Business Analytics with SAS Project on

Finding Attrition rate in a company



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

**Instructor: Dr. James Zhang**

**Project Members (Group 3)**

**Venkata Sarath Chandra Muktevi**

**Vinay Manideep Cheviti**

**Vamshi Enishetty**

**Harshavardhan Yarlagadda**

Contents

[1. Data Mining Objective and Motivation: 3](#_Toc529898343)

[1.1 Objective: 3](#_Toc529898344)

[1.2 Motivation: 3](#_Toc529898345)

[1.3 Executive Summary: 3](#_Toc529898346)

[1.4 Data set: 3](#_Toc529898347)

[1.5 Data Preprocessing: 5](#_Toc529898348)

[2. Predictive Analysis: 10](#_Toc529898349)

[2.1 Decision Tree: 10](#_Toc529898350)

[2.2 Logistic Regression: 15](#_Toc529898351)

[2.2.1 Logistic Regression with Selection Model as None: 15](#_Toc529898352)

[2.2.2 Logistic Regression- Forward Regression: 18](#_Toc529898353)

[2.2.3 Logistic Regression – Backward Regression: 20](#_Toc529898354)

[2.2.4 Logistic Regression – Step-wise Regression: 22](#_Toc529898355)

[2.3 Neural Network: 24](#_Toc529898356)

[2.4 Gradient Boosting: 26](#_Toc529898357)

[2.5 Final Model: 28](#_Toc529898358)

[2.5.1 Model Comparison: 28](#_Toc529898359)

[3. Conclusion: 30](#_Toc529898360)

[4. References: 31](#_Toc529898361)

# Data Mining Objective and Motivation:

# 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.

# 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.

# 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.

# 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.

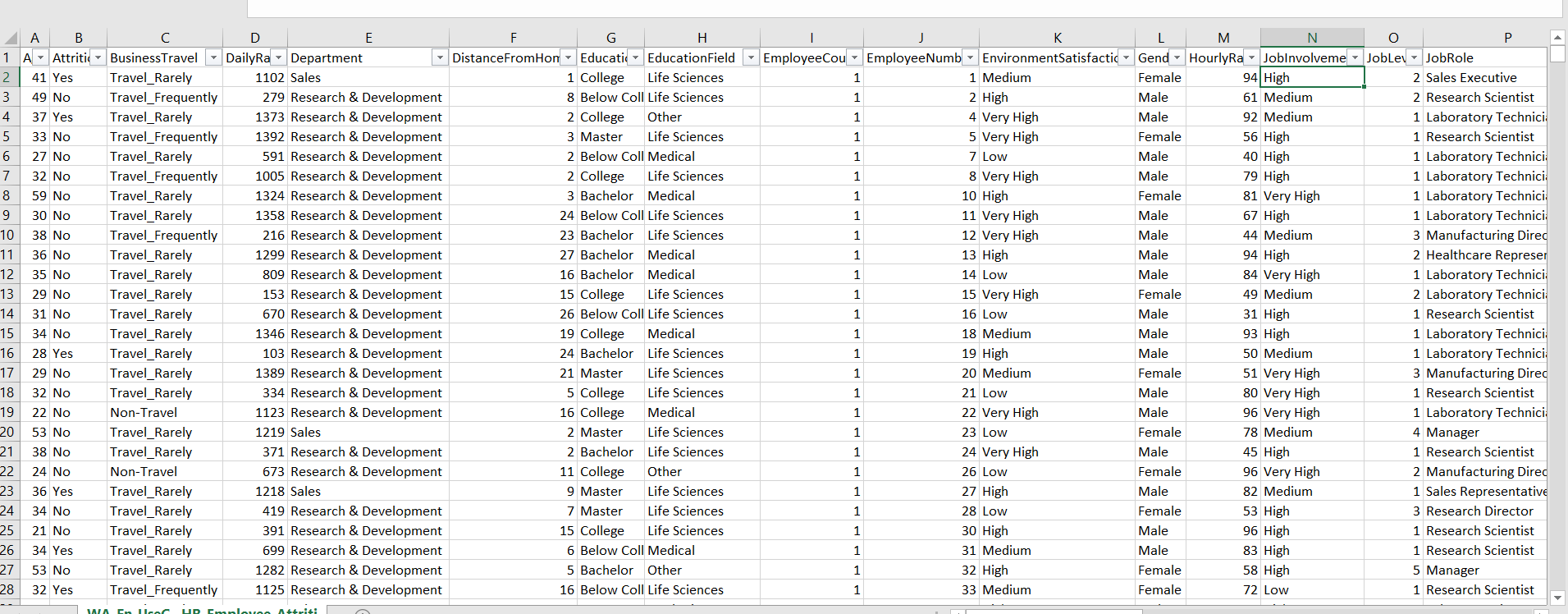
* TIME: Yearsatcompany, Yearsincurrentrole, YearsSinceLastPromotion, YearsWithCurrManager
* EMPLOYEE DETAILS: Age, Education, EducationField, Gender, WorkLifeBalance, DistanceFromHome, JobSatisfaction, MaritalStatus, NumCompaniesWorked
* INCOME: MonthlyIncome, MonthlyRate, DailyRate, HourlyRate, StockOptionLevel,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

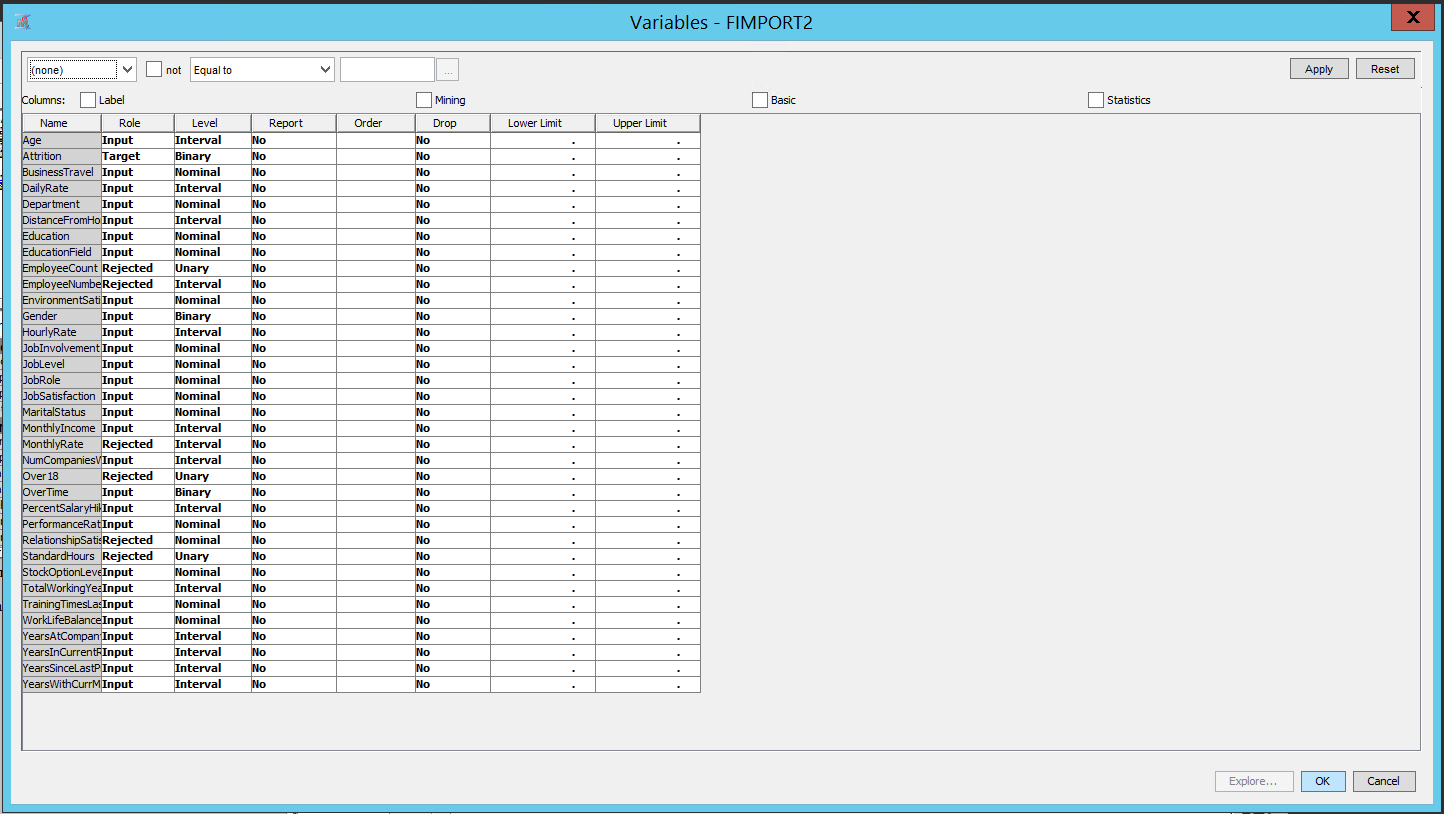
|  |  |  |
| --- | --- | --- |
| Age | Age of employee | INTERVAL |
| Attrition | Weather employee quits or not | BINARY |
| BusinessTravel | Frequency of Travel | NOMINAL |
| DailyRate | The amount of money employees is paid per day | INTERVAL |
| Department | Name of the department employee work | 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 |

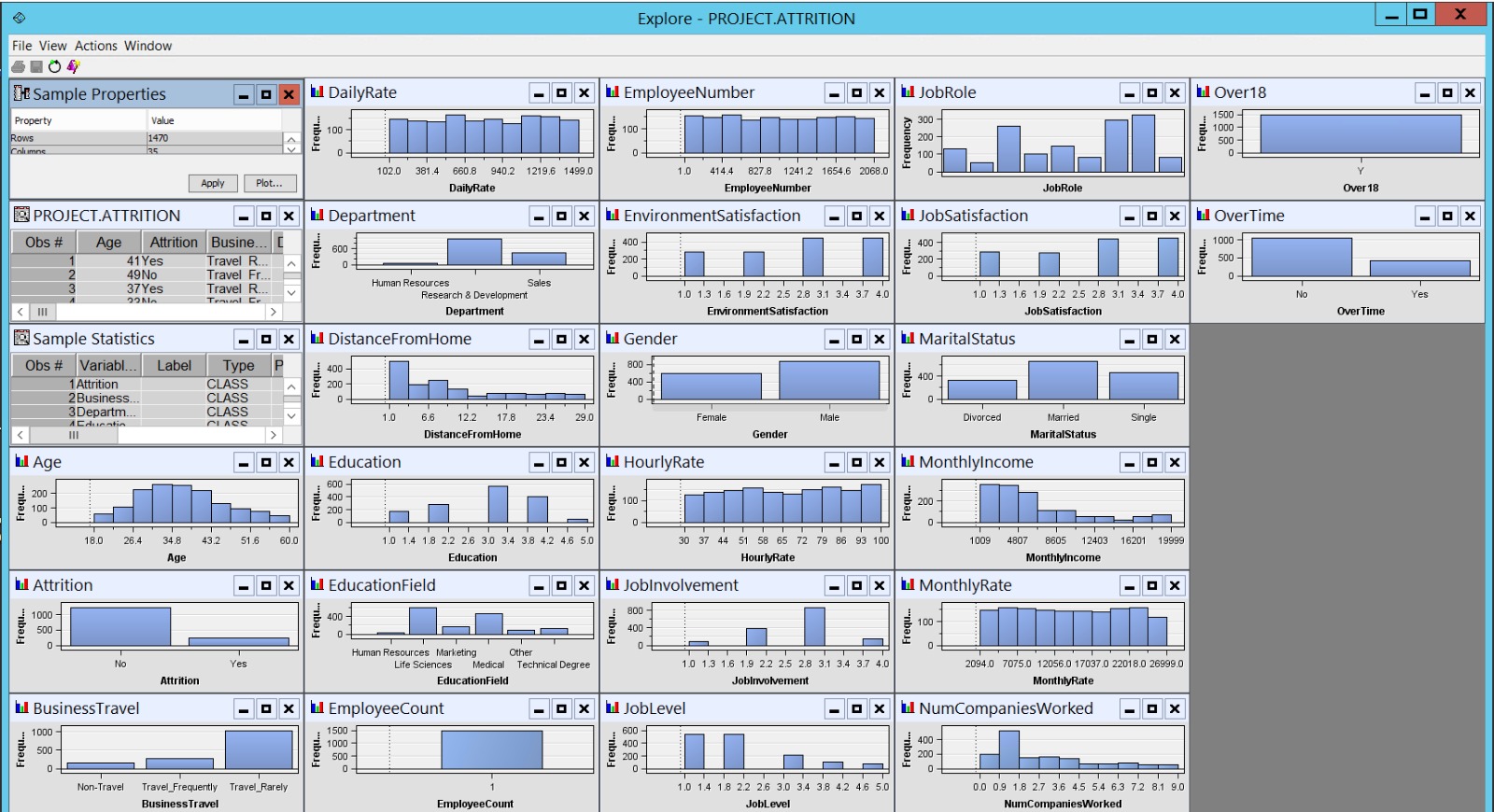
# Data Preprocessing:

Following steps are performed for Data Preprocessing –

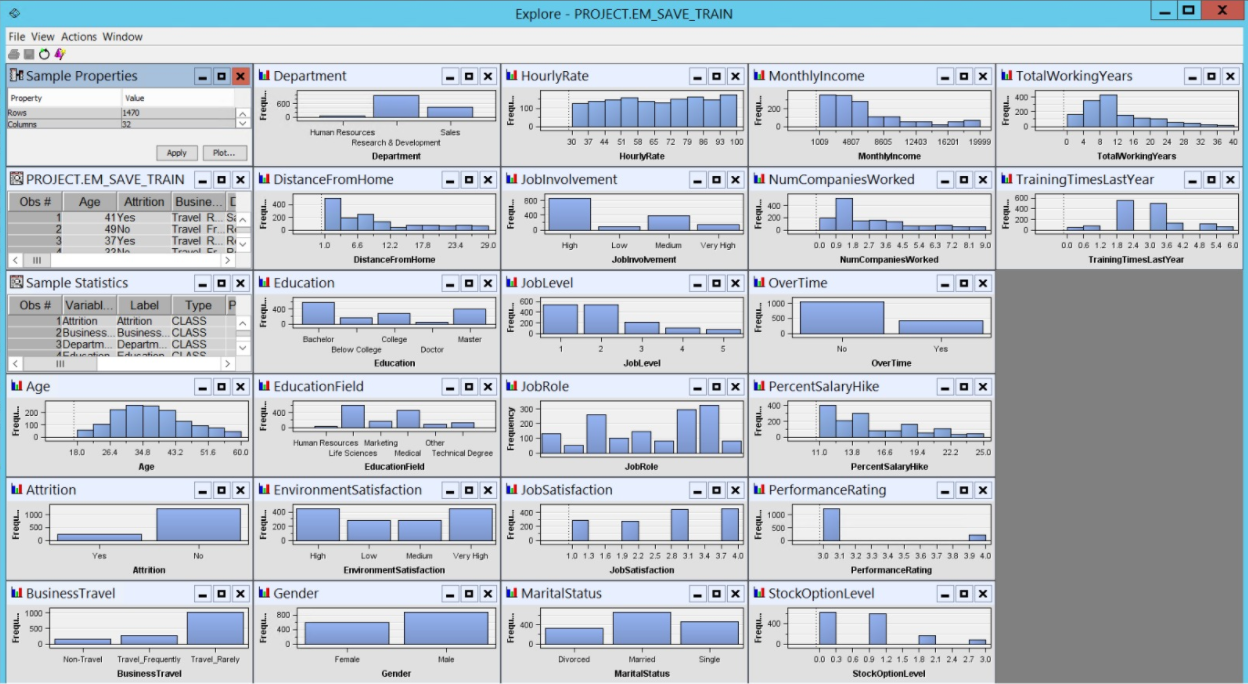
1. We changed the data format of Education Column from 1,2,3,4,5 to 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
2. We changed the data format of Environment Satisfaction from 1,2,3,4,5 to 1 'Low' 2 'Medium' 3 'High' 4 'Very High'.
3. We changed the data format of Job Involvement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





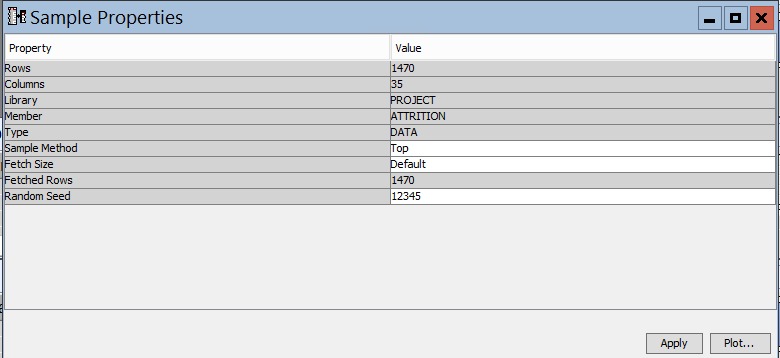
Before Preprocessing

After Preprocessing



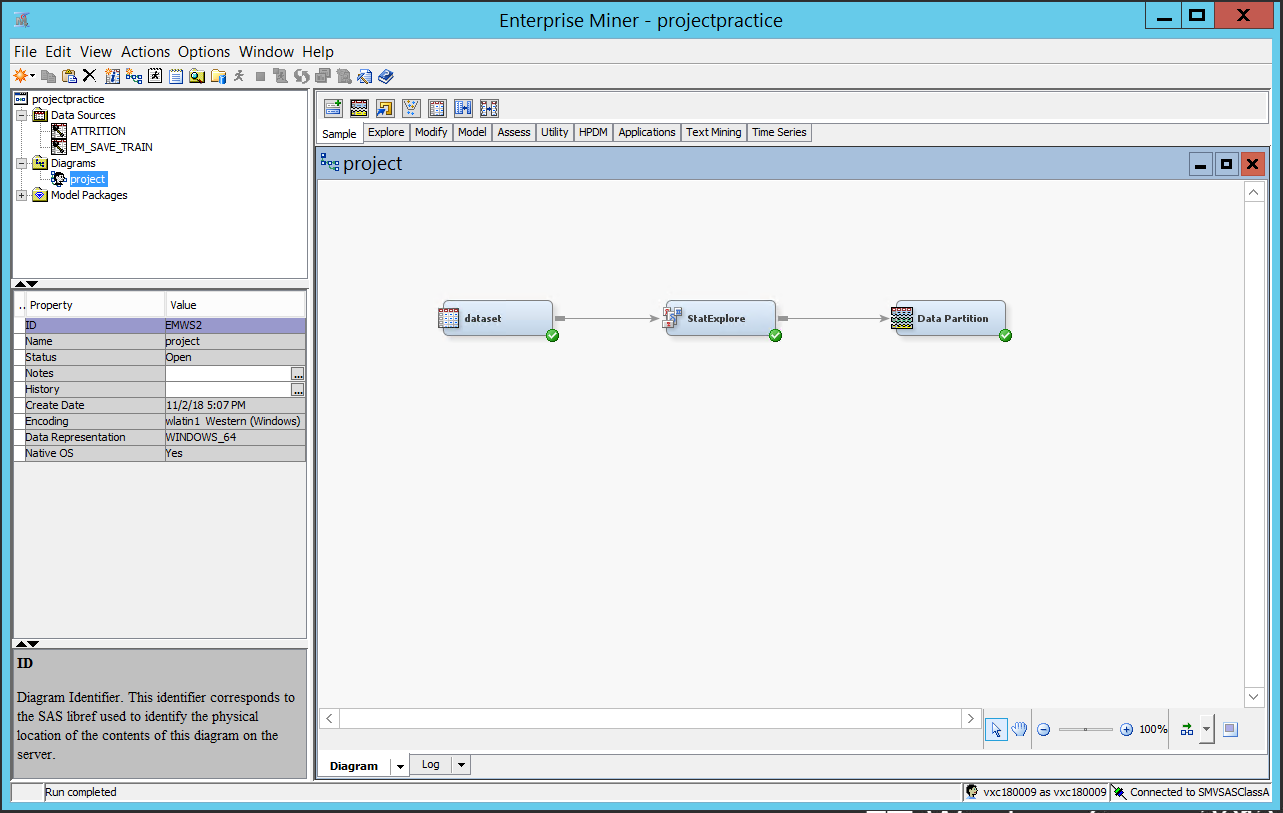
1. Data summary

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



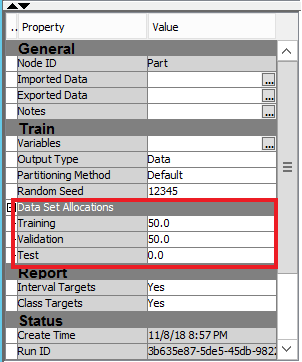
1. Final Data Preprocessing:

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.



1. 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.



# Predictive Analysis:

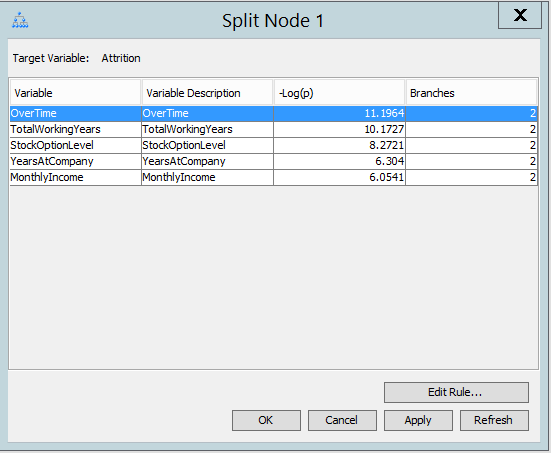
Attrition rate is predicted based on different input variables. Since our target variable – attrition is binary Yes/No, we have used four different models for predictive analysis:

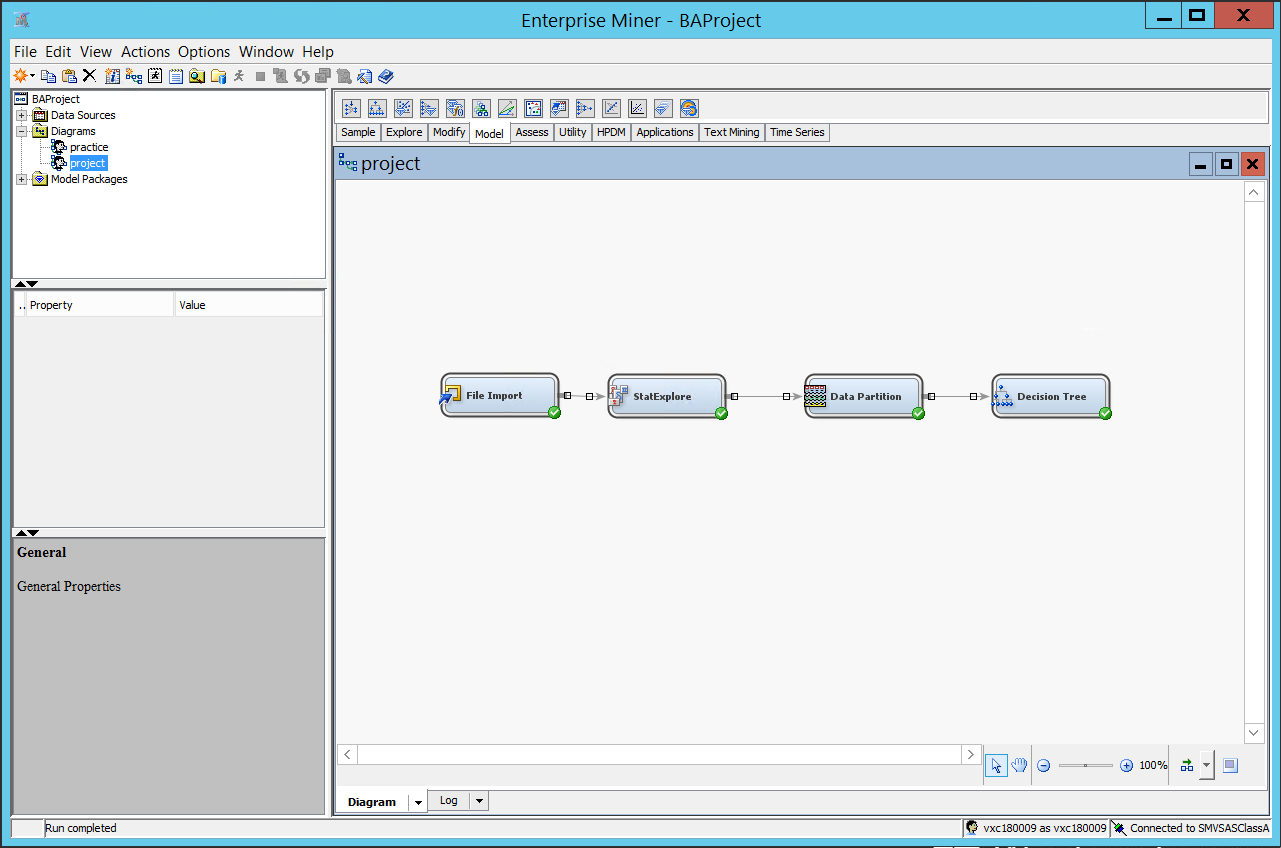
1. **Decision Tree**
2. **Logistic Regression**
3. **Neural Network**
4. **Gradient Boosting**

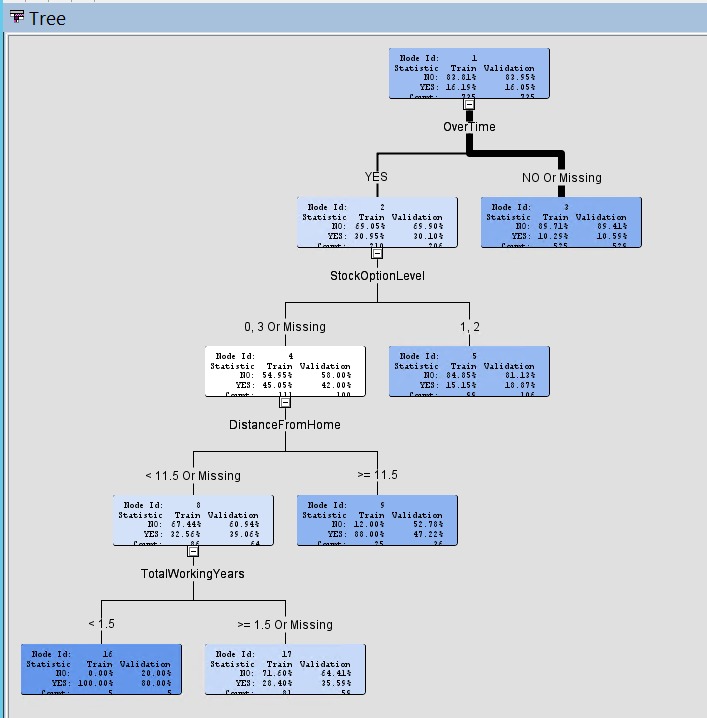
# Decision Tree:

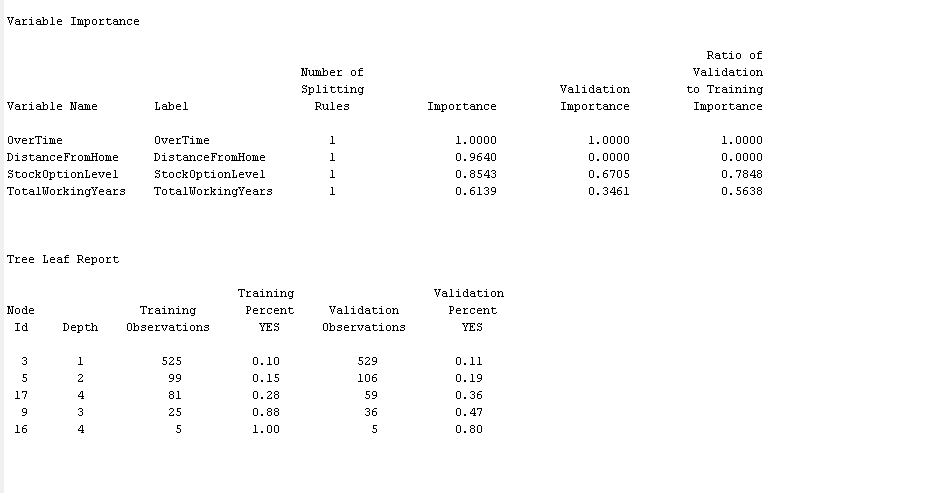
To find the important variables in the Dataset we ran decision tree on processed data. We selected the variables based on the ‘Variable Importance’ table in the output which are Overetime, Totalworkingyears, stockoptionlevel, Yearsatcompany, MonthlyIncome

the results are as follow:



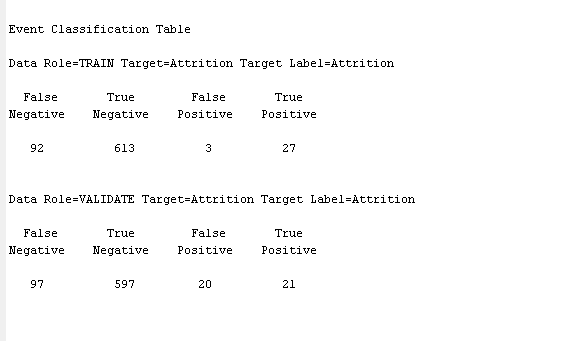






Confusion Matrix:

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

****

From above we can plot confusion matrix as below -

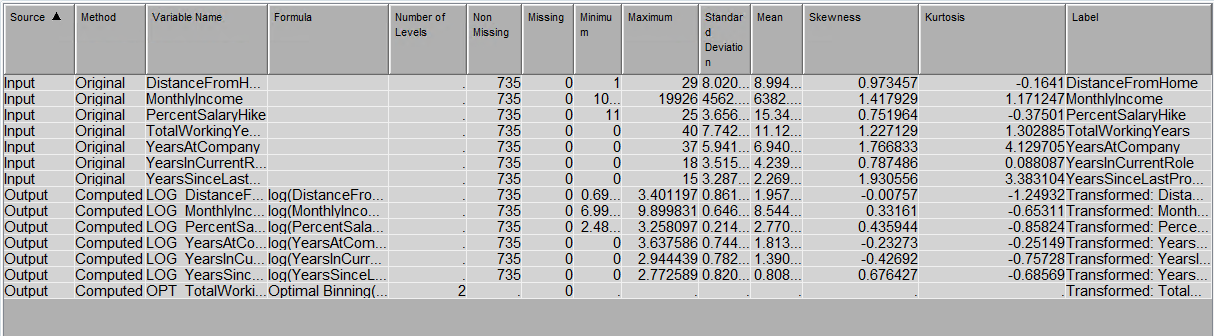
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | **Positive** | **Negative** |
| **Actual** | **Positive** | 21 | 20 |
| **Negative** | 97 | 597 |

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

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

# Logistic Regression:

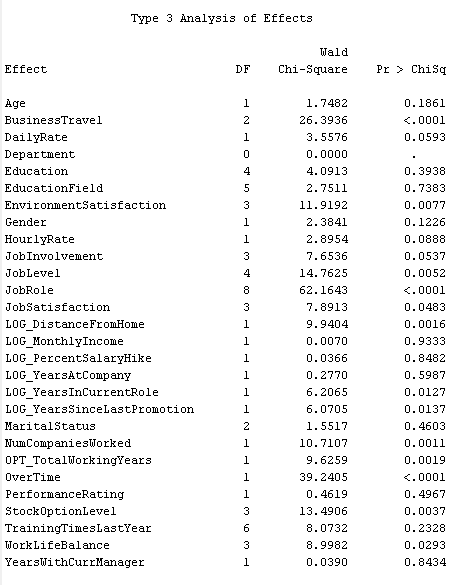
The skewness in the data set will affect the overall prediction of Attrition rate. To avoid it, we performed some data transformations on the variables which affect the target variable to reduce the skewness of the data. For performing the transformation, we used the available Transformation node in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

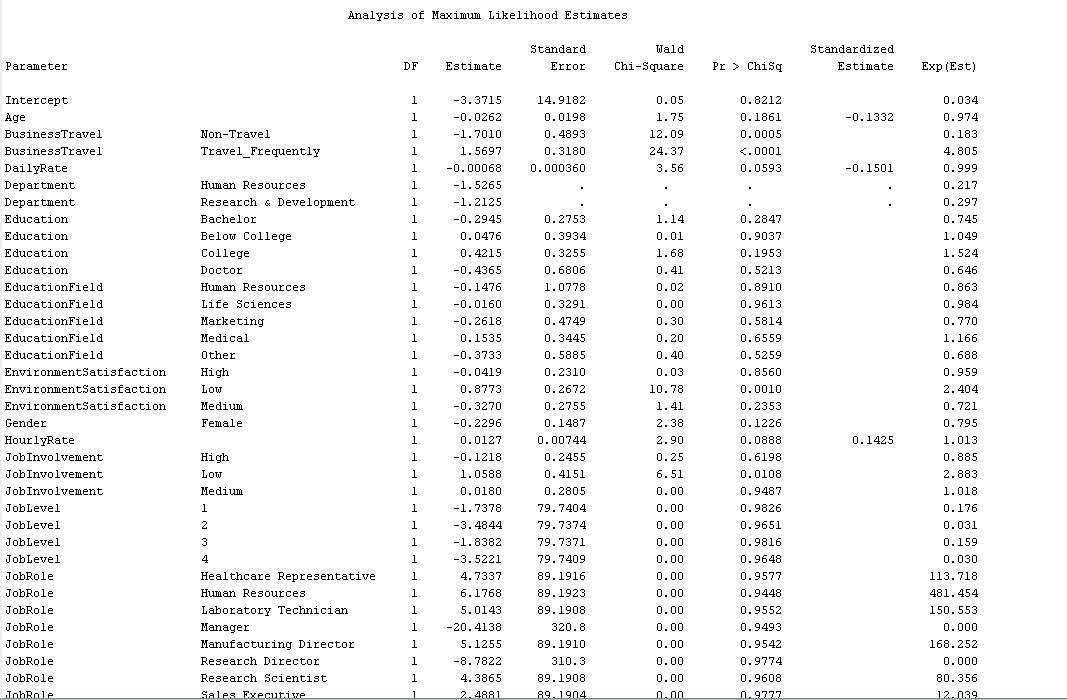


# Logistic Regression with Selection Model as None:

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:

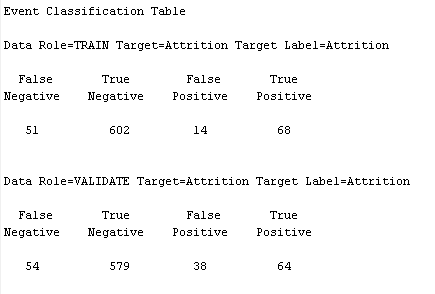




By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, JobRole, OverTime.

Confusion Matrix:

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



From above we can plot confusion matrix as below.

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

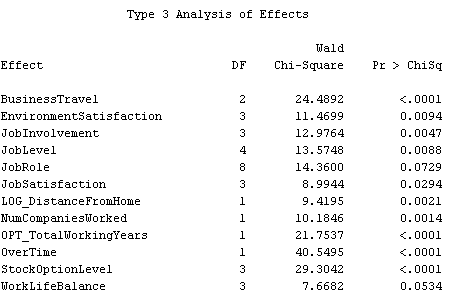
We can find the Accuracy of the model from Confusion matrix using following formula.

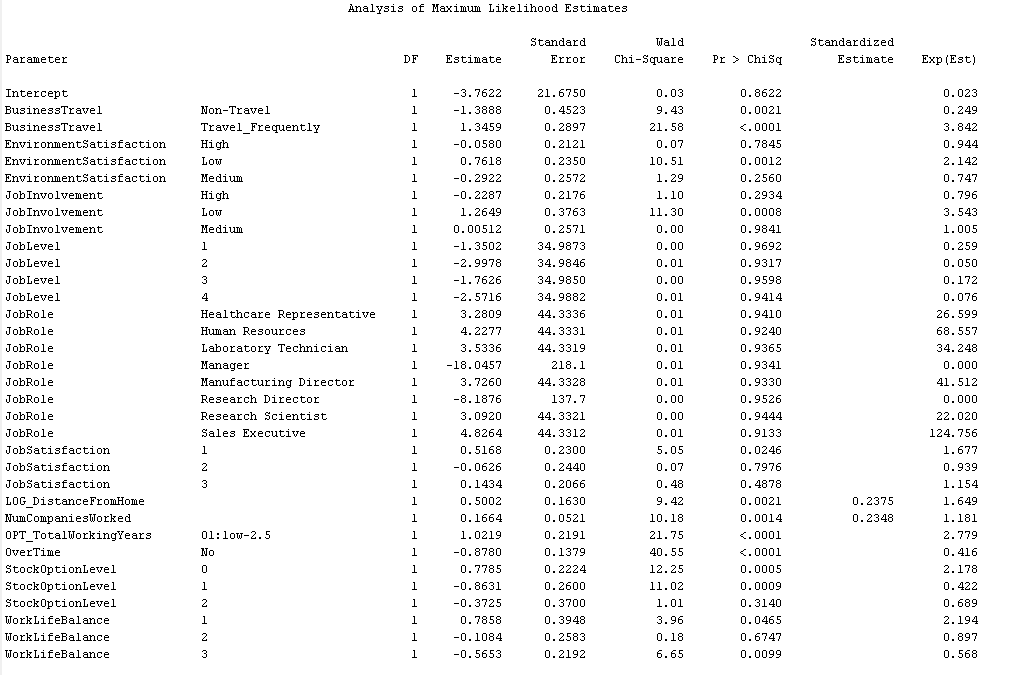
Accuracy of the model is 0.8748 i.e. 87.48%

# 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:

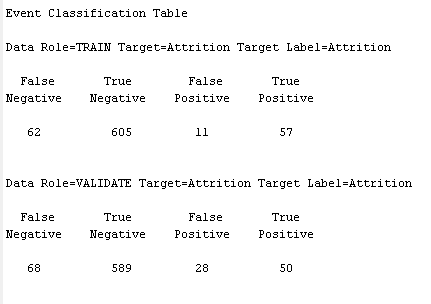




By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, OPT\_TotalWorkingYears, OverTime, StockOptionLevel.

Confusion Matrix:

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



From above we can plot confusion matrix as below.

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

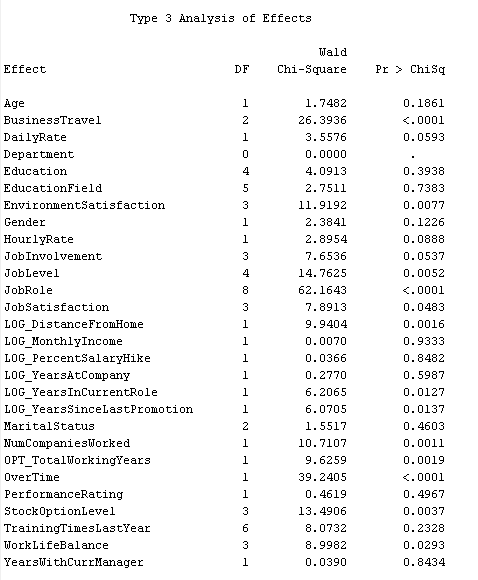
We can find the Accuracy of the model from Confusion matrix using following formula.

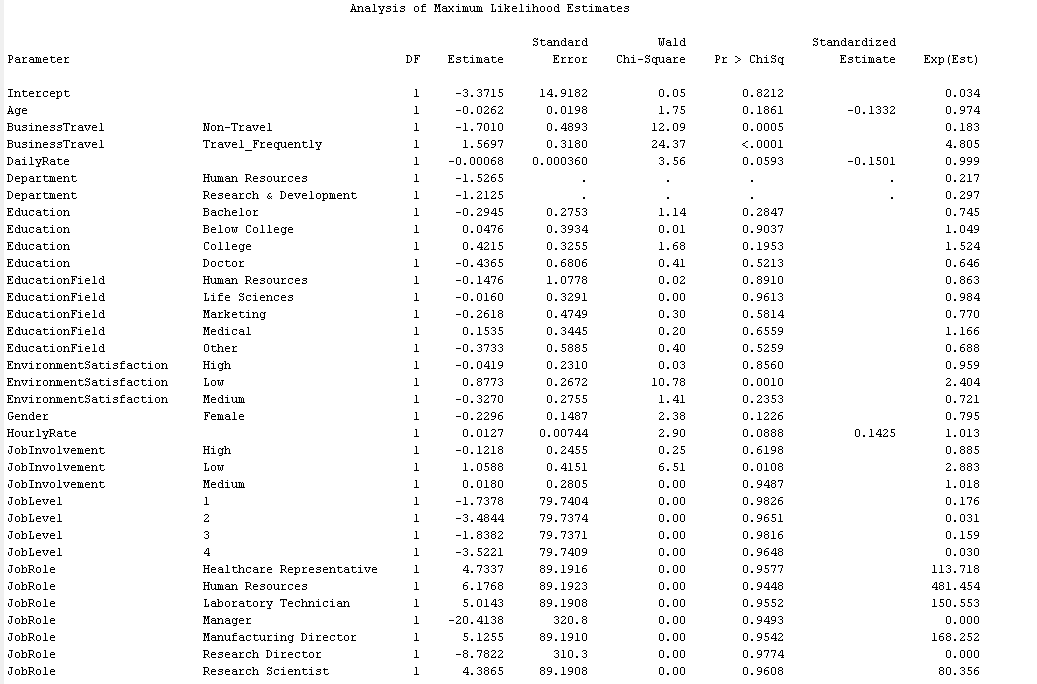
Accuracy of the model is 0.8694 i.e. 86.94%

# 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:

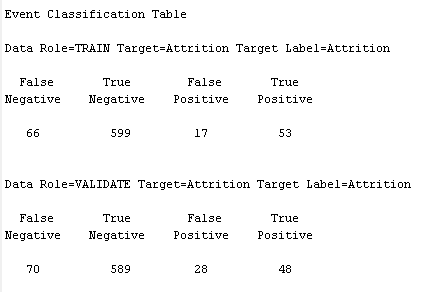




By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, JobRole, OverTime.

Confusion Matrix:

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



From above we can plot confusion matrix as below.

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

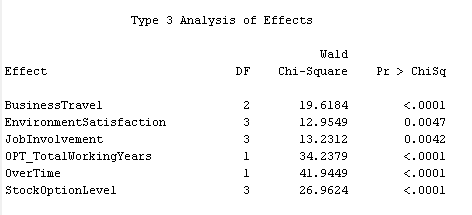
We can find the Accuracy of the model from Confusion matrix using following formula.

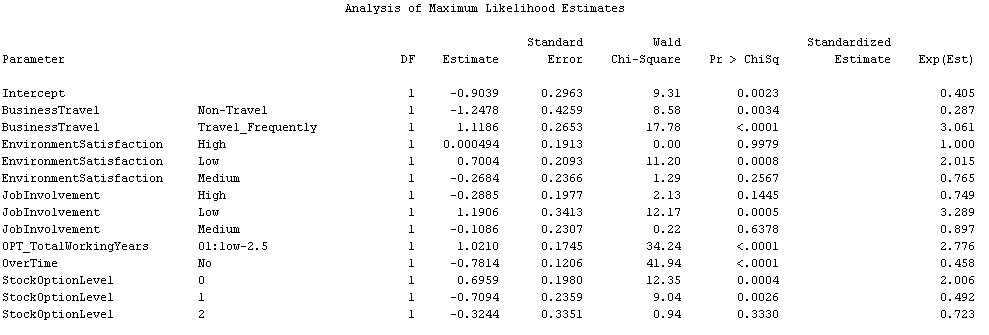
Accuracy of the model is 0.8667 i.e. 86.67%

# 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:

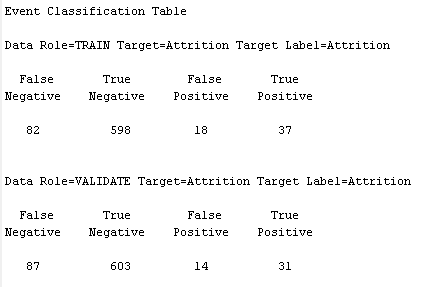




By observing the values under column Pr > ChiSq, we can conclude most significant variables – BusinessTravel, OPT\_TotalWorkingYears, OverTime, StockOptionLevel.

Confusion Matrix:

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



From above we can plot confusion matrix as below.

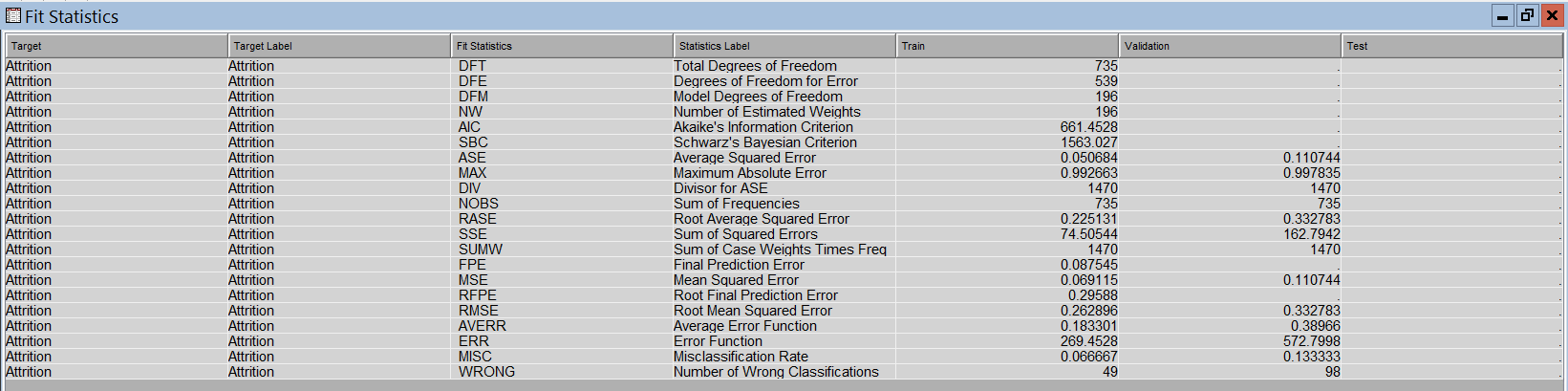
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | **Positive** | **Negative** |
| **Actual** | **Positive** | 31 | 14 |
| **Negative** | 87 | 603 |

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

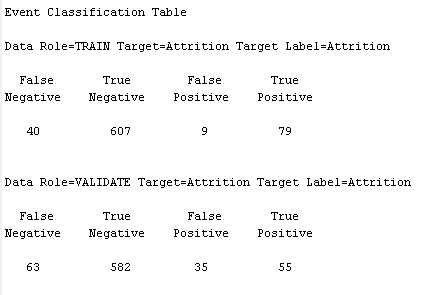
Accuracy of the model is 0.8626 i.e. 86.26%

# Neural Network:

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



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



From above we can plot confusion matrix as below.

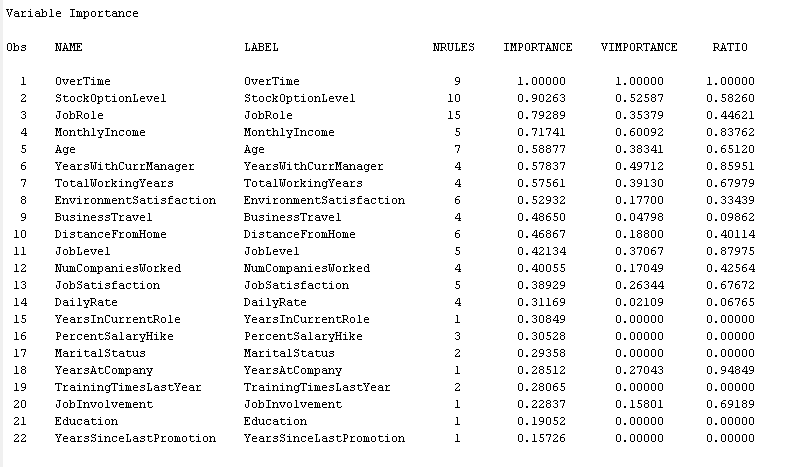
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | **Positive** | **Negative** |
| **Actual** | **Positive** | 55 | 35 |
| **Negative** | 63 | 582 |

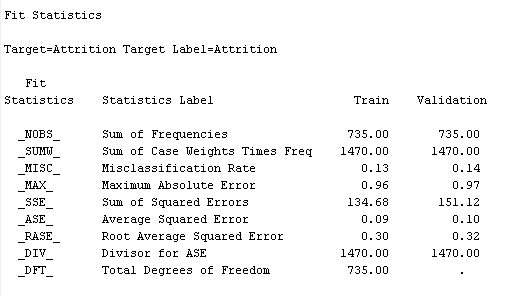
We can find the Accuracy of the model from Confusion matrix using following formula.

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

# Gradient Boosting:

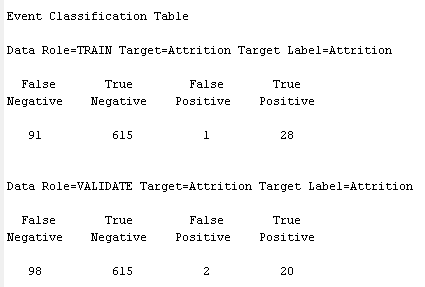
To find the important variables we ran Gradient Boosting on processed data. We selected the variables based on the ‘Variable Importance’ table in the output which are OverTime, StockOptionLevel, JobRole, MonthlyIncome, Age. We found the results as follow:





Confusion Matrix:

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



From above we can plot confusion matrix as below -

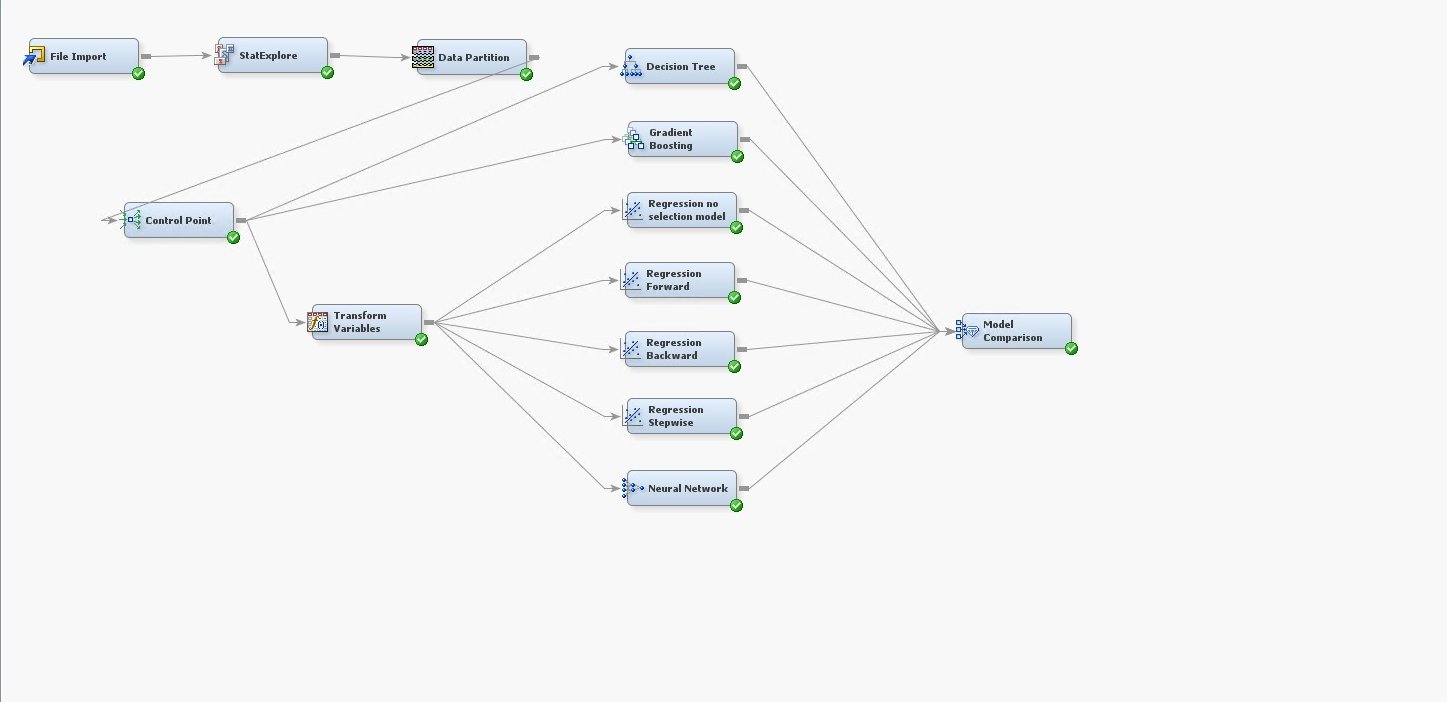
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | **Positive** | **Negative** |
| **Actual** | **Positive** | 20 | 2 |
| **Negative** | 98 | 615 |

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

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

# Final Model:

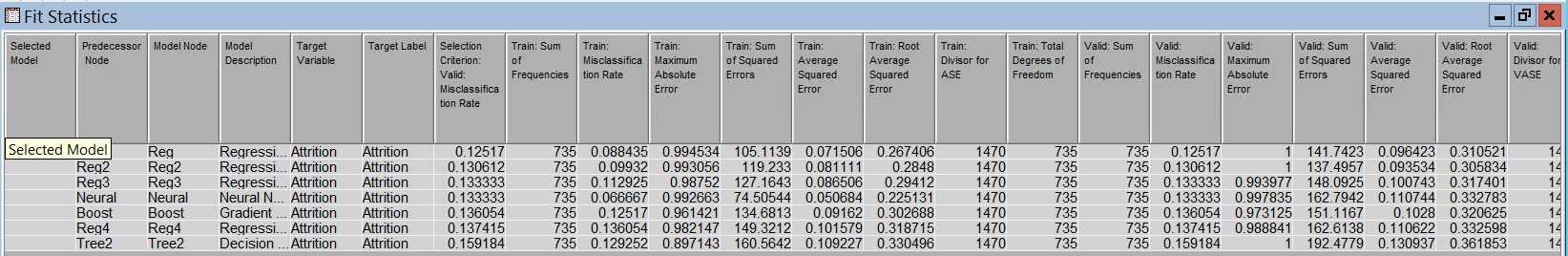
Following snapshot shows the final model. It involves all the nodes used for Data preprocessing, Descriptive Analysis and Predictive Modeling.

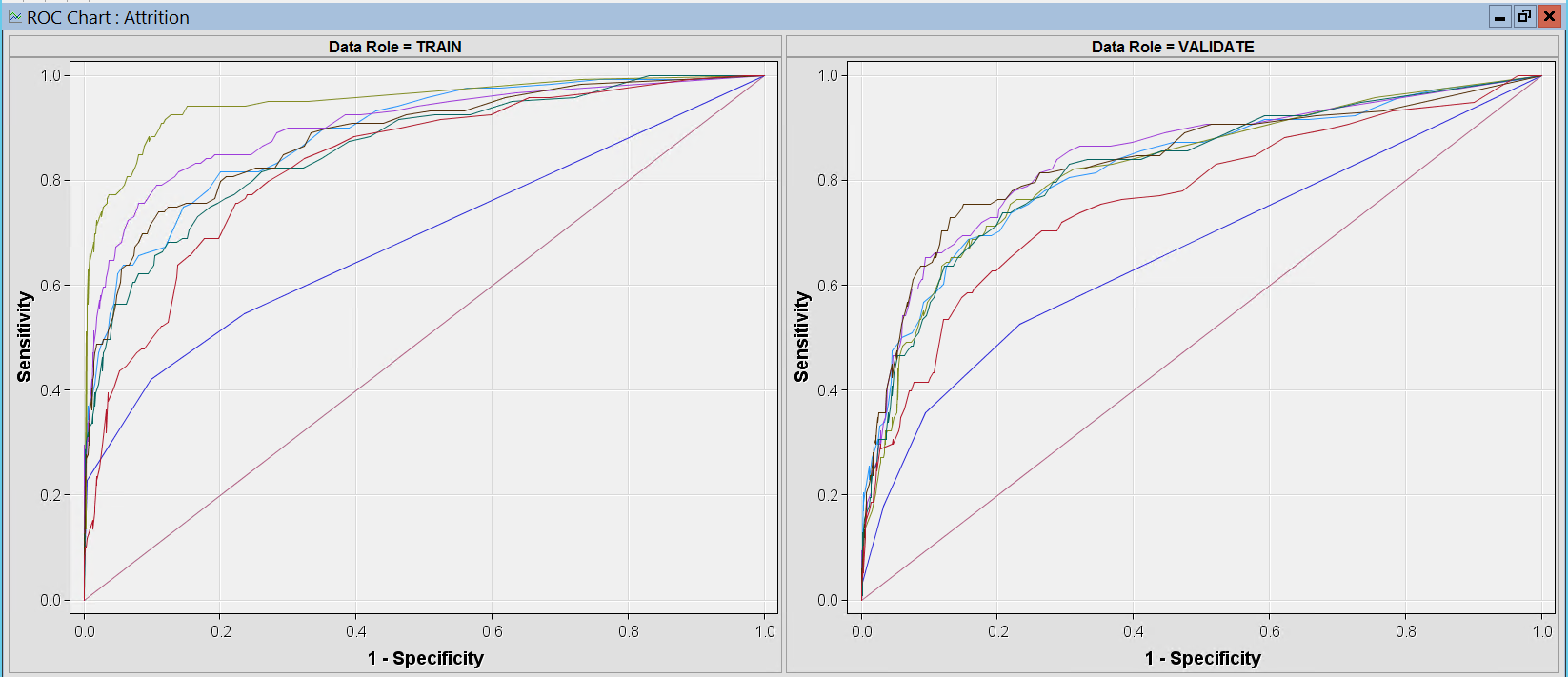


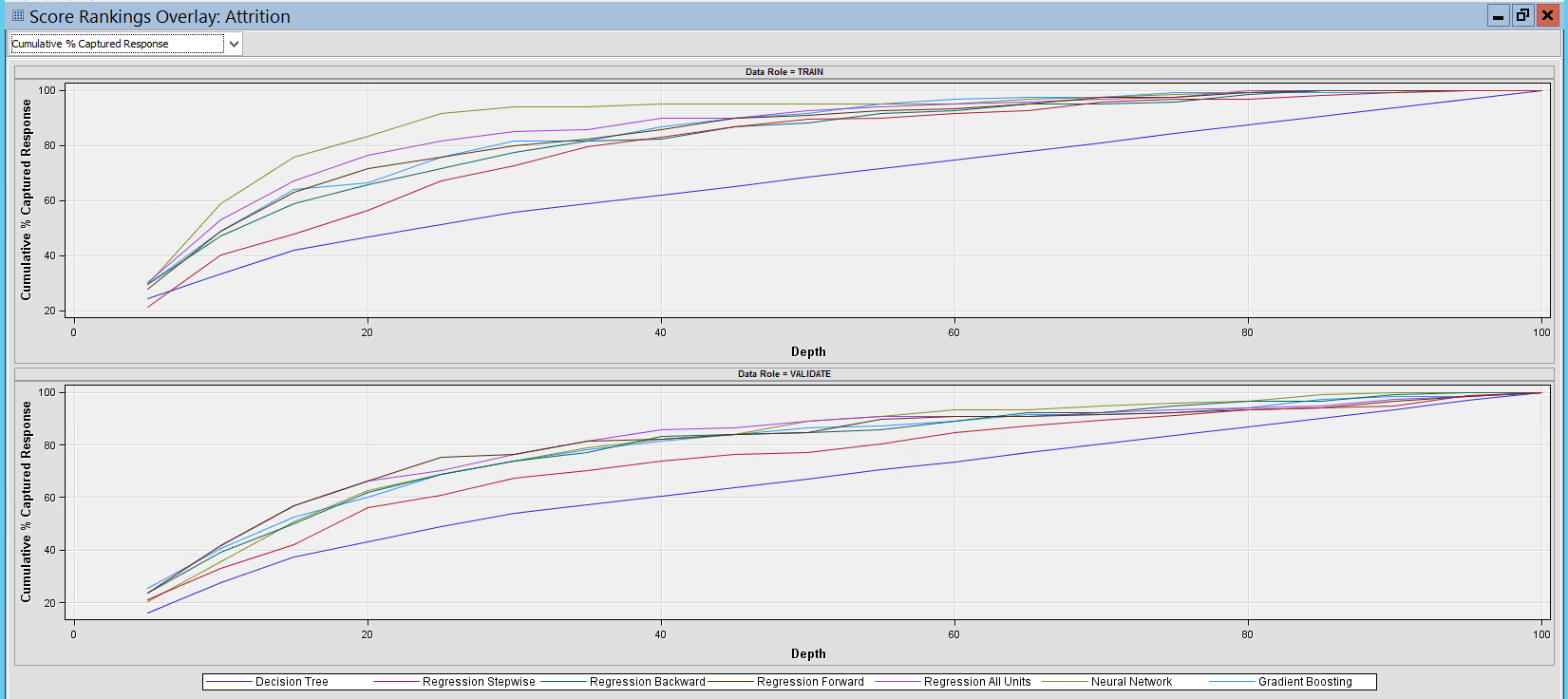
# Model Comparison:

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 –







# 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.

# 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>