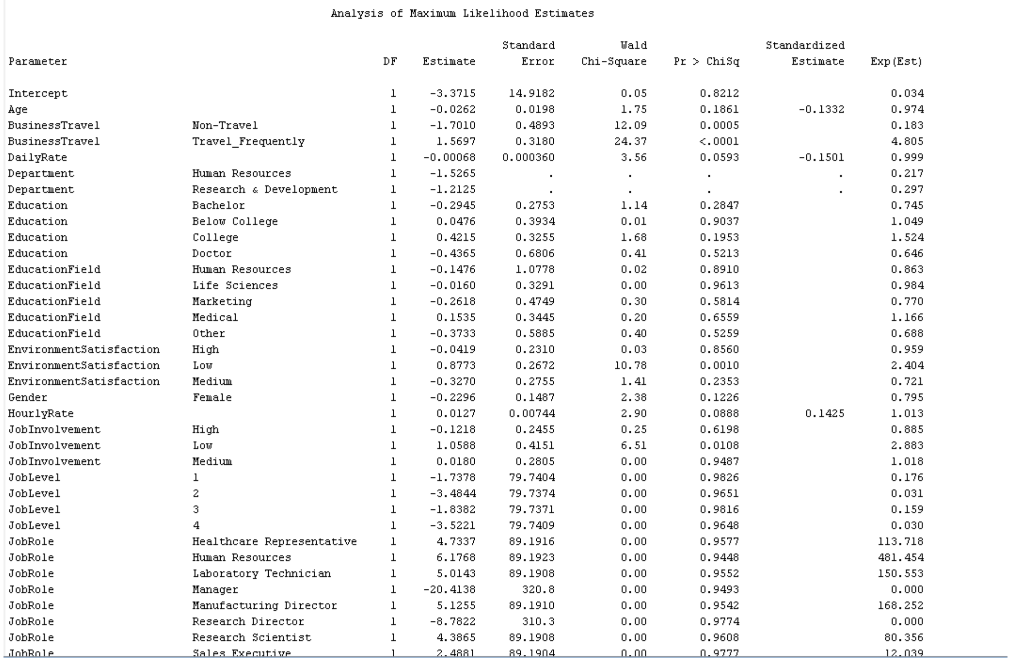
2.2.1 Logistic Regression with Selection Model as None:

To find the important variables we ran Logistic Regression with selected model as 'None' on processed data. We ranked the variables based on the 'Variable Importance' table in the output.

We found the results as follow:



```
                   Type 3 Analysis of Effects

                                         Wald
Effect                         DF    Chi-Square    Pr > ChiSq

Age                             1        1.7482        0.1861
BusinessTravel                  2       26.3936        <.0001
DailyRate                       1        3.5576        0.0593
Department                      0        0.0000         .
Education                       4        4.0913        0.3938
EducationField                  5        2.7511        0.7383
EnvironmentSatisfaction         3       11.9192        0.0077
Gender                          1        2.3841        0.1226
HourlyRate                      1        2.8954        0.0888
JobInvolvement                  3        7.6536        0.0537
JobLevel                        4       14.7625        0.0052
JobRole                         8       62.1643        <.0001
JobSatisfaction                 3        7.8913        0.0483
LOG_DistanceFromHome            1        9.9404        0.0016
LOG_MonthlyIncome               1        0.0070        0.9333
LOG_PercentSalaryHike           1        0.0366        0.8482
LOG_YearsAtCompany              1        0.2770        0.5987
LOG_YearsInCurrentRole          1        6.2065        0.0127
LOG_YearsSinceLastPromotion     1        6.0705        0.0137
MaritalStatus                   2        1.5517        0.4603
NumCompaniesWorked              1       10.7107        0.0011
OPT_TotalWorkingYears           1        9.6259        0.0019
OverTime                        1       39.2405        <.0001
PerformanceRating               1        0.4619        0.4967
StockOptionLevel                3       13.4906        0.0037
TrainingTimesLastYear           6        8.0732        0.2328
WorkLifeBalance                 3        8.9982        0.0293
YearsWithCurrManager            1        0.0390        0.8434
```

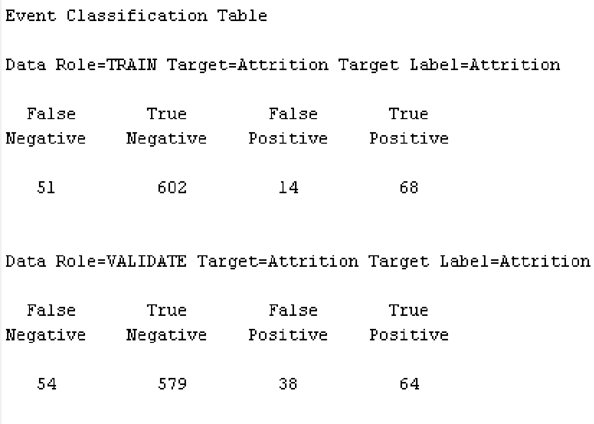


```
                                        Analysis of Maximum Likelihood Estimates

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

Intercept                                       1     -3.3715    14.9182          0.05        0.8212                       0.034
Age                                             1     -0.0262     0.0198          1.75        0.1861         -0.1332       0.974
BusinessTravel            Non-Travel            1     -1.7010     0.4893         12.09        0.0005                       0.183
BusinessTravel            Travel_Frequently     1      1.5697     0.3180         24.37        <.0001                       4.805
DailyRate                                       1    -0.00068   0.000360          3.56        0.0593         -0.1501       0.999
Department                Human Resources       1     -1.5265          .             .             .               .       0.217
Department                Research & Development 1    -1.2125          .             .             .               .       0.297
Education                 Bachelor              1     -0.2945     0.2753          1.14        0.2847                       0.745
Education                 Below College         1      0.0476     0.3934          0.01        0.9037                       1.049
Education                 College               1      0.4215     0.3255          1.68        0.1953                       1.524
Education                 Doctor                1     -0.4365     0.6806          0.41        0.5213                       0.646
EducationField            Human Resources       1     -0.1476     1.0778          0.02        0.8910                       0.863
EducationField            Life Sciences         1     -0.0160     0.3291          0.00        0.9613                       0.984
EducationField            Marketing             1     -0.2618     0.4749          0.30        0.5814                       0.770
EducationField            Medical               1      0.1535     0.3445          0.20        0.6559                       1.166
EducationField            Other                 1     -0.3733     0.5885          0.40        0.5259                       0.688
EnvironmentSatisfaction   High                  1     -0.0419     0.2310          0.03        0.8560                       0.959
EnvironmentSatisfaction   Low                   1      0.8773     0.2672         10.78        0.0010                       2.404
EnvironmentSatisfaction   Medium                1     -0.3270     0.2755          1.41        0.2353                       0.721
Gender                    Female                1     -0.2296     0.1487          2.38        0.1226                       0.795
HourlyRate                                      1      0.0127    0.00744          2.90        0.0888          0.1425       1.013
JobInvolvement            High                  1     -0.1218     0.2455          0.25        0.6198                       0.885
JobInvolvement            Low                   1      1.0588     0.4151          6.51        0.0108                       2.883
JobInvolvement            Medium                1      0.0180     0.2805          0.00        0.9487                       1.018
JobLevel                  1                     1     -1.7378    79.7404          0.00        0.9826                       0.176
JobLevel                  2                     1     -3.4844    79.7374          0.00        0.9651                       0.031
JobLevel                  3                     1     -1.8382    79.7371          0.00        0.9816                       0.159
JobLevel                  4                     1     -3.5221    79.7409          0.00        0.9648                       0.030
JobRole                   Healthcare Representative 1  4.7337    89.1916          0.00        0.9577                     113.718
JobRole                   Human Resources       1      6.1768    89.1923          0.00        0.9448                     481.454
JobRole                   Laboratory Technician 1      5.0143    89.1908          0.00        0.9552                     150.553
JobRole                   Manager               1    -20.4138      320.8          0.00        0.9493                       0.000
JobRole                   Manufacturing Director 1     5.1255    89.1910          0.00        0.9542                     168.252
JobRole                   Research Director     1     -8.7822      310.3          0.00        0.9774                       0.000
JobRole                   Research Scientist    1      4.3865    89.1908          0.00        0.9608                      80.356
JobRole                   Sales Executive       1      2.4881    89.1904          0.00        0.9777                      12.039
```

By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, JobRole, OverTime.

Confusion Matrix:

We can calculate the Confusion matrix from below table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   51         602         14          68


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   54         579         38          64
```

From above we can plot confusion matrix as below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	64	38
	Negative	54	579

We can find the Accuracy of the model from Confusion matrix using following formula.

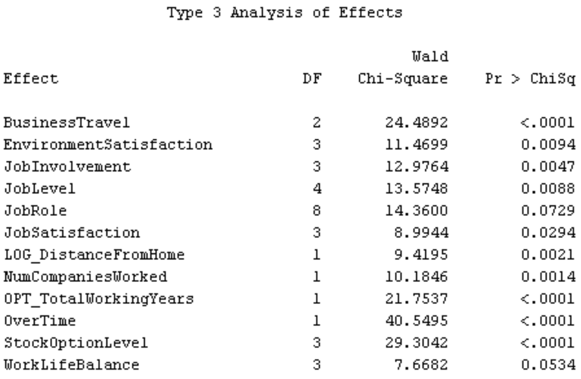
$$Accuracy = \frac{(64 + 579)}{(64 + 38 + 579 + 54)}$$

Accuracy of the model is 0.8748 i.e. 87.48%

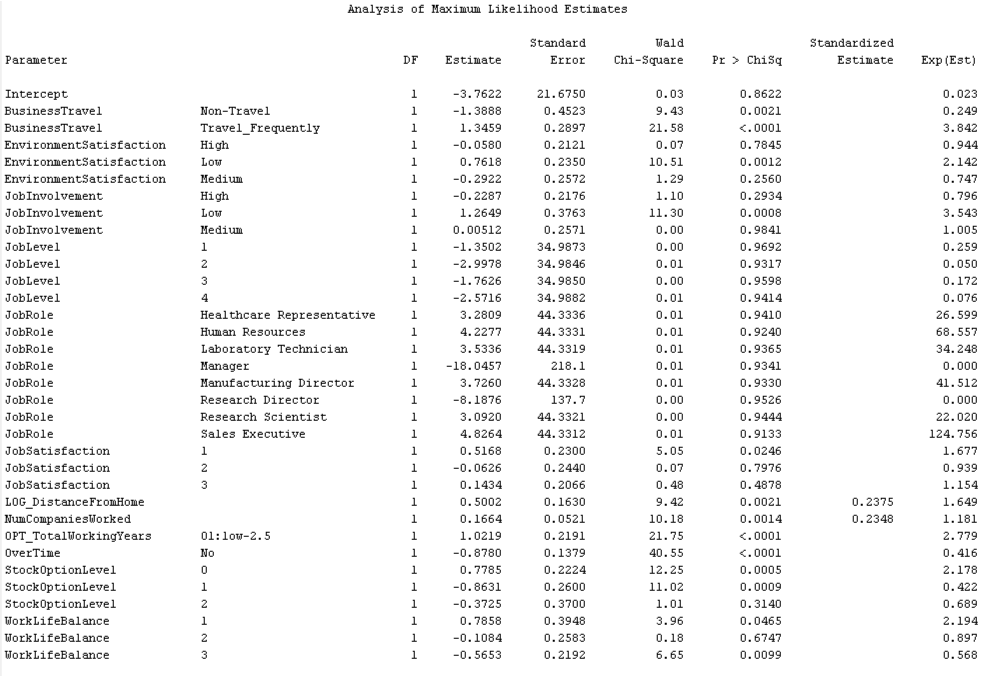
2.2.2 Logistic Regression- Forward Regression:

To find the important variables we ran Logistic Regression with selected model as ‘Forward’ on processed data. We selected the variables based on the ‘Variable Importance’ table in the output.

We found the results as follow:



```
                 Type 3 Analysis of Effects

                                          Wald
Effect                        DF    Chi-Square    Pr > ChiSq

BusinessTravel                 2       24.4892        <.0001
EnvironmentSatisfaction        3       11.4699        0.0094
JobInvolvement                 3       12.9764        0.0047
JobLevel                       4       13.5748        0.0088
JobRole                        8       14.3600        0.0729
JobSatisfaction                3        8.9944        0.0294
LOG_DistanceFromHome           1        9.4195        0.0021
NumCompaniesWorked             1       10.1846        0.0014
OPT_TotalWorkingYears          1       21.7537        <.0001
OverTime                       1       40.5495        <.0001
StockOptionLevel               3       29.3042        <.0001
WorkLifeBalance                3        7.6682        0.0534
```

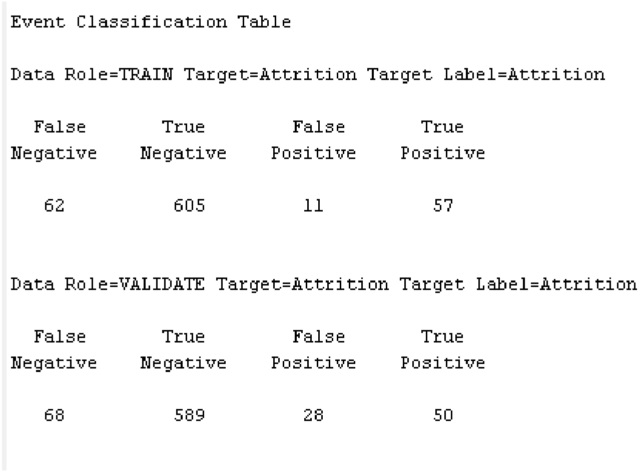


```
                                            Analysis of Maximum Likelihood Estimates

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

Intercept                                               1     -3.7622   21.6750          0.03        0.8622                       0.023
BusinessTravel           Non-Travel                     1     -1.3888    0.4523          9.43        0.0021                       0.249
BusinessTravel           Travel_Frequently              1      1.3459    0.2897         21.58        <.0001                       3.842
EnvironmentSatisfaction  High                           1     -0.0580    0.2121          0.07        0.7845                       0.944
EnvironmentSatisfaction  Low                            1      0.7618    0.2350         10.51        0.0012                       2.142
EnvironmentSatisfaction  Medium                         1     -0.2922    0.2572          1.29        0.2560                       0.747
JobInvolvement           High                           1     -0.2287    0.2176          1.10        0.2934                       0.796
JobInvolvement           Low                            1      1.2649    0.3763         11.30        0.0008                       3.543
JobInvolvement           Medium                         1     0.00512    0.2571          0.00        0.9841                       1.005
JobLevel                 1                              1     -1.3502   34.9873          0.00        0.9692                       0.259
JobLevel                 2                              1     -2.9978   34.9846          0.01        0.9317                       0.050
JobLevel                 3                              1     -1.7626   34.9850          0.00        0.9598                       0.172
JobLevel                 4                              1     -2.5716   34.9882          0.01        0.9414                       0.076
JobRole                  Healthcare Representative      1      3.2809   44.3336          0.01        0.9410                      26.599
JobRole                  Human Resources                1      4.2277   44.3331          0.01        0.9240                      68.557
JobRole                  Laboratory Technician          1      3.5336   44.3319          0.01        0.9365                      34.248
JobRole                  Manager                        1    -18.0457     218.1          0.01        0.9341                       0.000
JobRole                  Manufacturing Director         1      3.7260   44.3328          0.01        0.9330                      41.512
JobRole                  Research Director              1     -8.1876     137.7          0.00        0.9526                       0.000
JobRole                  Research Scientist             1      3.0920   44.3321          0.00        0.9444                      22.020
JobRole                  Sales Executive                1      4.8264   44.3312          0.01        0.9133                     124.756
JobSatisfaction          1                              1      0.5168    0.2300          5.05        0.0246                       1.677
JobSatisfaction          2                              1     -0.0626    0.2440          0.07        0.7976                       0.939
JobSatisfaction          3                              1      0.1434    0.2066          0.48        0.4878                       1.154
LOG_DistanceFromHome                                    1      0.5002    0.1630          9.42        0.0021          0.2375       1.649
NumCompaniesWorked                                      1      0.1664    0.0521         10.18        0.0014          0.2348       1.181
OPT_TotalWorkingYears    01:low-2.5                     1      1.0219    0.2191         21.75        <.0001                       2.779
OverTime                 No                             1     -0.8780    0.1379         40.55        <.0001                       0.416
StockOptionLevel         0                              1      0.7785    0.2224         12.25        0.0005                       2.178
StockOptionLevel         1                              1     -0.8631    0.2600         11.02        0.0009                       0.422
StockOptionLevel         2                              1     -0.3725    0.3700          1.01        0.3140                       0.689
WorkLifeBalance          1                              1      0.7858    0.3948          3.96        0.0465                       2.194
WorkLifeBalance          2                              1     -0.1084    0.2583          0.18        0.6747                       0.897
WorkLifeBalance          3                              1     -0.5653    0.2192          6.65        0.0099                       0.568
```

By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, OPT_TotalWorkingYears, OverTime, StockOptionLevel.

Confusion Matrix:

We can calculate the Confusion matrix from below table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False      True      False     True
Negative   Negative   Positive  Positive

   62        605        11        57


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False      True      False     True
Negative   Negative   Positive  Positive

   68        589        28        50
```

From above we can plot confusion matrix as below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	50	28
	Negative	68	589

We can find the Accuracy of the model from Confusion matrix using following formula.

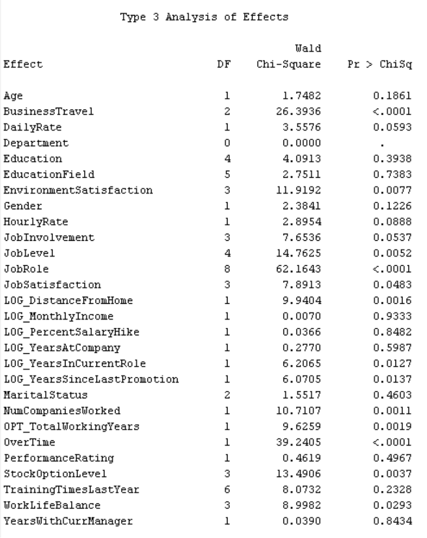
$$Accuracy = \frac{(50 + 589)}{(50 + 589 + 68 + 28)}$$

Accuracy of the model is 0.8694 i.e. 86.94%

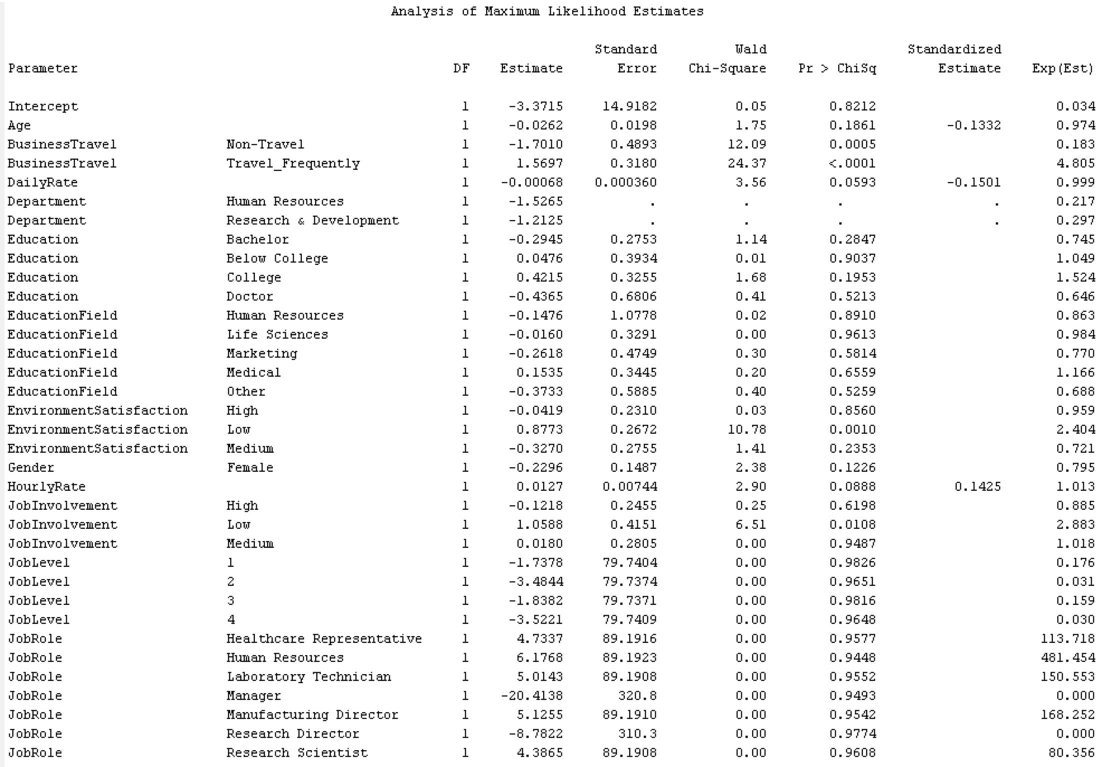
2.2.3 Logistic Regression – Backward Regression:

To find the important variables we ran Logistic Regression with Selected model as ‘Backward’ on processed data. We selected the variables based on the ‘Variable Importance’ table in the output.

We found the results as follow:



```
                 Type 3 Analysis of Effects

                                             Wald
Effect                          DF     Chi-Square     Pr > ChiSq

Age                              1         1.7482         0.1861
BusinessTravel                   2        26.3936         <.0001
DailyRate                        1         3.5576         0.0593
Department                       0         0.0000          .
Education                        4         4.0913         0.3938
EducationField                   5         2.7511         0.7383
EnvironmentSatisfaction          3        11.9192         0.0077
Gender                           1         2.3841         0.1226
HourlyRate                       1         2.8954         0.0888
JobInvolvement                   3         7.6536         0.0537
JobLevel                         4        14.7625         0.0052
JobRole                          8        62.1643         <.0001
JobSatisfaction                  3         7.8913         0.0483
LOG_DistanceFromHome             1         9.9404         0.0016
LOG_MonthlyIncome                1         0.0070         0.9333
LOG_PercentSalaryHike            1         0.0366         0.8482
LOG_YearsAtCompany               1         0.2770         0.5987
LOG_YearsInCurrentRole           1         6.2065         0.0127
LOG_YearsSinceLastPromotion      1         6.0705         0.0137
MaritalStatus                    2         1.5517         0.4603
NumCompaniesWorked               1        10.7107         0.0011
OPT_TotalWorkingYears            1         9.6259         0.0019
OverTime                         1        39.2405         <.0001
PerformanceRating                1         0.4619         0.4967
StockOptionLevel                 3        13.4906         0.0037
TrainingTimesLastYear            6         8.0732         0.2328
WorkLifeBalance                  3         8.9982         0.0293
YearsWithCurrManager             1         0.0390         0.8434
```

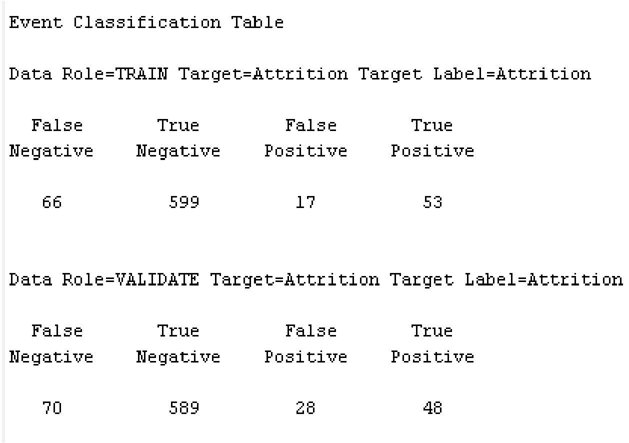


```
                                           Analysis of Maximum Likelihood Estimates

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

Intercept                                                1     -3.3715     14.9182          0.05        0.8212                       0.034
Age                                                      1     -0.0262      0.0198          1.75        0.1861         -0.1332       0.974
BusinessTravel             Non-Travel                    1     -1.7010      0.4893         12.09        0.0005                       0.183
BusinessTravel             Travel_Frequently             1      1.5697      0.3180         24.37        <.0001                       4.805
DailyRate                                                1    -0.00068    0.000360          3.56        0.0593         -0.1501       0.999
Department                 Human Resources               1     -1.5265           .             .             .               .       0.217
Department                 Research & Development        1     -1.2125           .             .             .               .       0.297
Education                  Bachelor                      1     -0.2945      0.2753          1.14        0.2847                       0.745
Education                  Below College                 1      0.0476      0.3934          0.01        0.9037                       1.049
Education                  College                       1      0.4215      0.3255          1.68        0.1953                       1.524
Education                  Doctor                        1     -0.4365      0.6806          0.41        0.5213                       0.646
EducationField             Human Resources               1     -0.1476      1.0778          0.02        0.8910                       0.863
EducationField             Life Sciences                 1     -0.0160      0.3291          0.00        0.9613                       0.984
EducationField             Marketing                     1     -0.2618      0.4749          0.30        0.5814                       0.770
EducationField             Medical                       1      0.1535      0.3445          0.20        0.6559                       1.166
EducationField             Other                         1     -0.3733      0.5885          0.40        0.5259                       0.688
EnvironmentSatisfaction    High                          1     -0.0419      0.2310          0.03        0.8560                       0.959
EnvironmentSatisfaction    Low                           1      0.8773      0.2672         10.78        0.0010                       2.404
EnvironmentSatisfaction    Medium                        1     -0.3270      0.2755          1.41        0.2353                       0.721
Gender                     Female                        1     -0.2296      0.1487          2.38        0.1226                       0.795
HourlyRate                                               1      0.0127     0.00744          2.90        0.0888          0.1425       1.013
JobInvolvement             High                          1     -0.1218      0.2455          0.25        0.6198                       0.885
JobInvolvement             Low                           1      1.0588      0.4151          6.51        0.0108                       2.883
JobInvolvement             Medium                        1      0.0180      0.2805          0.00        0.9487                       1.018
JobLevel                   1                             1     -1.7378     79.7404          0.00        0.9826                       0.176
JobLevel                   2                             1     -3.4844     79.7374          0.00        0.9651                       0.031
JobLevel                   3                             1     -1.8382     79.7371          0.00        0.9816                       0.159
JobLevel                   4                             1     -3.5221     79.7409          0.00        0.9648                       0.030
JobRole                    Healthcare Representative     1      4.7337     89.1916          0.00        0.9577                     113.718
JobRole                    Human Resources               1      6.1768     89.1923          0.00        0.9448                     481.454
JobRole                    Laboratory Technician         1      5.0143     89.1908          0.00        0.9552                     150.553
JobRole                    Manager                       1    -20.4138       320.8          0.00        0.9493                       0.000
JobRole                    Manufacturing Director        1      5.1255     89.1910          0.00        0.9542                     168.252
JobRole                    Research Director             1     -8.7822       310.3          0.00        0.9774                       0.000
JobRole                    Research Scientist            1      4.3865     89.1908          0.00        0.9608                      80.356
```

By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, JobRole, OverTime.

Confusion Matrix:

We can calculate the Confusion matrix from below table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   66          599         17          53


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   70          589         28          48
```

From above we can plot confusion matrix as below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	48	28
	Negative	70	589

We can find the Accuracy of the model from Confusion matrix using following formula.

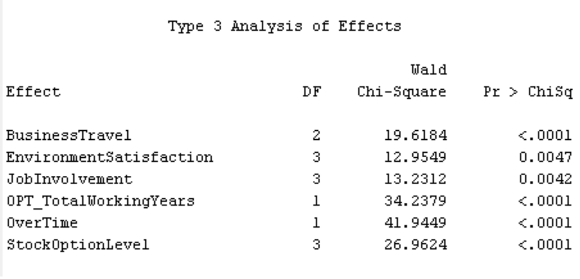
$$Accuracy = \frac{(48 + 589)}{(48 + 589 + 70 + 28)}$$

Accuracy of the model is 0.8667 i.e. 86.67%

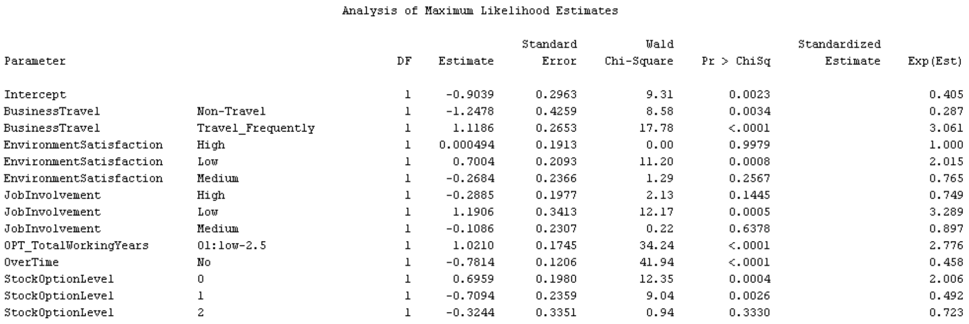
2.2.4 Logistic Regression – Step-wise Regression:

To find the important variables we ran Logistic Regression with Selected model as ‘Step-Wise’ on processed data. We selected the variables based on the ‘Variable Importance’ table in the output.

We found the results as follow:



```
                 Type 3 Analysis of Effects

                                           Wald
Effect                        DF    Chi-Square    Pr > ChiSq

BusinessTravel                 2       19.6184        <.0001
EnvironmentSatisfaction        3       12.9549        0.0047
JobInvolvement                 3       13.2312        0.0042
OPT_TotalWorkingYears          1       34.2379        <.0001
OverTime                       1       41.9449        <.0001
StockOptionLevel               3       26.9624        <.0001
```

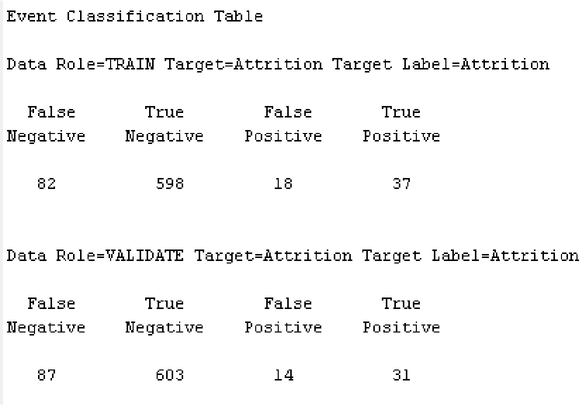


```
                                           Analysis of Maximum Likelihood Estimates

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

Intercept                                           1     -0.9039      0.2963          9.31        0.0023                      0.405
BusinessTravel              Non-Travel              1     -1.2478      0.4259          8.58        0.0034                      0.287
BusinessTravel              Travel_Frequently       1      1.1186      0.2653         17.78        <.0001                      3.061
EnvironmentSatisfaction     High                    1    0.000494      0.1913          0.00        0.9979                      1.000
EnvironmentSatisfaction     Low                     1      0.7004      0.2093         11.20        0.0008                      2.015
EnvironmentSatisfaction     Medium                  1     -0.2684      0.2366          1.29        0.2567                      0.765
JobInvolvement              High                    1     -0.2885      0.1977          2.13        0.1445                      0.749
JobInvolvement              Low                     1      1.1906      0.3413         12.17        0.0005                      3.289
JobInvolvement              Medium                  1     -0.1086      0.2307          0.22        0.6378                      0.897
OPT_TotalWorkingYears       01:low-2.5              1      1.0210      0.1745         34.24        <.0001                      2.776
OverTime                    No                      1     -0.7814      0.1206         41.94        <.0001                      0.458
StockOptionLevel            0                       1      0.6959      0.1980         12.35        0.0004                      2.006
StockOptionLevel            1                       1     -0.7094      0.2359          9.04        0.0026                      0.492
StockOptionLevel            2                       1     -0.3244      0.3351          0.94        0.3330                      0.723
```

By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, OPT_TotalWorkingYears, OverTime, StockOptionLevel.

Confusion Matrix:

We can calculate the Confusion matrix from below table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False       True       False       True
Negative    Negative    Positive    Positive

   82         598          18          37


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False       True       False       True
Negative    Negative    Positive    Positive

   87         603          14          31
```

From above we can plot confusion matrix as below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	31	14
	Negative	87	603

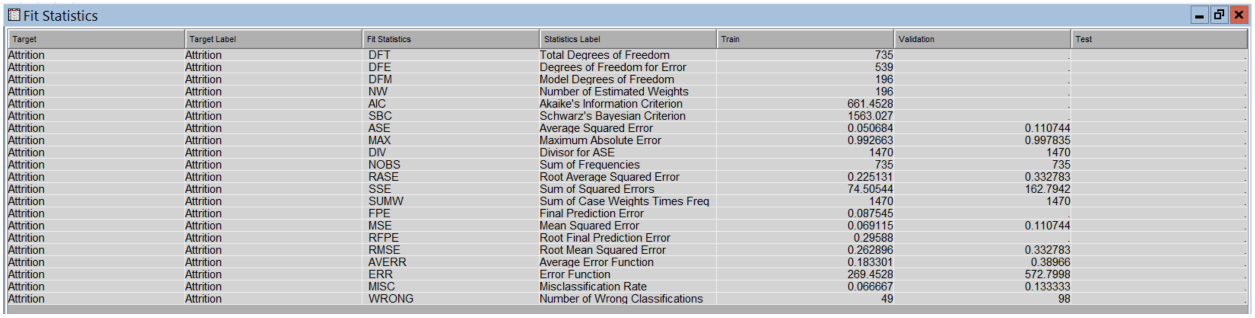
We can find the Accuracy of the model from Confusion matrix using following formula.

$$Accuracy = \frac{(31 + 603)}{(31 + 603 + 14 + 87)}$$

Accuracy of the model is 0.8626 i.e. 86.26%

2.3 Neural Network:

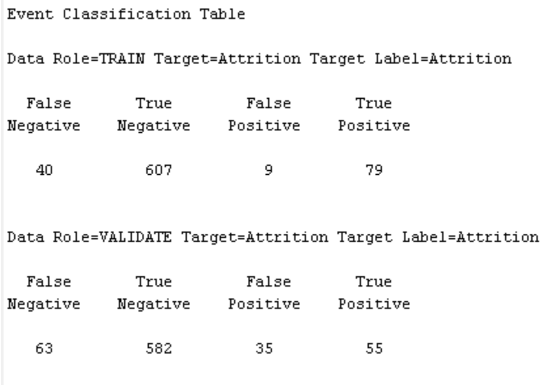
We ran the Neural Network model and found the misclassification rate as follows –



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition	Attrition	DFT	Total Degrees of Freedom	735	.	.
Attrition	Attrition	DFE	Degrees of Freedom for Error	539	.	.
Attrition	Attrition	DFM	Model Degrees of Freedom	196	.	.
Attrition	Attrition	NW	Number of Estimated Weights	196	.	.
Attrition	Attrition	AIC	Akaike's Information Criterion	661.4528	.	.
Attrition	Attrition	SBC	Schwarz's Bayesian Criterion	1563.027	.	.
Attrition	Attrition	ASE	Average Squared Error	0.050684	0.110744	.
Attrition	Attrition	MAX	Maximum Absolute Error	0.992663	0.997835	.
Attrition	Attrition	DIV	Divisor for ASE	1470	1470	.
Attrition	Attrition	NOBS	Sum of Frequencies	735	735	.
Attrition	Attrition	RASE	Root Average Squared Error	0.225131	0.332783	.
Attrition	Attrition	SSE	Sum of Squared Errors	74.50544	162.7942	.
Attrition	Attrition	SUMW	Sum of Case Weights Times Freq	1470	1470	.
Attrition	Attrition	FPE	Final Prediction Error	0.087545	.	.
Attrition	Attrition	MSE	Mean Squared Error	0.069115	0.110744	.
Attrition	Attrition	RFPE	Root Final Prediction Error	0.29588	.	.
Attrition	Attrition	RMSE	Root Mean Squared Error	0.262896	0.332783	.
Attrition	Attrition	AVERR	Average Error Function	0.183301	0.38966	.
Attrition	Attrition	ERR	Error Function	269.4528	572.7998	.
Attrition	Attrition	MISC	Misclassification Rate	0.066667	0.133333	.
Attrition	Attrition	WRONG	Number of Wrong Classifications	49	98	.

Confusion Matrix: We can calculate the Confusion matrix from below table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False      True      False      True
Negative   Negative   Positive   Positive

   40        607         9         79


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False      True      False      True
Negative   Negative   Positive   Positive

   63        582        35         55
```

From above we can plot confusion matrix as below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	55	35
	Negative	63	582

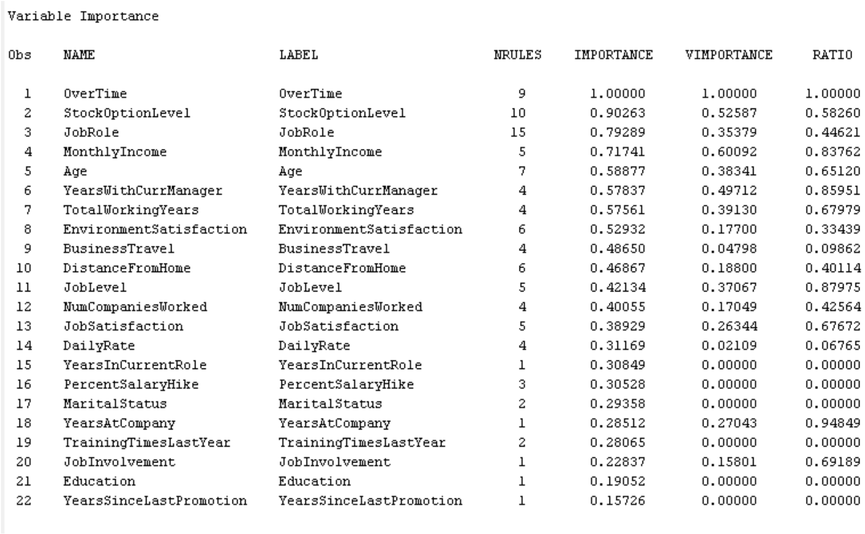
We can find the Accuracy of the model from Confusion matrix using following formula.

$$Accuracy = \frac{(55 + 582)}{(55 + 582 + 63 + 35)}$$

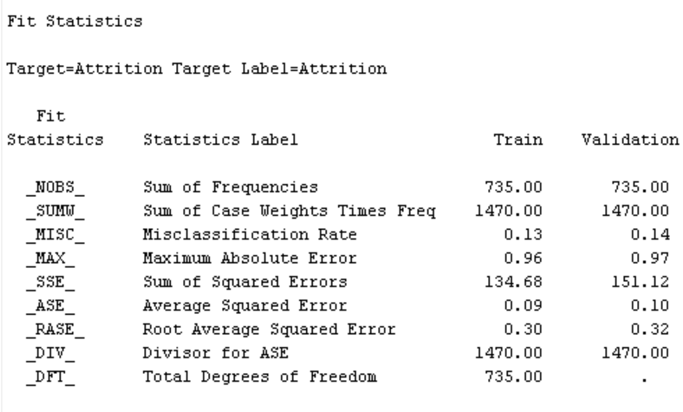
Accuracy of the model come to 0.8667 i.e. 86.67%

2.4 Gradient Boosting:

To find the important variables we ran Gradient Boosting on processed data. We selected the variables based on the 'Variable Importance' table in the output which are OverTime, StockOptionLevel, JobRole, MonthlyIncome, Age. We found the results as follow:



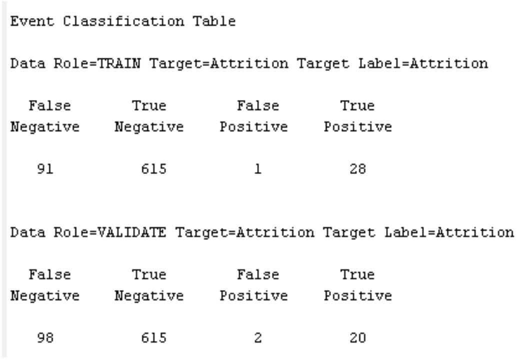
```
Variable Importance

Obs    NAME                         LABEL                        NRULES    IMPORTANCE    VIMPORTANCE     RATIO

  1    OverTime                     OverTime                        9        1.00000       1.00000      1.00000
  2    StockOptionLevel             StockOptionLevel               10        0.90263       0.52587      0.58260
  3    JobRole                      JobRole                        15        0.79289       0.35379      0.44621
  4    MonthlyIncome                MonthlyIncome                   5        0.71741       0.60092      0.83762
  5    Age                          Age                             7        0.58877       0.38341      0.65120
  6    YearsWithCurrManager         YearsWithCurrManager            4        0.57837       0.49712      0.85951
  7    TotalWorkingYears            TotalWorkingYears               4        0.57561       0.39130      0.67979
  8    EnvironmentSatisfaction      EnvironmentSatisfaction         6        0.52932       0.17700      0.33439
  9    BusinessTravel               BusinessTravel                  4        0.48650       0.04798      0.09862
 10    DistanceFromHome             DistanceFromHome                6        0.46867       0.18800      0.40114
 11    JobLevel                     JobLevel                        5        0.42134       0.37067      0.87975
 12    NumCompaniesWorked           NumCompaniesWorked              4        0.40055       0.17049      0.42564
 13    JobSatisfaction              JobSatisfaction                 5        0.38929       0.26344      0.67672
 14    DailyRate                    DailyRate                       4        0.31169       0.02109      0.06765
 15    YearsInCurrentRole           YearsInCurrentRole              1        0.30849       0.00000      0.00000
 16    PercentSalaryHike            PercentSalaryHike               3        0.30528       0.00000      0.00000
 17    MaritalStatus                MaritalStatus                   2        0.29358       0.00000      0.00000
 18    YearsAtCompany               YearsAtCompany                  1        0.28512       0.27043      0.94849
 19    TrainingTimesLastYear        TrainingTimesLastYear           2        0.28065       0.00000      0.00000
 20    JobInvolvement               JobInvolvement                  1        0.22837       0.15801      0.69189
 21    Education                    Education                       1        0.19052       0.00000      0.00000
 22    YearsSinceLastPromotion      YearsSinceLastPromotion         1        0.15726       0.00000      0.00000
```



```
Fit Statistics

Target=Attrition Target Label=Attrition

   Fit
Statistics    Statistics Label                  Train    Validation

  _NOBS_      Sum of Frequencies               735.00       735.00
  _SUMW_      Sum of Case Weights Times Freq  1470.00      1470.00
  _MISC_      Misclassification Rate             0.13         0.14
  _MAX_       Maximum Absolute Error             0.96         0.97
  _SSE_       Sum of Squared Errors            134.68       151.12
  _ASE_       Average Squared Error              0.09         0.10
  _RASE_      Root Average Squared Error         0.30         0.32
  _DIV_       Divisor for ASE                 1470.00      1470.00
  _DFT_       Total Degrees of Freedom         735.00          .
```

Confusion Matrix:

We can calculate the Confusion matrix from Classification table present in the output.



```
Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   91          615          1           28


Data Role=VALIDATE Target=Attrition Target Label=Attrition

  False       True        False       True
Negative    Negative    Positive    Positive

   98          615          2           20
```

From above we can plot confusion matrix as below -

		Predicted	
		Positive	**Negative**
Actual	**Positive**	20	2
	Negative	98	615

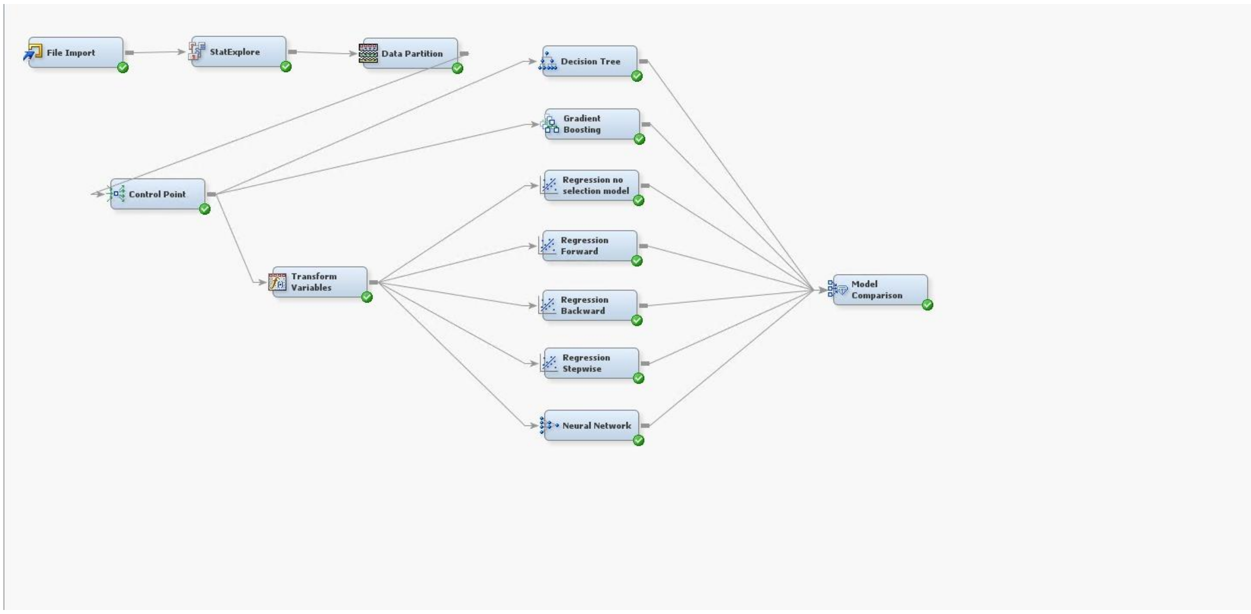
We can find the Accuracy of the model from Confusion matrix using following formula -

$$Accuracy = \frac{(20 + 615)}{(20 + 615 + 98 + 2)}$$

Accuracy of the model is 0.8639 i.e. 86.39%

2.5 Final Model:

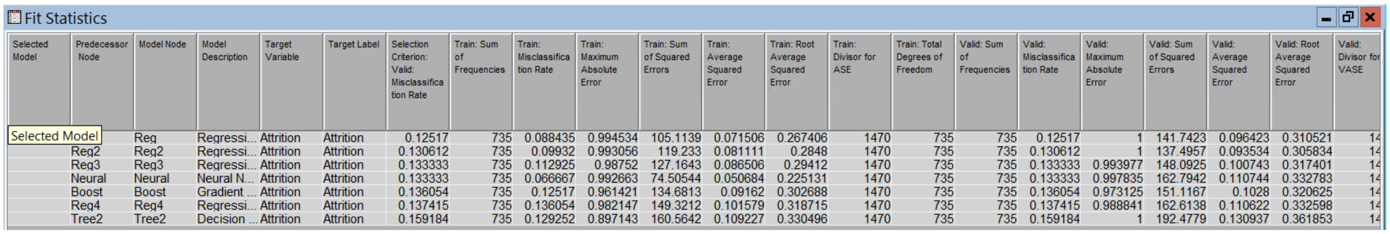
Following snapshot shows the final model. It involves all the nodes used for Data preprocessing, Descriptive Analysis and Predictive Modeling.



2.5.1 Model Comparison:

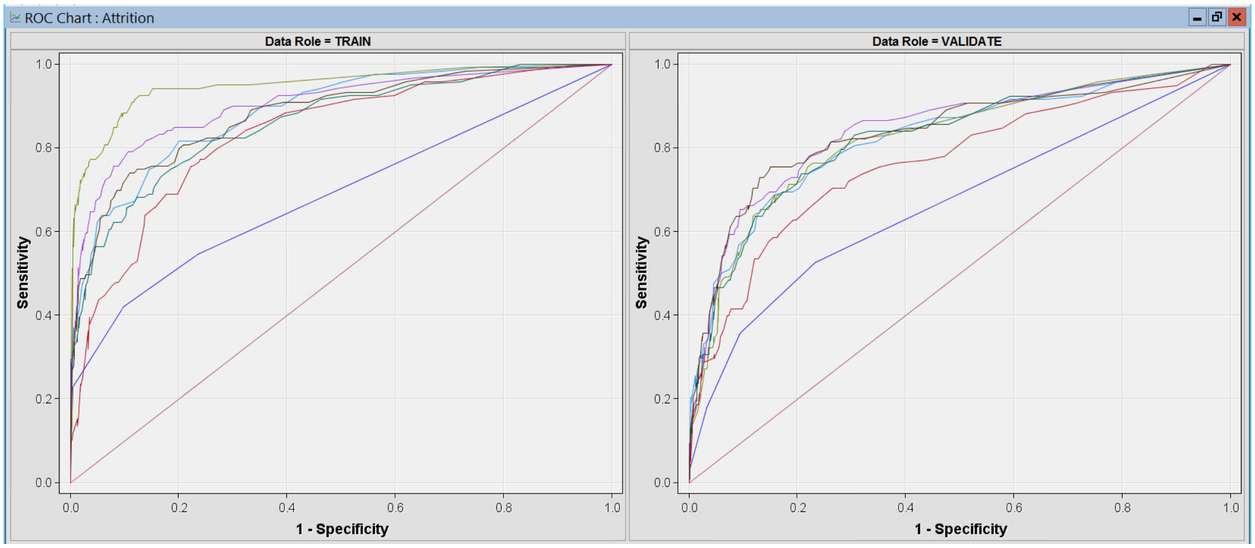
We ran 7 models/Classifiers on the data set. We connected the Classifiers to Model Comparison node to find the best suitable model to find the attrition rate. Considering the misclassification rates for Train and Validation data sets and accuracy calculated using Confusion Matrix as well as misclassification rate, we can conclude that 'Regression with all inputs' is the best model for predicting attrition.

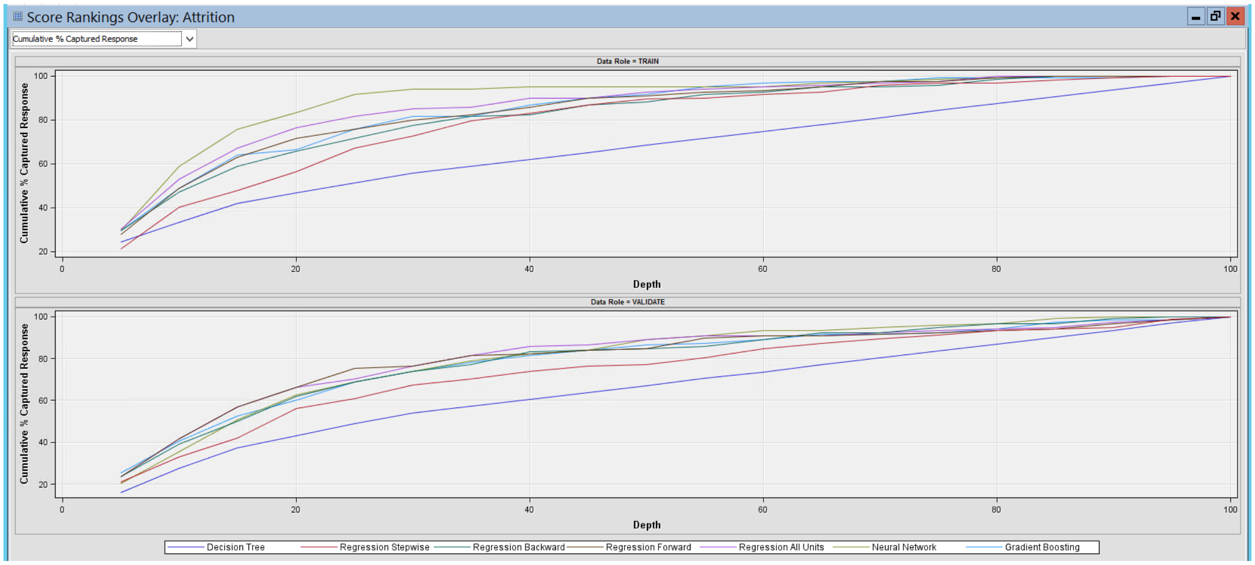
Below is the snapshot of Model Comparison Output –



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	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: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Selected Model		Reg	Regressi...	Attrition	Attrition	0.12517	735	0.088435	0.994534	105.1139	0.071506	0.267406	1470	735	735	0.12517	1	141.7423	0.096423	0.310521	14
	Reg2	Reg2	Regressi...	Attrition	Attrition	0.130612	735	0.09932	0.993056	119.233	0.081111	0.2848	1470	735	735	0.130612	1	137.4957	0.093534	0.305834	14
	Reg3	Reg3	Regressi...	Attrition	Attrition	0.133333	735	0.112925	0.98752	127.1643	0.086506	0.29412	1470	735	735	0.133333	0.993977	148.0925	0.100743	0.317401	14
	Neural	Neural	Neural N...	Attrition	Attrition	0.133333	735	0.066667	0.992663	74.50544	0.050684	0.225131	1470	735	735	0.133333	0.997835	162.7942	0.110744	0.332783	14
	Boost	Boost	Gradient ...	Attrition	Attrition	0.136054	735	0.12517	0.961421	134.6813	0.09162	0.302688	1470	735	735	0.136054	0.973125	151.1167	0.1028	0.320625	14
	Reg4	Reg4	Regressi...	Attrition	Attrition	0.137415	735	0.136054	0.982147	149.3212	0.101579	0.318715	1470	735	735	0.137415	0.988841	162.6138	0.110622	0.332598	14
	Tree2	Tree2	Decision ...	Attrition	Attrition	0.159184	735	0.129252	0.897143	160.5642	0.109227	0.330496	1470	735	735	0.159184	1	192.4779	0.130937	0.361853	14


ROC Chart : Attrition
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0


Score Rankings Overlay: Attrition
Cumulative % Captured Response
Data Role = TRAIN
Data Role = VALIDATE
Cumulative % Captured Response
Depth
0
20
40
60
80
100
Decision Tree
Regression Stepwise
Regression Backward
Regression Forward
Regression All Units
Neural Network
Gradient Boosting

3. Conclusion:

After running 7 statistical models on the processed dataset, we found that OverTime, BusinessTravel, StockOptionLevel are the three significant variables which have greater impact on an Employee leaving a company. Taking Confusion Matrix and Misclassification Rate. We concluded that for predicting the likelihood of attrition, **Regression with Selection Model None** is the best possible model.

4. References:

http://support.sas.com/documentation/cdl/en/emgsj/66018/HTML/default/viewer.htm#p03iy98sk0c9bvn1r6x7ppx8uj08.htm

https://support.sas.com/kb/24/205.html

https://support.sas.com/resources/papers/proceedings15/SAS1965-2015.pdf

http://support.sas.com/publishing/pubcat/chaps/57587.pdf