**Aathira Puthenpurayil -P2736418**

Data Mining Problem Assignment

FRAMINGHAM DATASET

**Table of contents**

Table of Contents

[Abstract 2](#_Toc135192013)

[Business Problem 2](#_Toc135192014)

[Methodology - SEMMA framework 2](#_Toc135192015)

[Data Exploration 2](#_Toc135192016)

[Data partition creation of model sets 5](#_Toc135192017)

[Data Modification 6](#_Toc135192018)

[Data Modelling 8](#_Toc135192019)

[Regression 8](#_Toc135192020)

[Development of models 8](#_Toc135192021)

[Model performance 10](#_Toc135192022)

[Chosen Regression equation. 11](#_Toc135192023)

[Decision Tree 12](#_Toc135192024)

[Development of models 12](#_Toc135192025)

[Performance of models 13](#_Toc135192026)

[Critical path of best model 15](#_Toc135192027)

[Overfitting analysis 15](#_Toc135192028)

[Analysis of the best model 15](#_Toc135192029)

[A discussion of how the chosen topic from Assessment 1 might be useful for cardiovascular disease research. 17](#_Toc135192030)

[Conclusion 18](#_Toc135192031)

[Recommendation 18](#_Toc135192032)

[My Reflections on the process – What did I learn from this exercise? 18](#_Toc135192033)

[References and Bibliography 18](#_Toc135192034)

[Appendix 19](#_Toc135192035)

## Abstract

Cardiovascular Disease is a major cause of death in US. The Framingham data study has been conducted to analyse the risk factors. As the data miners, task is to analyse the provided dataset to predict best model for CVD analysis.

The report focuses on the methodology used, including data pre-processing, exploratory data analysis, feature selection, and modelling techniques. Model performance was evaluated using appropriate metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (ROC).

The technical report concludes by recommending the best model so that the NHS can be use them to have better insight for the dataset for the development of prevention and targeting purposes.

# Business Problem

The data set provided for analysis is FRAMINGHAM dataset with 5209 observations and 13 variables. The key challenge for NHS includes the identification of risk factors as well as to find good model to predict better insights. By addressing these issues, it will help the NHS to prevent the cardiovascular disease by proving early treatment and taken precautions against it in society. There are 11 input, death age as rejected and status as target variable in the dataset. As a data analyst needs to build a best model to predict the risk of status (death or alive).

# Methodology - SEMMA framework

SEMMA stands for sample, Explore, Modify, Model and Assess. Sample node can be used to import the data, data partition, filter, merge and append. Explore node is used for exploratory data analysis to find the missing values, outliers as well as to visualize the data through charts, summary statistics and graphs. Modify is used to perform feature engineering, addressing missing values through imputations. Model node is used to select appropriate model for CVD and train the model and evaluate the model performance metrices such as accuracy, precision, recall, and area under the receiver operating characteristic curve (ROC). Finally, the Asses model is used to do the model comparison.

# Data Exploration

Firstly, import Framingham dataset using file import node.

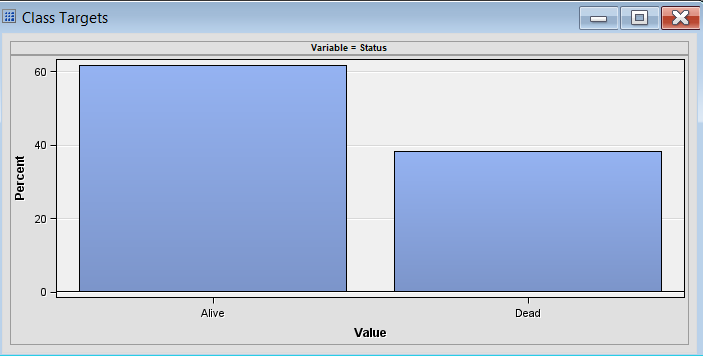


Figure : Target variable

This is an imbalanced dataset as the number of defaults is lower than the number of non-defaults. So, need to consider sampling and performance matrix accordingly.

**Missing values**

* death age - 61.78%
* Cholesterol and cholesterol status - 2.92%
* Height, smoking status, smoking, weight and weight status < 1%
* Bpressure status, diastolic, status and systolic - 0%

Weight is most normally distributed. Systolic has most skewed. Systolic is having fattest tails as it shows high kurtosis.



Figure : Explore Graph

**Correlation**

PCA needs to be considered if there is any correlation found.

A screenshot of a computer

Description automatically generated with medium confidence

Figure : Correlation Matrix

**The Multiplot, Graph Explore and Stats Explore nodes**

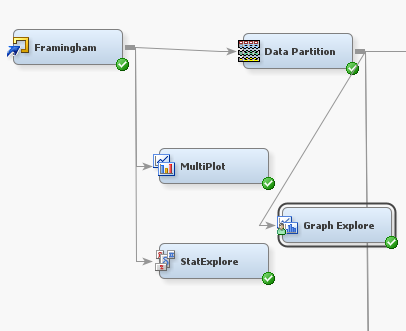


Figure : Graph Explore nodes

Multiplot node automatically creates charts with input and target variable.

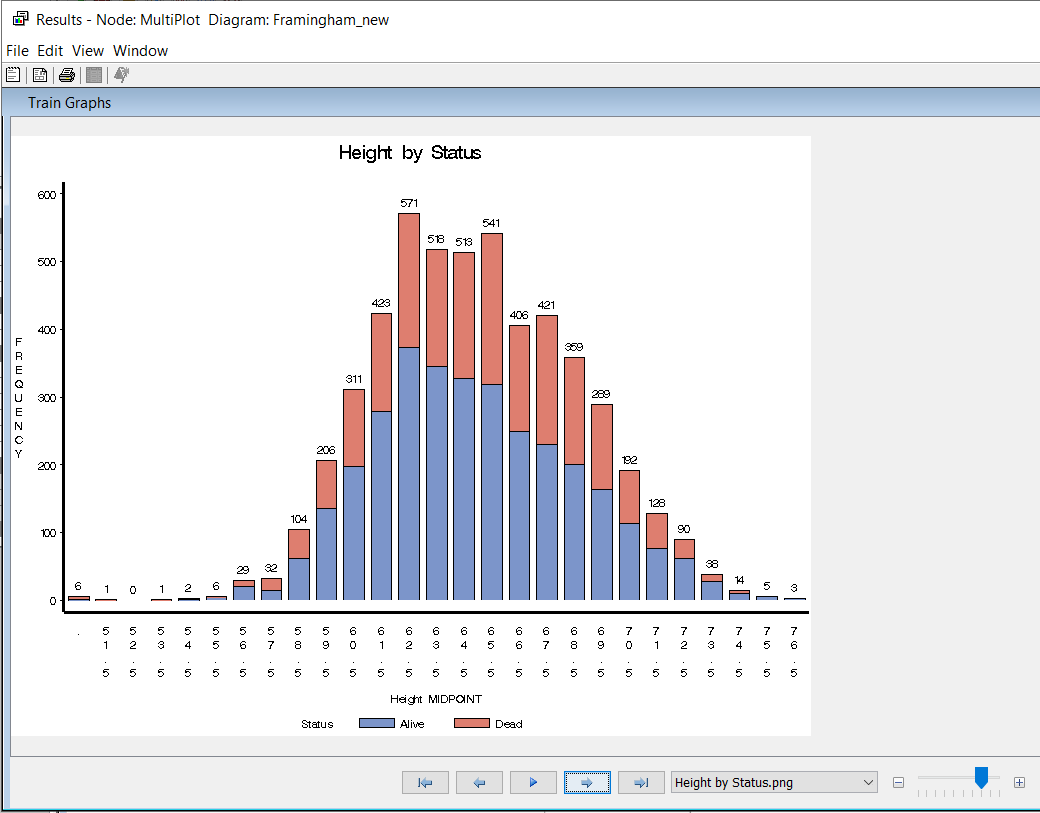


Figure :Multiplot

Graph explore mode used to explore samples of data through different graphs. Sample method of Random is provided. There is new column created called dataobs

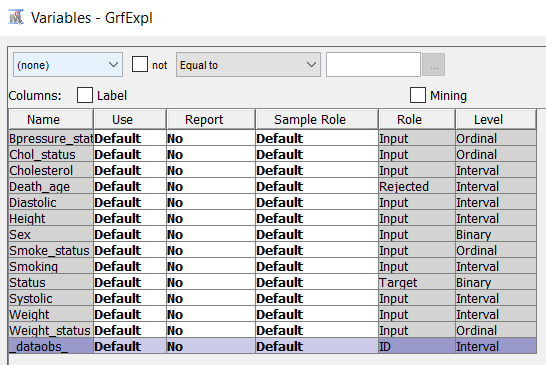


Figure : Graph explore

StatExplore is used to find which input variables are closely related to the target. Here Systolic is closely related to target followed by diastolic and Bpressure status. Height and weight status are least related to the target variable.

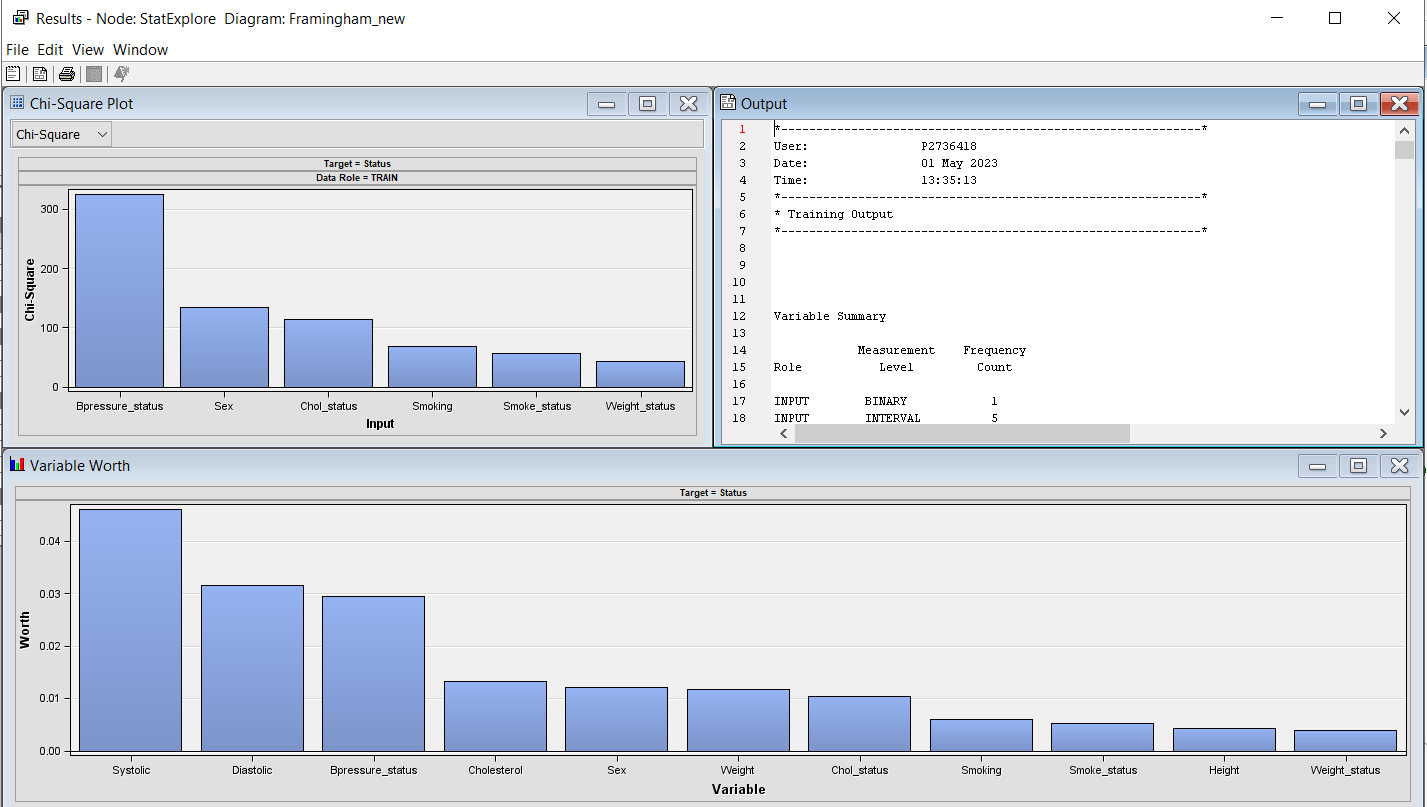


Figure : Stat Explore

Based on the above information tried plotting the box plot to get more detailed of closely related inputs systolic and diastolic. People with systolic has higher chance of death due to CVD followed by diastolic. Similarly various other chart types can also be used to explore the data in more graphical representation.

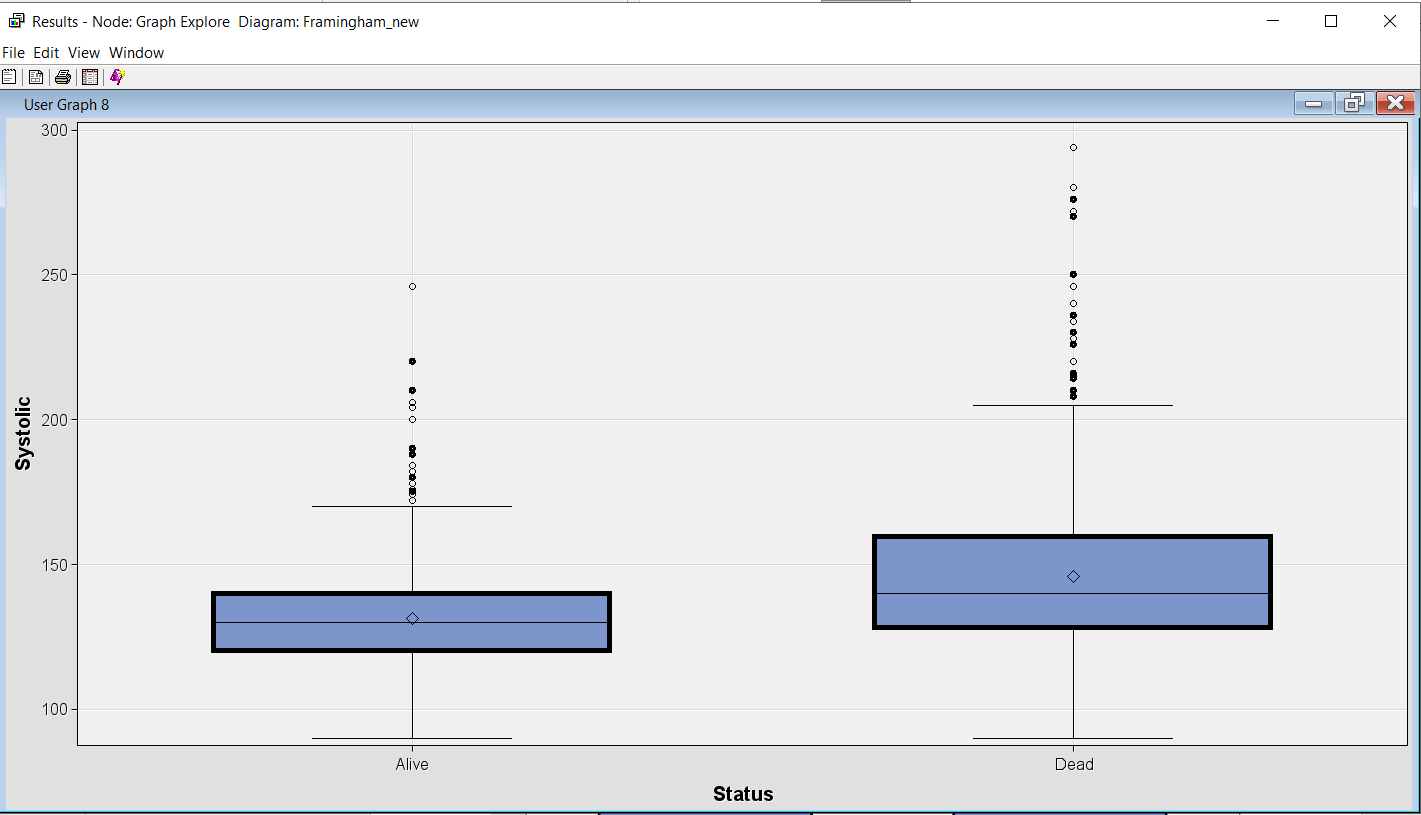
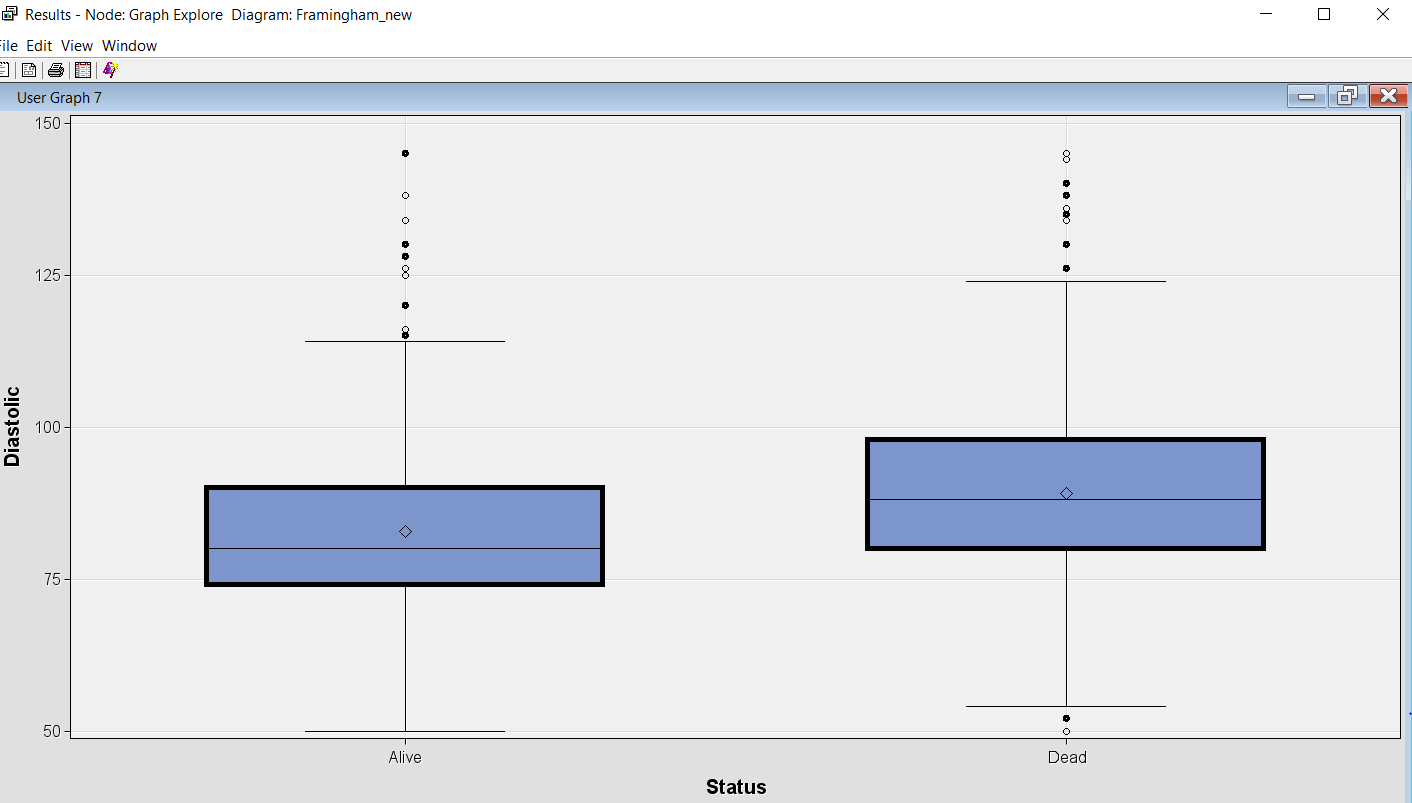


Figure : BOX PLOT

# Data partition creation of model sets

Data partition splits the sample data into training, test and validation with 40 ,30 and 30. By default the target is already identifies so giving data as stratified on target value.



Figure : Data Partition node

Stratified sampling is used as the dataset is imbalanced

## Data Modification

**Impute node**

Data Modifications can be done to deal with any outliers or with the skewed distributions. As we have rejected death age which has most missing values and the remaining variables has very few it is not necessary to use filter node which will remove the outliers. Instead of filter node impute node can be used to fill the missing values to improve model.

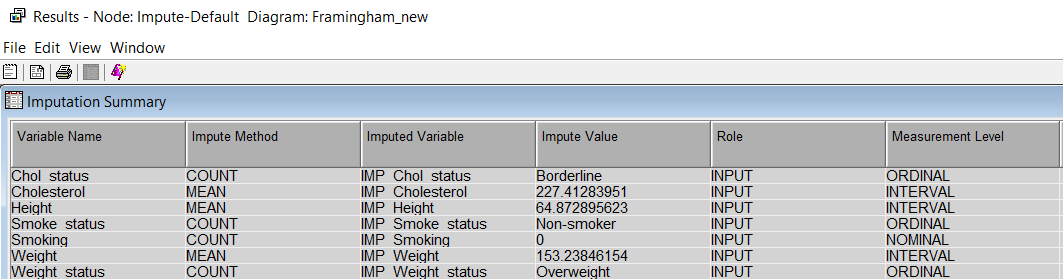


Figure : Impute default

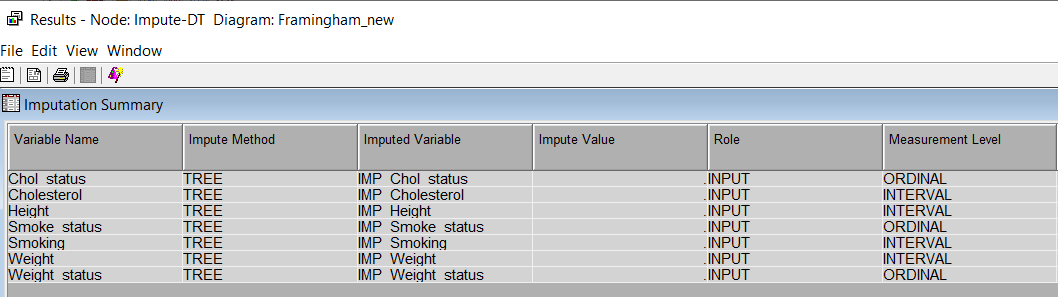


Figure : Impute Decision tree

**Transform node**

Skewness can be deal with transform variable to make the skewed variable to a normal distribution. Here 2 variables Systolic with skewness has been transformed using log function and smoking contains most of the values zero, it is transformed using smoking>0 and changed variable type to binary.

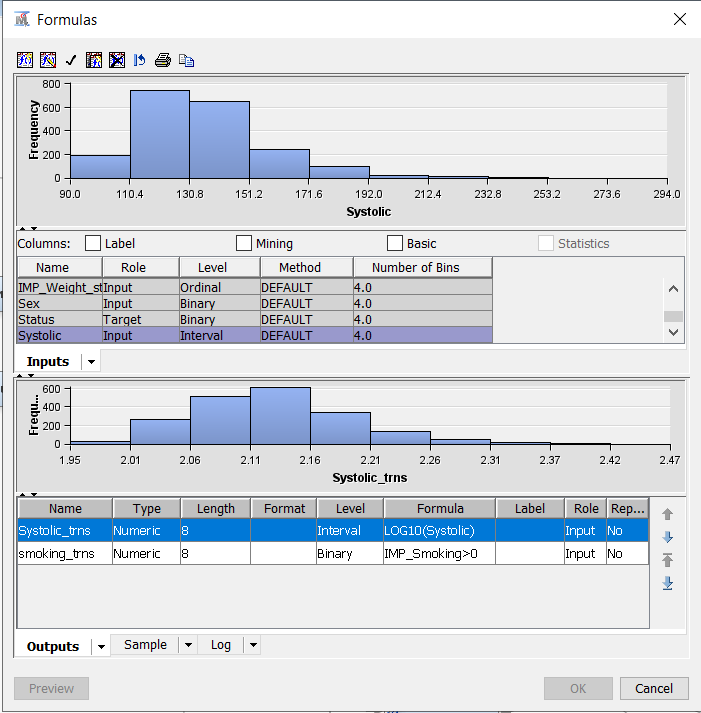


Figure : Transform variable for systolic

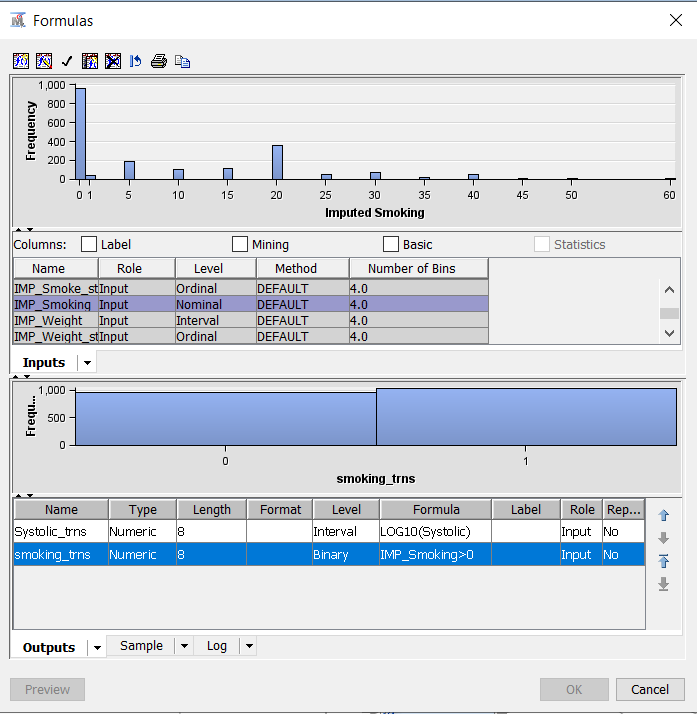


Figure : Transform variable for smoking

**Variable selection node.**

It is used to make preliminary selection of variables that are closely related to target variables. Rsquare selection method is used and only selected the closely related variables such as systolic, Bpressure status and imp smoke status.

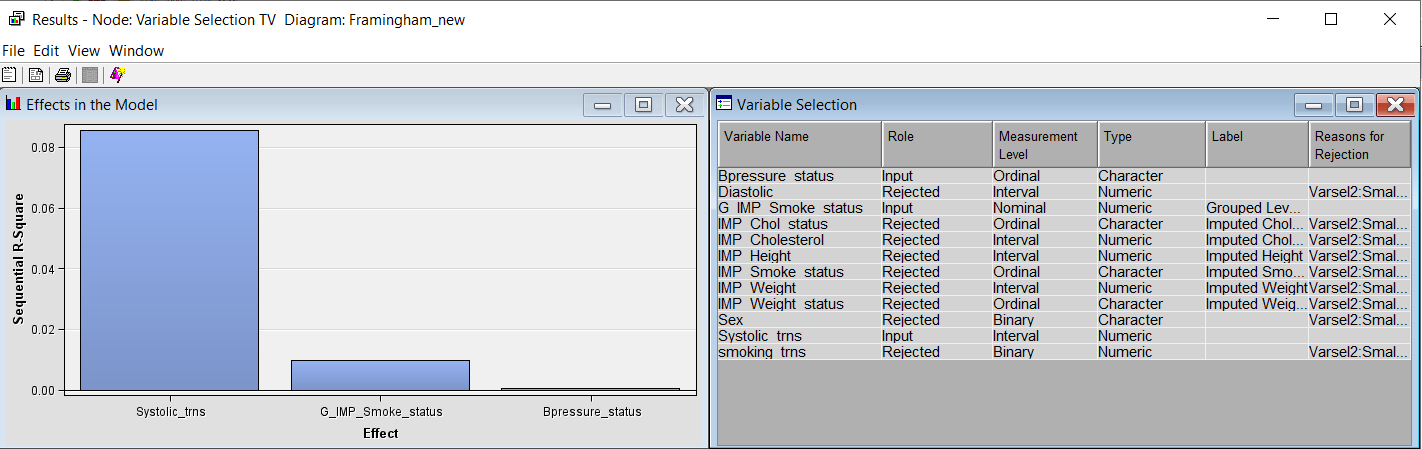


Figure : Variable selection

**Principal Component node.**

This node will help to reduce the number of variables in the dataset by retaining only the important information. The DMEURL procedure shows majority of the observations have high (42%) or normal (42%) blood pressure and optimal blood pressure (14%) is less common. For cholesterol borderline (37%) and high (34%) have most of the observations and desirable (27%) is less relevant. Non-smokers make up the largest portion of the observations (48.32%), followed by heavy smokers (20.57%), moderate smokers (11.05%), light smokers (11.43%), and very heavy smokers. Overweight individual (68%) has most of the observations followed by normal weight and underweight. Females (54%) are more compared to male (45%).

Eigen values shows in descending order. Higher the eigen values, there will be more components.

Summary shows there are maximum of 11 principal components is taken as input variables and shows a total variation of 93%.

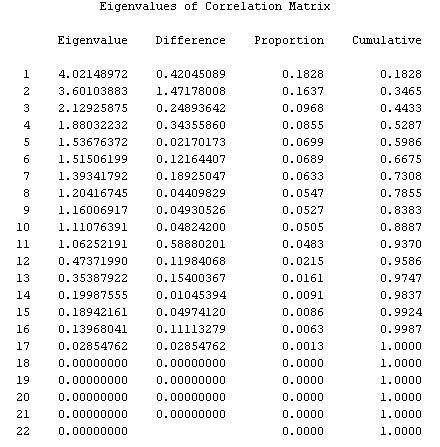
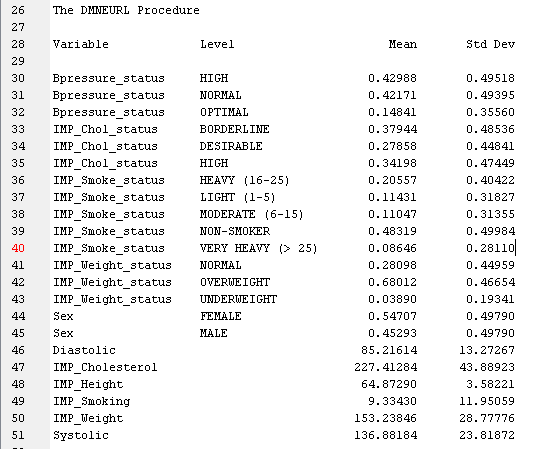


Figure : Principal component

## Data Modelling

## Regression

### Development of models

Impute node with default and tree variations has been taken for the regression model development.

**Steps with Impute Default:**

* Impute Default ->Regression Default and stepwise -> Best Model Regression default.
* Impute Default ->Principal Component ->Regression PC and Default -> Best Model Regression default.
* Impute Default ->Transform Variable->Regression TV and Default -> Best Model Regression default.
* Impute Default ->Transform Variable->Variable selection->Regression TV-VS and Default -> Best Model Regression default.

So final model of impute Default is regression is with default settings.

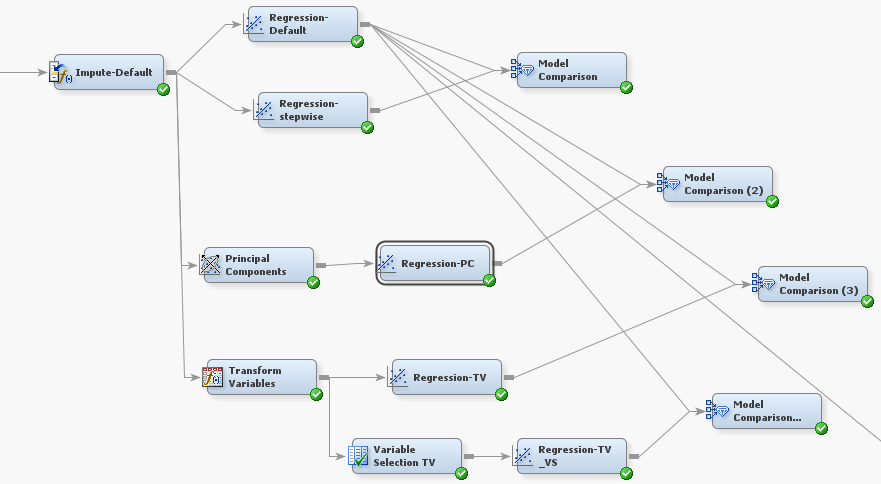


Figure : Pipeline with impute default

**Steps with Impute Tree:**

* Impute Tree->Regression Default and stepwise -> Best Model Regression default.
* Impute Tree ->Principal Component ->Regression PC and Default -> Best Model Regression default.
* Impute Tree ->Transform Variable->Regression TV and Default -> Best Model Regression Transform variable.
* Impute Tree ->Transform Variable->Variable selection->Regression TV-VS and Default -> Best Model Regression Transform variable.

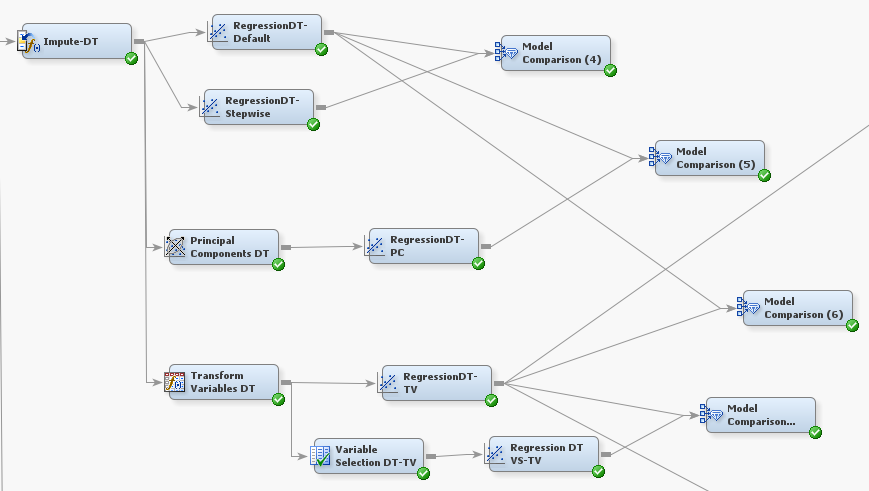
.

Figure : Pipeline with impute tree

From the above two best models, impute with regression default and impute decision tree with transform variable is again compared and got the final best model for regression as impute decision tree with transform variable.

**Best Model selection steps Using ROC curve**

1. Roc curve is analyzed whether any line closer to the curve 0,1 here both are almost in same line so not able to predict from this.
2. In Fit Statistics Regression Default Roc index is having higher value compared to compared to Impute Regression DT-TV but the difference is very small so not able to finalize the prediction model.
3. Cumulative lift also lies in same line so not able to predict model.
4. Finally, the output need to check the False negative where impute Regression DT transform variable shows false negative less compared to regression default. So taken impute Regression DT with transform variable as the best model.

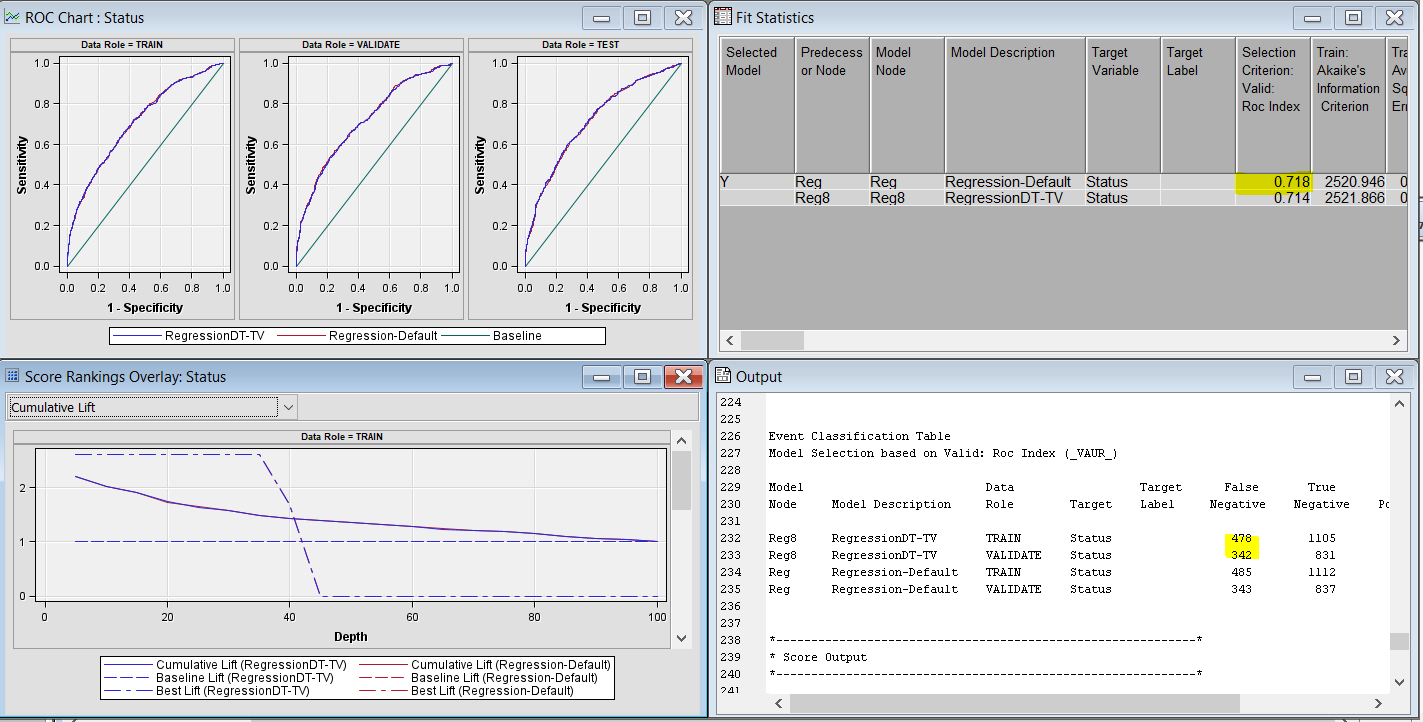


Figure : Roc chart

### Model performance

**Best Model performance using %Response, Cumulative % Response, Lift, and Cumulative Lift.**

1. High non-cumulative % response scope
2. High cumulative Lift
3. Check the scope %
4. Low False negative.

Considering all these factors the Final model for regression shown as Regression with impute tree with transform variable.

**Table 1: Regression models performance**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model variations** | **Non Cu. % Res. Scope** | **Cu. Lift** | **Scope %** | **True –**  **%** | **False –**  **%** | **True +**  **%** | **False +**  **%** |
| Impute Default-Regression default | 56.64% | 1.48 | 40% | 53.62% | 21.97% | 16.21% | 8.19% |
| Impute default-Regression Stepwise | 55.68% | 1.45 | 40% | 54.26% | 22.42% | 15.76% | 7.56% |
| Impute default-Principal component | 54.24% | 1.42 | 40% | 53.11% | 23.45% | 14.73% | 8.71% |
| Impute default-Transform variable | 56.16% | 1.47 | 40% | 53.11% | 21.91% | 16.27% | 8.71% |
| Impute default-transform variable-variable selection | 53.76% | 1.40 | 40% | 54.13% | 24.15% | 14.03% | 7.69% |
| Impute Tree-Regression default | 56.16% | 1.47 | 40% | 53.75% | 22.29% | 15.89% | 8.07% |
| Impute Tree-Regression Stepwise | 56.16% | 1.47 | 40% | 54.32% | 22.42% | 15.76% | 7.49% |
| Impute Tree-Principal component | 54.4% | 1.42 | 40% | 52.98% | 23.25% | 14.93% | 23.25% |
| Impute Tree-Transform variable | 56.48% | 1.48 | 40% | 53.23% | 21.90% | 16.27% | 8.58% |
| Impute default-transform variable-variable selection | 54.35% | 1.42 | 40% | 54.19% | 24.21% | 13.96% | 7.62% |

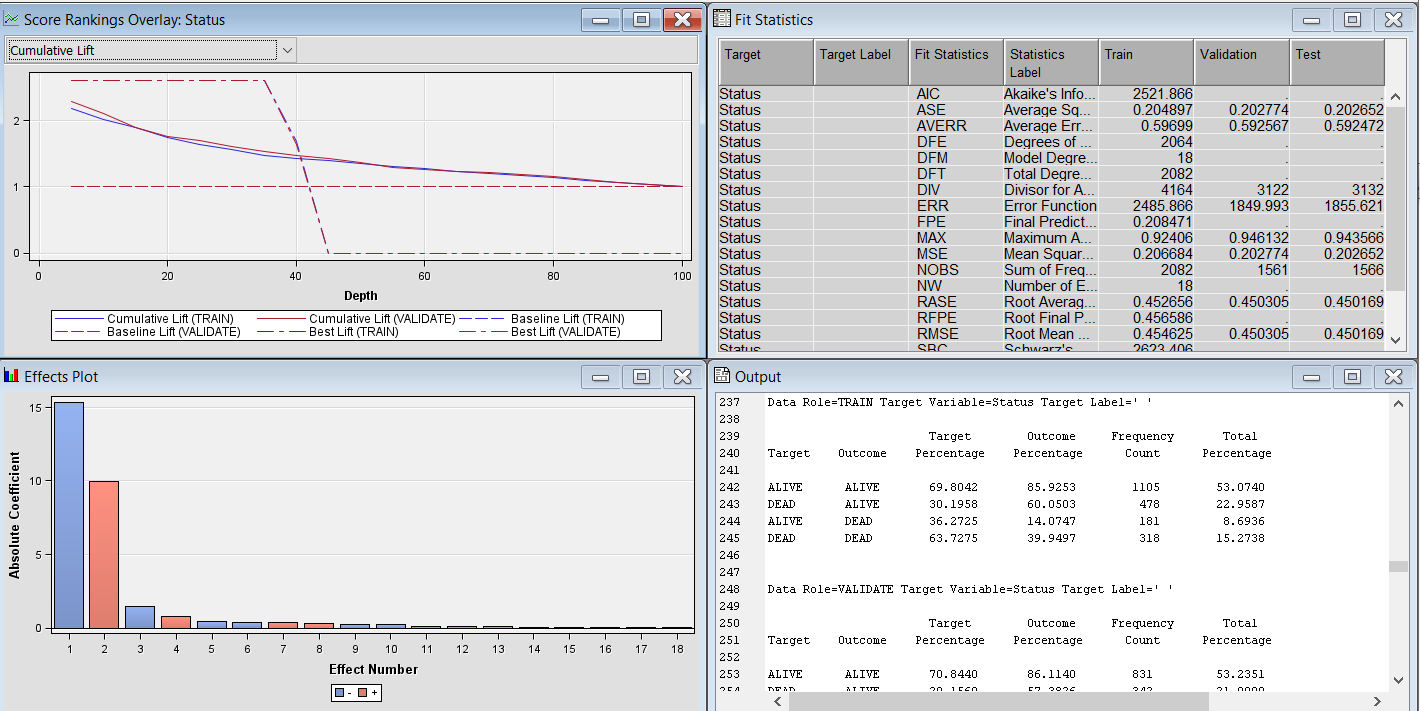


Figure : Model performance Graph

### Chosen Regression equation.

**logit p =a + b1X1 + b2X2 + b3X3 + ………**

Target variable logit p is probability of target event

a is the intercept

b1**,**b2,b3 are Gradient with respect to independent variable X1 ,x2,x3

logit p = -15.3617 + (-0.0337) \* Bpressure\_status\_High + (-0.0414) \* Bpressure\_status\_Normal + (-0.00730) \* Diastolic + 0.0915 \* IMP\_Chol\_status\_Borderline + (-0.2429) \* IMP\_Chol\_status\_Desirable + 0.000888 \* IMP\_Cholesterol + (-0.0987) \* IMP\_Height + 0.3359 \* IMP\_Smoke\_status\_Heavy + 0.1319 \* IMP\_Smoke\_status\_Light + 0.3931 \* IMP\_Smoke\_status\_Moderate + (-1.4868) \* IMP\_Smoke\_status\_Non-smoker + 0.00918 \* IMP\_Weight + (-0.2433) \* IMP\_Weight\_status\_Normal + (-0.4178) \* IMP\_Weight\_status\_Overweight + (-0.4027) \* Sex\_Female + 0.8023 \* TRANS\_0 + 9.9336 \* TRANS\_systolic

## Decision Tree

### Development of models

**Steps with Decision Tree:**

* Impute Tree->2way and 3 way tree -> Best Model 3way Tree.
* Impute Tree->3 way tree and 3 way pruned-> Best Model 3way Tree.
* Impute Tree->3 way average square error, miss classification and lift -> Best Model 3 way average square error.
* Impute Tree ->Principal Component 3 way average square error and 3 way average square error -> Best Model 3 way average square error.
* Impute Tree ->Transform variable 3 way average square error and 3 way average square error -> Best Model Transform variable 3 way average square error.
* Impute Tree ->Transform variable -variable selection 3 way average square error and 3 way average square error -> Best Model Transform variable 3 way average square error.

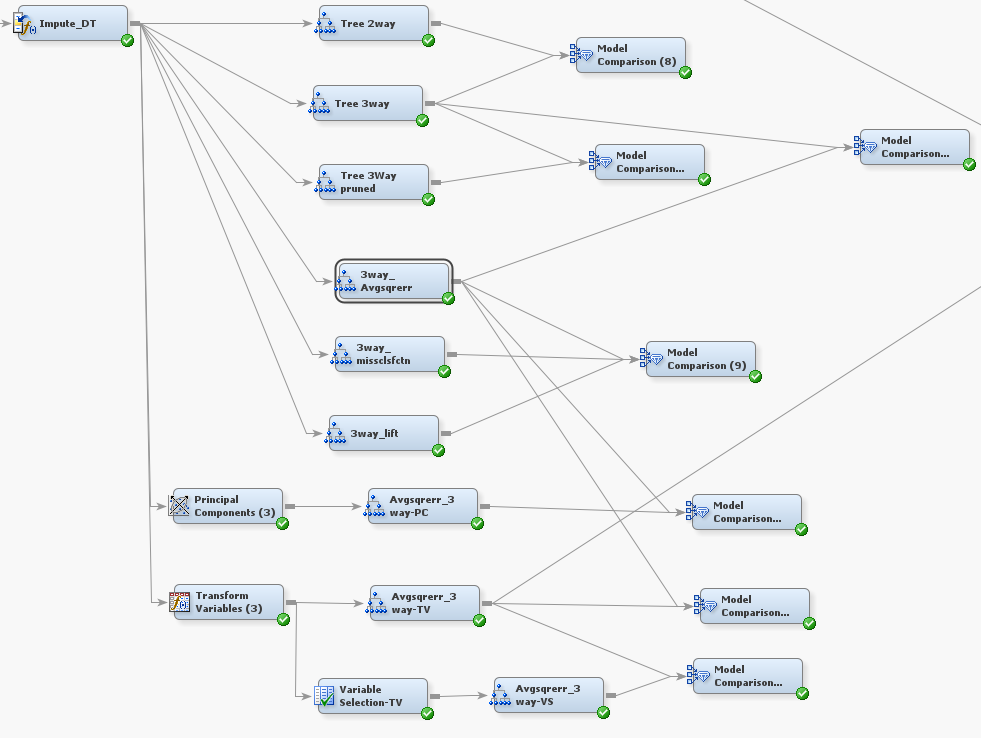


Figure : Impute tree pipeline

**Best Model selection steps Using ROC curve**

1. Roc curve is analyzed whether any line closer to the curve 0,1 here both are almost in same line so not able to predict from this.
2. In Fit Statistics for 3 way Avgsqrerr is having Roc index higher compared to 3 way Avgsqrerr TV but the difference is very small so not able to finalize the prediction model.
3. Cumulative lift also lies in same line so not able to predict model.
4. Finally, the output needs to check the False negative where 3-way Avgsqrerr TV shows false negative less compared to 3 way Avgsqrerr. So taken impute 3-way Avgsqrerr TV with transform variable as the best model.

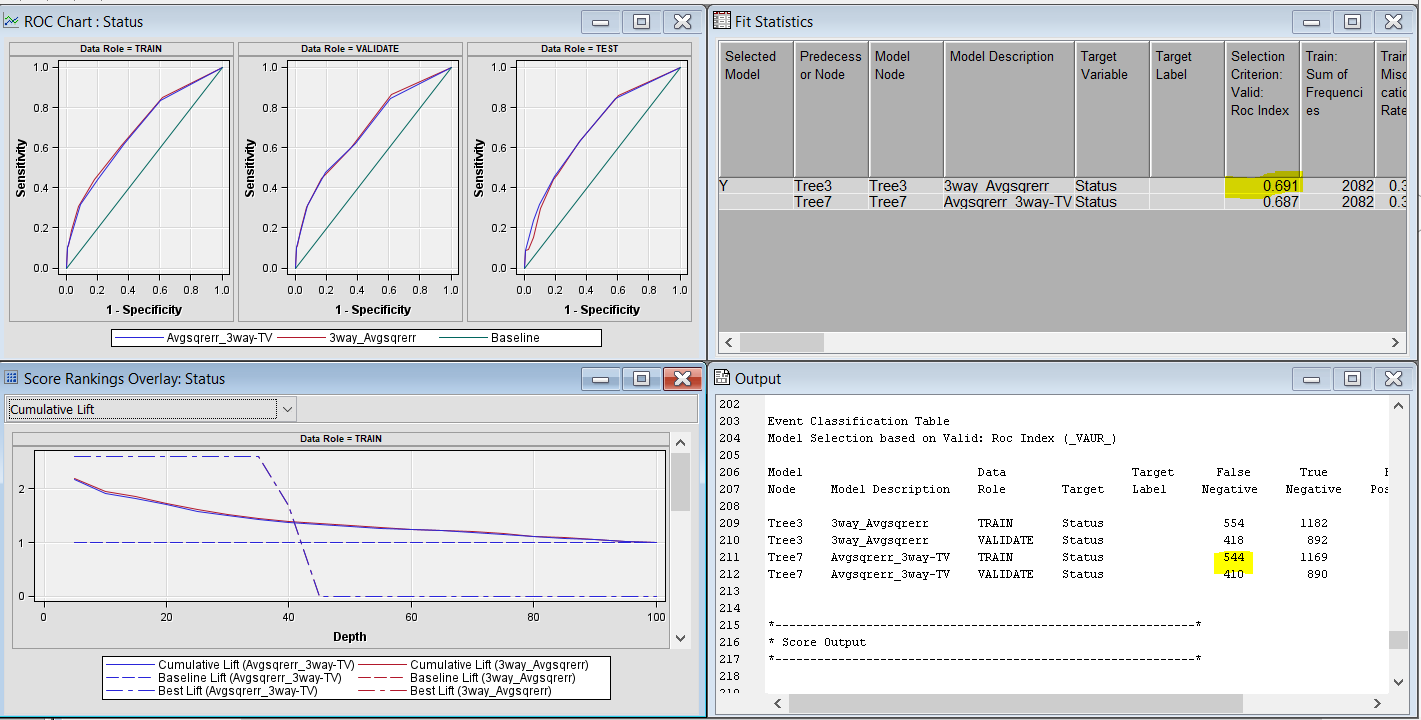


Figure : ROC Curve

### Performance of models

**Best Model performance using %Response, Cumulative % Response, Lift, and Cumulative Lift.**

Final model for Decision tree shown as Impute Tree with Transform variable and Avgsqrerr 3way with high non-cumulative % response scope, high cumulative Lift, scope % and low False negative.

**Table 2: Decision tree models performance**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model variations** | **Non Cu. % Res. Scope** | **Cu. Lift** | **Scope %** | **True –**  **%** | **False –**  **%** | **True +**  **%** | **False +**  **%** |
| Impute Tree-Decision tree 2 way | 48.14% | 1.26 | 40% | 59.83% | 29.47% | 8.71% | 1.98% |
| Impute Tree-Decision tree 3 way | 52.58% | 1.38 | 40% | 57.46% | 26.84% | 11.34% | 4.36% |
| Impute Tree-Decision tree 3 way pruned | 52.24% | 1.37 | 40% | 58.49% | 27.67% | 10.51% | 3.33% |
| Impute Tree-Decision tree 3way Avgsqrerr | 52.82% | 1.38 | 40% | 57.14% | 26.78% | 11.40% | 4.68% |
| Impute Tree-Decision tree 3way missclsfn | 52.58% | 1.38 | 40% | 57.46% | 26.84% | 11.34% | 4.36% |
| Impute Tree-Decision tree 3way lift | 52.24% | 1.37 | 40% | 57.72% | 27.16% | 11.02% | 4.09% |
| Impute Tree – Principal component – Decision tree Avgsqrerr 3way | 50.96% | 1.33 | 40% | 56.24% | 27.87% | 10.31% | 5.57% |
| Impute Tree –Transform variable– Decision tree Avgsqrerr 3way | 53.04% | 1.39 | 40% | 57.01% | 26.26% | 11.91% | 4.80% |
| Impute Tree –Transform variable– Variable selection-Decision tree Avgsqrerr 3way | 51.98% | 1.36 | 40% | 57.14% | 27.67% | 10.51% | 4.68% |

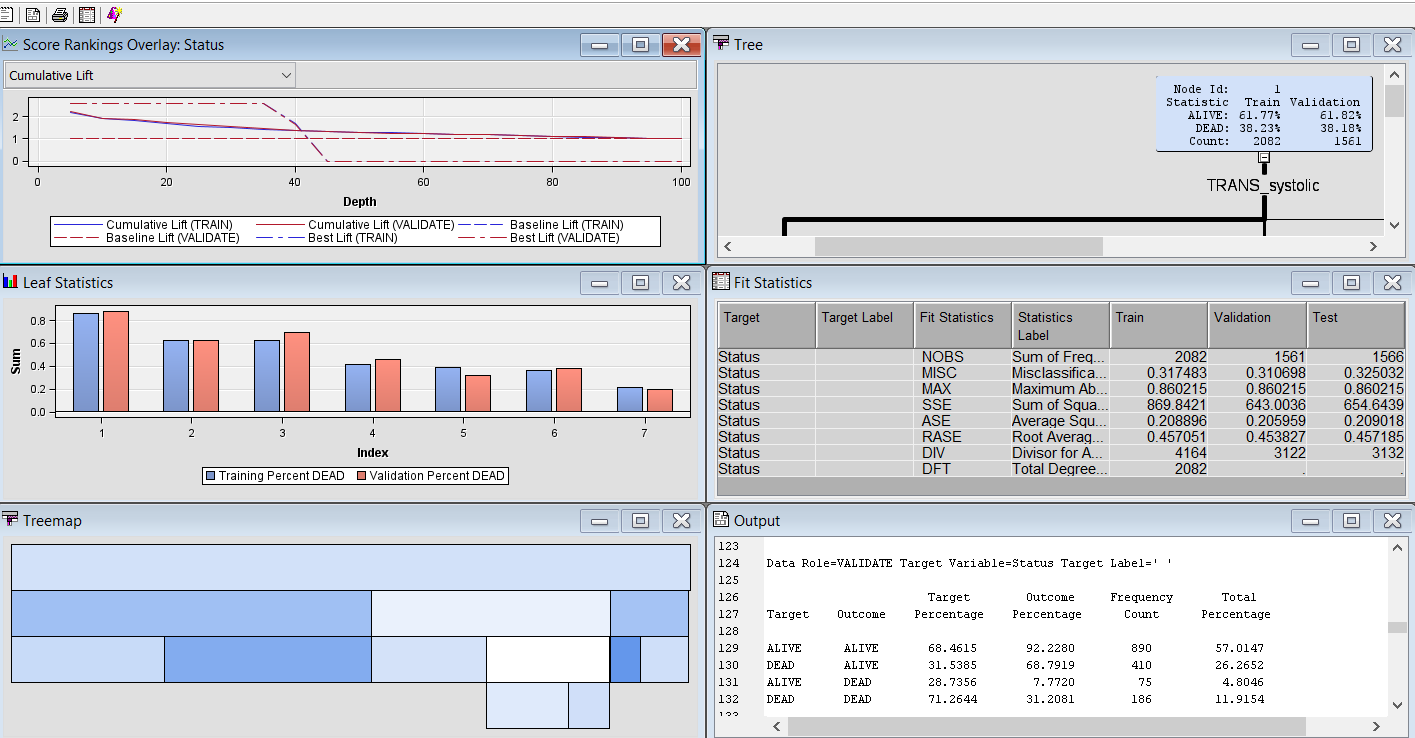


Figure : Decision tree model chart

In tree map top bar shows the root node and each node below shows each levels. Darker colour represents best node.

### Critical path of best model

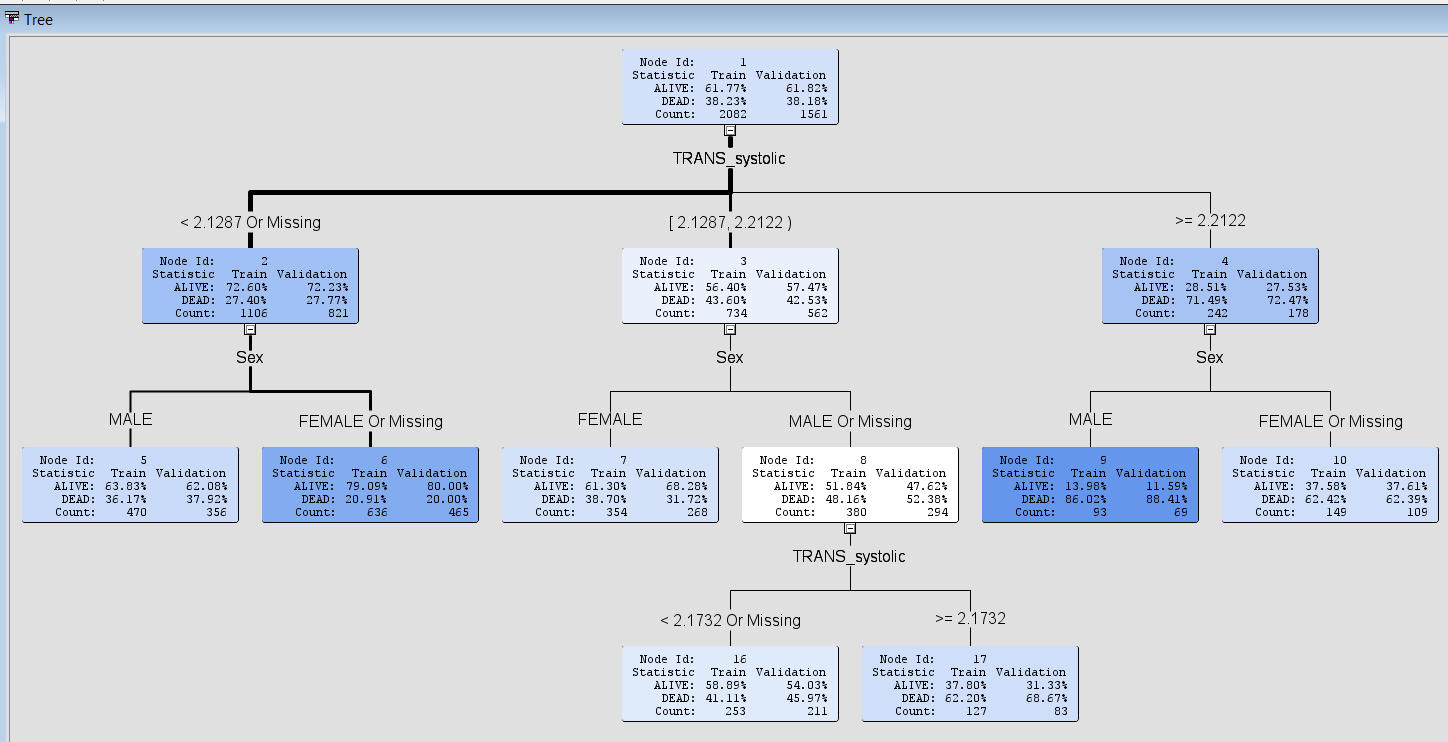


Figure : Decision Tree diagram

TRANS\_systolic <2.1287 or Missing and sex FEMALE Or Missing

### Overfitting analysis

For overfitting analysis subtree assessment plot has been observed with misclassification rate. It is expected both test and validate shows similar behaviour and it starts to diverge if there is no similarity. In this case pruning identified a decision tree with number of leaves as 7.

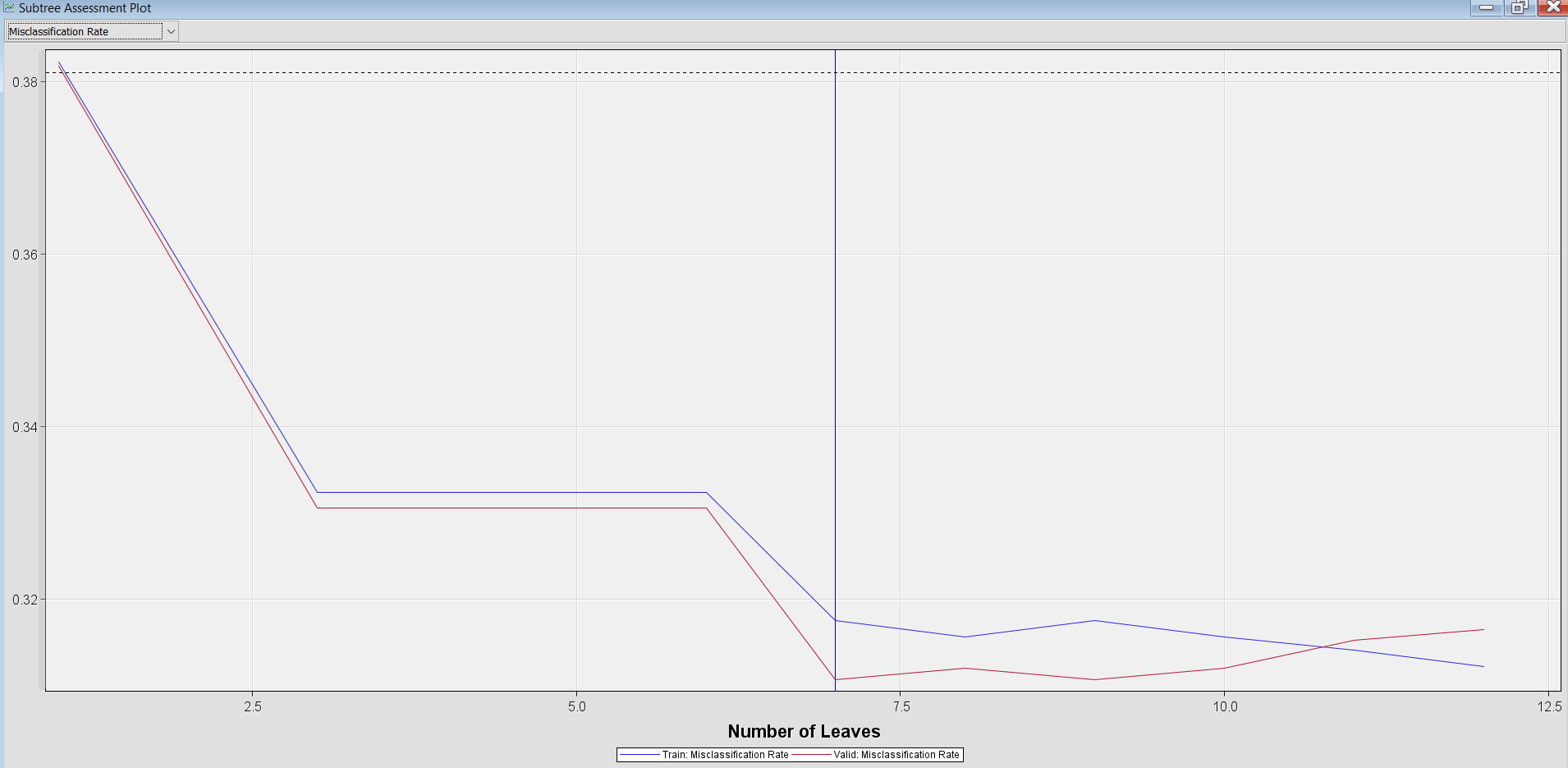


Figure : Subtree assessment plot

# Analysis of the best model

Both the best model from regression and decision tree is compared and got the final best model as the model as impute decision tree with transform variable. Summary table shows regression with impute tree transform variable scored best model with high non-cumulative % response ,cumulative lift and less false negatives compared to decision tree model.

**Table 3 Summary results of the best performing models**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non Cu. % Res Scope** | **Cu. Lift** | **Scope %** | **True –**  **%** | **False –**  **%** | **True +**  **%** | **False +**  **%** |
| Regression(Impute tree- transform variable) | 56.48% | 1.48 | 40% | 53.23% | 21.91% | 16.27% | 8.58% |
| Decision Tree(impute tree-transform variable-3way-Average square error ) | 53.04% | 1.39 | 40% | 57.01% | 26.26% | 11.91% | 4.80% |

**Chart Showing Cumulative Lift and Scope%**

Cumulative lift is the lift and scope is the depth. As per the observation cumulative lift is high forRegression (Impute tree- transform variable) and considered to be the best model. Scope is 40% throughout the model comparisons and the model has 40% likely to default.

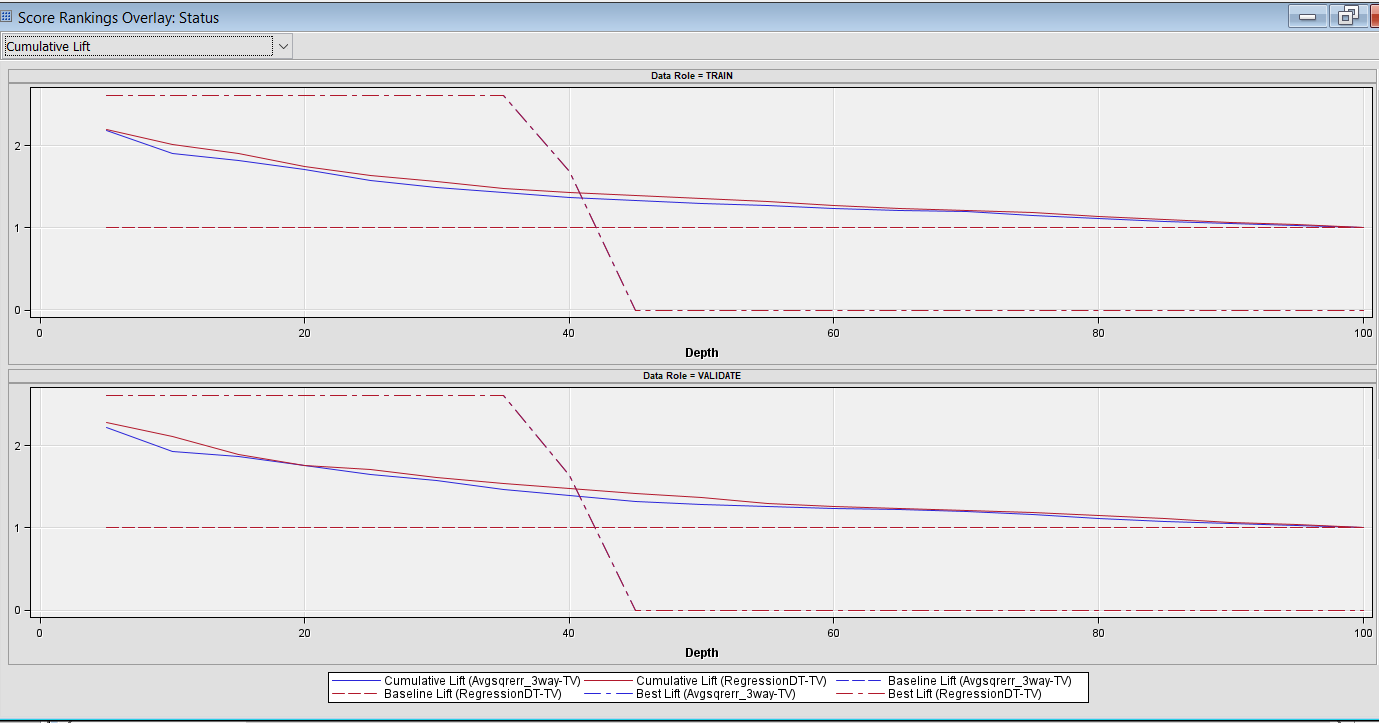


Figure : Cumulative Lift and Scope%

**Chart Showing Non-cumulative % response scope**

No cumulative % response is the cumulative % response and the model with high cumulative response is considered to be the best model. HereRegression (Impute tree- transform variable) has 56.48% of cumulative % response and considered to be the best model.

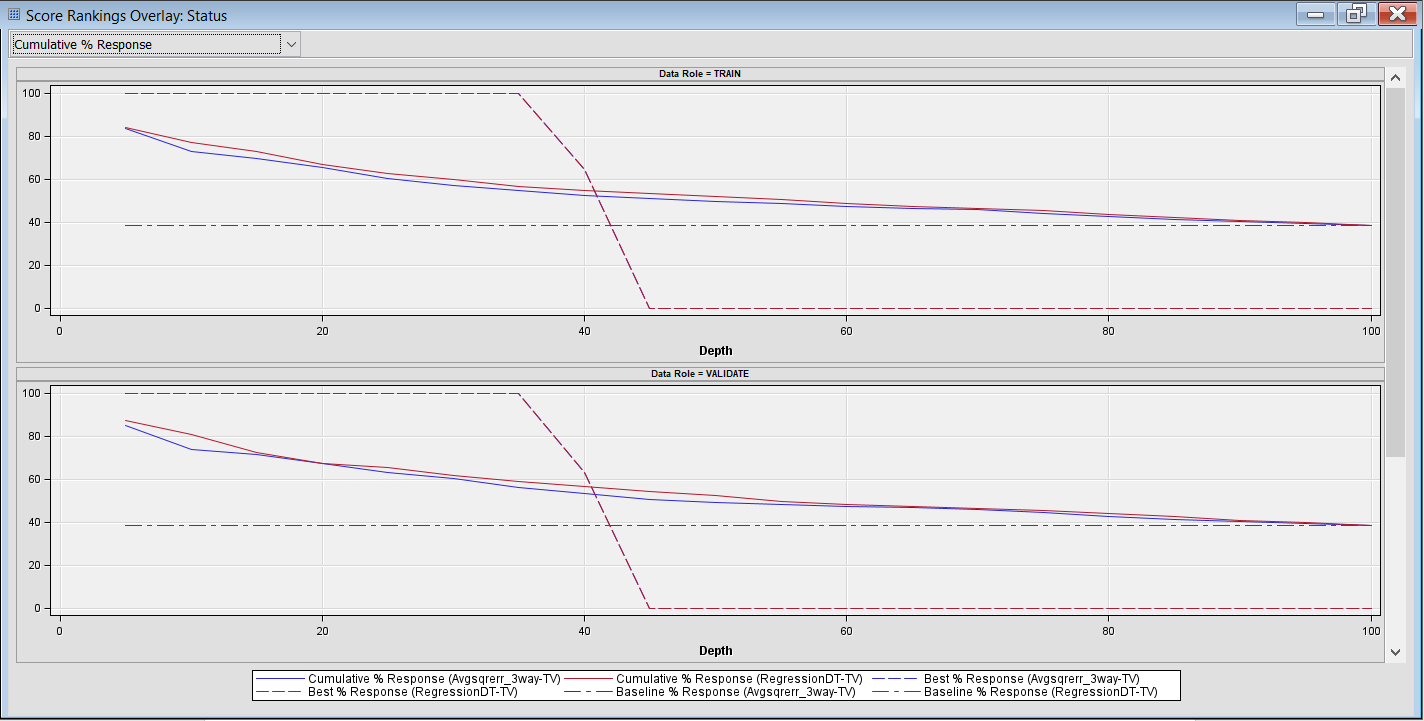


Figure : Non-cumulative % response scope

**Chart Showing False +, False -, True +, True –**

Considering the Target as Alive and Dead these can be distributed as follows:

* Alive Alive: True Negative
* Alive Dead: False Positive
* Dead Dead: True Positive
* Dead Alive: False Negative

Model with less false negative % is considered as best model. In Framingham dataset false negative is less for Regression (Impute tree- transform variable) with 16.27%.

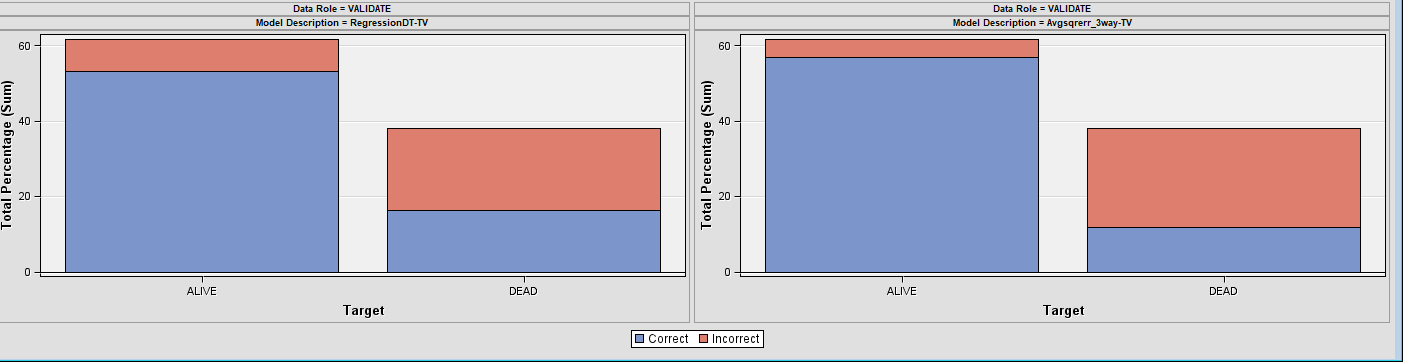


Figure : False +, False -, True +, True –

# A discussion of how the chosen topic from Assessment 1 might be useful for cardiovascular disease research.

The Chosen topic for Assessment 1 was judgmental forecasting. Judgmental Forecasting is used in healthcare to forecast based on human intuition and expert knowledge [4]. It can be useful for cardiovascular disease research in many ways including the expert opinions from doctors or researchers in this field can be taken for prediction based on their previous experiences [2]. It will also help in identifying new emerging risk factors that are not captured in historical data. Scenario analysis can also be conducted to analyze any changes in lifestyle, medical treatment, mortality rate and healthcare allocation [5]. It is also helpful for identifying complex data sets manually with the help of experts and help in bridging gap between research and practice [1]. Moreover, Judgmental forecasting is quick in analyzing the data where if emergency analysis required for cardiovascular disease in health field. Cardiovascular disease may also have some unstructured data such as reports, trends etc. which is difficult to quantify and can be made use of Judgmental forecasting.

Overall, as cardiovascular disease analysis comes under the healthcare sector [3]it is very helpful to get an initial insight for national health service from expert opinions to have the forecasting on the disease to prevent or reduce the impact on individuals and society.

## Conclusion

FRAMINGHAM dataset has been analysed using the various regression and decision tree models with different combinations of variations. As per the analysis, the best model obtained is Regression with dataset being imputed and transformed. Analyses of best model has been carried out by checking ROC Curve, Fit statistics, Correlation matrix as well as the cumulative lift and response %.

Overall, the predicted model has more than 50% of non-cumulative % response and

Model is 40% likely to default. False negative is also less than 25% which makes it suitable model with high precision.

# Recommendation

To reduce the risk of cardiovascular death regular check up needs to be carried by medical departments specially for the individuals having Blood pressure, cholesterol, overweight and smoking habits. Educate the public about the healthy lifestyle especially for females as it shows higher percentage of death rates compared to males. NHS should continue various surveys and model evaluation to have better insights in the future as prevention is better than cure. It is the best option to build a regression model to predict cardiovascular diseases to have a continuous evaluation of the model with better accuracy.

# My Reflections on the process – What did I learn from this exercise?

I got a clear idea on how to predict a best model by SEMMA framework through various steps as follows.

1. How to conduct exploratory data analysis by analysing missing values, outliers.
2. How to perform data partition
3. Data modification with help of impute, transform, variable selection and principal component.
4. Data modelling with regression and decision tree with the above variations
5. Analyzing the best model by using ROC curve, Fit statistics, accuracy factors, cumulative lift and response %.

## References and Bibliography

1. National Heart, Lung, and Blood Institute. The Framingham Heart Study. Available at: https://www.framinghamheartstudy.org/.
2. SAS Institute Inc. 2003.Data Mining Using SAS Enterprise Miner. A case Study Approach, Second Edition. Cary, NC: SAS Institute Inc.
3. Sarma, Kattamuri S.2007.Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications. Cary, NC: SAS Institute Inc.
4. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol. 1979;110(3):281-290.
5. Dawber TR, Meadors GF, Moore Jr. FE. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health. 1951;41(3):279-281.

## Appendix

**Data Mining Roles**

Bpressure status, Chol status, Smoke status and Weight status is given as ordinal datatype as it has more than two levels in an implied order whereas cholesterol, death age, height, diastolic, systolic, smoking and weight as interval as the data is numeric. Binary for sex and status as it has two possible levels.

Status as dead or Alive is taken as target variable to meet the aim of the project against the risk of CVD deaths. Death age is rejected as there are many missing fields in the provided data and the remaining variables as input.

A screenshot of a computer

Description automatically generated with medium confidence

Figure : Import dataset

A screenshot of a computer

Description automatically generated with medium confidence

Figure : statistics chart

**Workflow diagram**

A picture containing line, diagram, text, parallel

Description automatically generated

A picture containing diagram, line, screenshot, plan

Description automatically generated

A picture containing diagram, screenshot, line, text

Description automatically generated