**Project 1**

**Data and Model Analytics using SAS Enterprise Miner**

**MIS 6334 Advanced BA with SAS**

**Group 7**

Ashwin Gajendra

Mayank Kothari

Priyankaa Jayaprakash

Shamsundar Shripad Kulkarni

Vipa Shashikant Patel

Contents

[Project Purpose 3](#_Toc443990359)

[Business Goal 3](#_Toc443990360)

[Part I. Data Pre-processing 3](#_Toc443990361)

[1. Edit Variables 3](#_Toc443990362)

[2. StatExplore 3](#_Toc443990363)

[3. Insights 5](#_Toc443990365)

[4. Replacement 6](#_Toc443990366)

[5. Imputation 7](#_Toc443990367)

[6. Dataset Partition 7](#_Toc443990368)

[Part II. Building Decision Trees 7](#_Toc443990369)

[1. Automatically pruned decision tree 7](#_Toc443990370)

[2. Decision Tree using Log worth (Interactive decision tree) 9](#_Toc443990371)

[Part III. Building Neural Networks and a Regression Model 10](#_Toc443990372)

[1. Transform Variables 10](#_Toc443990373)

[2. Regression Model 10](#_Toc443990374)

[3. Neural Network 11](#_Toc443990375)

[Part IV. Model Comparison and Champion Model Selection 12](#_Toc443990376)

[Part V. Improving Model Performance 13](#_Toc443990378)

[Managerial Implications and Learning 16](#_Toc443990379)

Project Purpose

The Expedia dataset provides detailed information regarding the bookings done by a customer on website. The purpose is to develop a highly-optimized business model along with managerial-relevant findings by walking through SEMMA process in SAS Enterprise Miner which provide us concrete details about various important factors considered while handling a booking session for a customer.

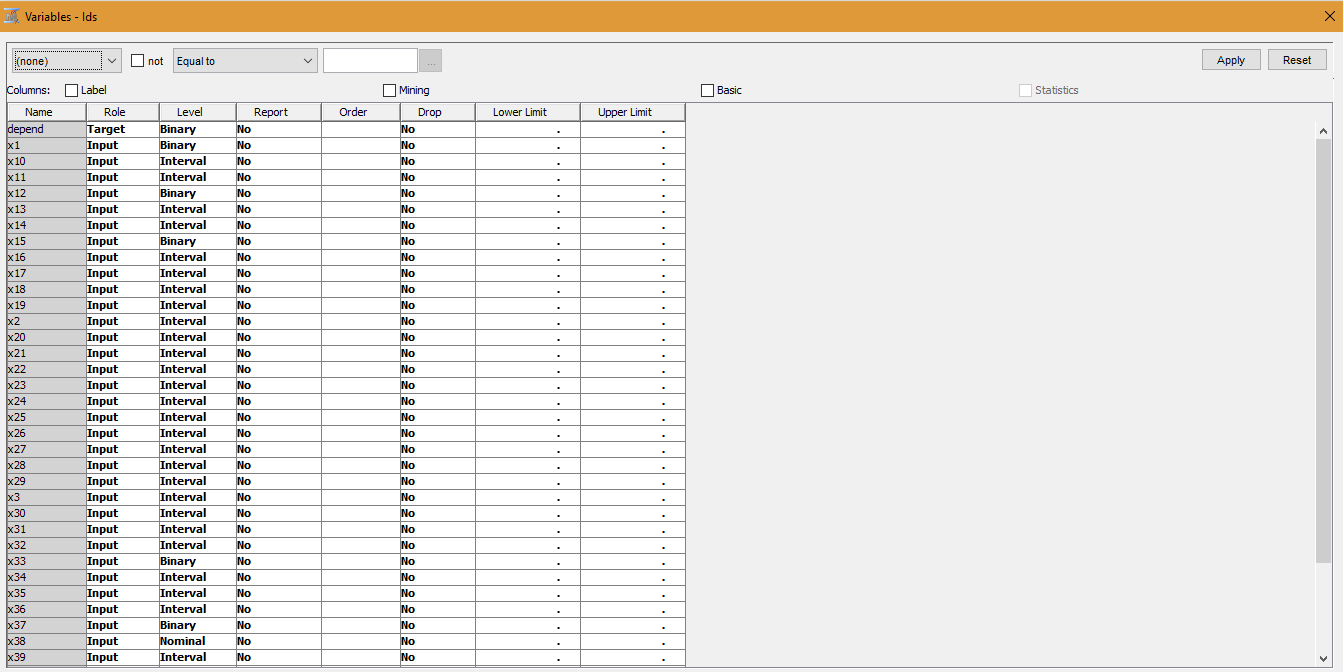
# Business Goal

To predict if the user is going to book in is active session.

# Part I. Data Pre-processing

## Edit Variables

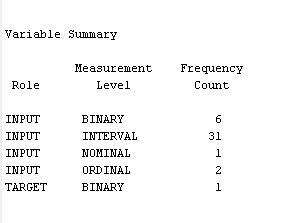
Several variables were classified into wrong data types, so we corrected them as follows:



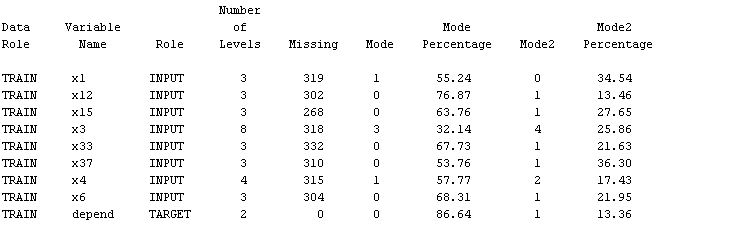
## StatExplore

## To explore the given dataset further, we used StatExplore node

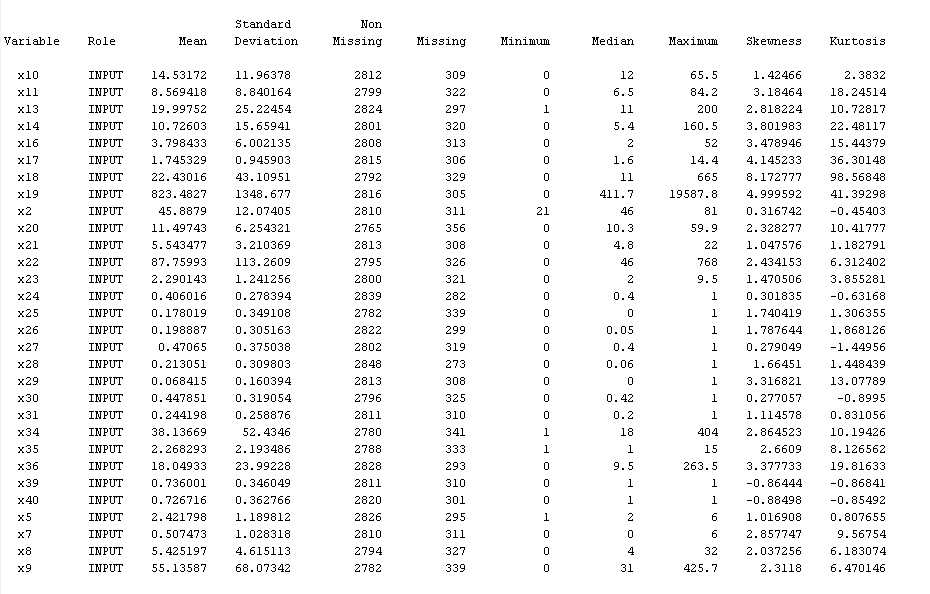
* We decided to reject X32 and X38 since the missing value is 50% of the total proportion.
* Rest all variables have the missing value percentage of about 10%.



**Class Variable Summary Statistics**



**Interval Variable Summary Statistics**



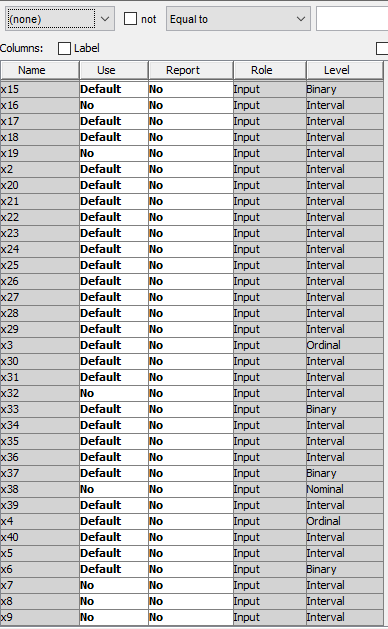
## Insights

While analysing the dataset, we found interesting association among few variables.

* X25 (Percentage of total bookings are to this site) can be derived from X7(No. of bookings the user made at this site in the past) and X16 (No. of past bookings of all sites so far) such that X7/X16 = X25. Likewise,
* X11(Average time spent per sessions to this site) = X9 (Time spent in this site so far in minutes) / X8 (No. of sessions to this site so far)
* X28 (Percentage of total minutes are to this site) = X9 (Time spent in this site so far in minutes) / X19 (Total minutes of all sites)

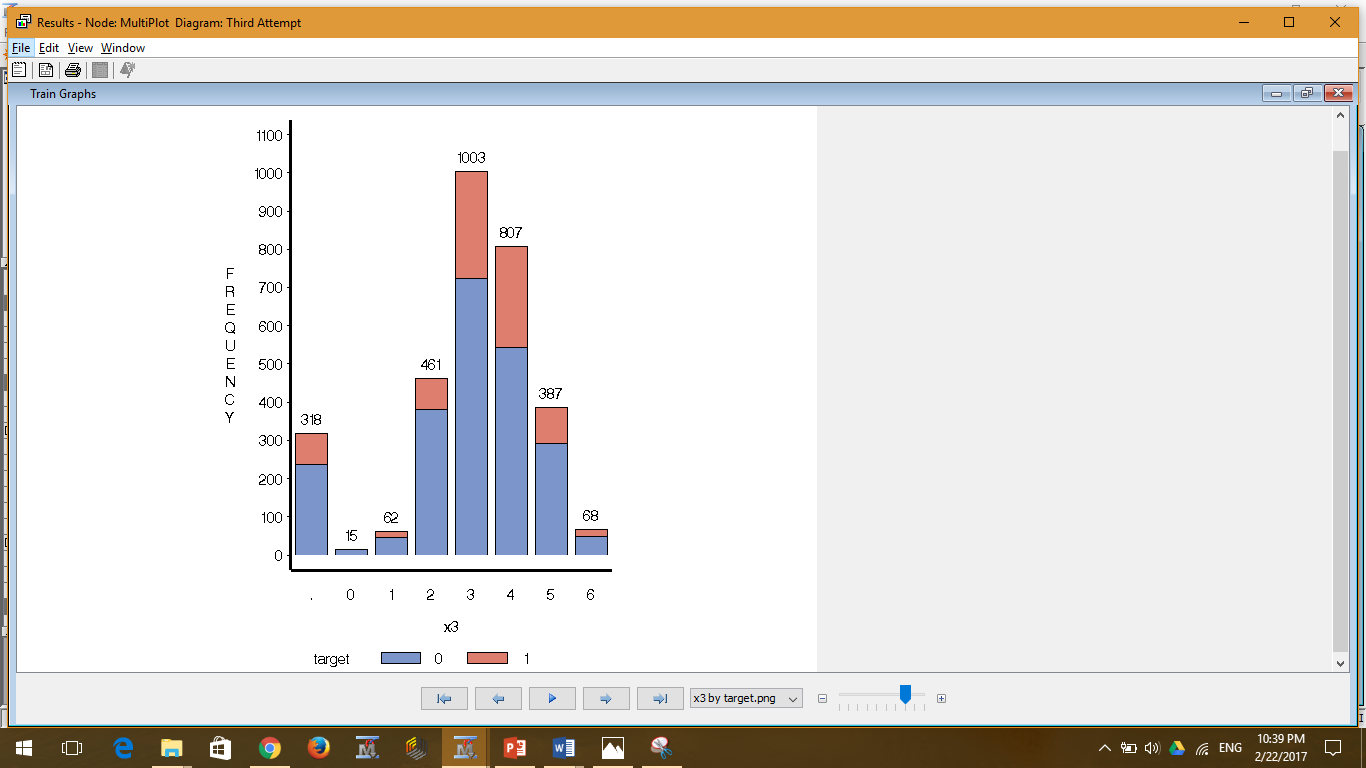
Hence, X9, X8, X19, X7, and X16 can be represented/denoted by X11, X28, and X25.

Also, the variables X7, X8 and X9 seem to be obvious, there is no surprising information which these variables convey. Therefore, we decided to drop these variables to reduce the redundancy.

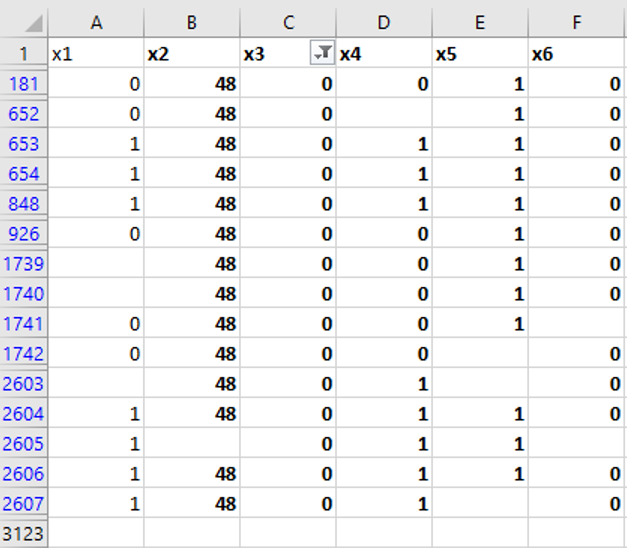


## Replacement

* We initially thought of replacing value for variable whose levels are undefined or misrepresented by \_MISSING\_ and later imputing them, but we observed that all the levels are represented properly.
* There are 15 observations for which income(X3) is zero.



* We analysed Age, Household Size, Child, Education variables for these observations and found that Income=0 seemed like an error or the users didn’t reveal their true income. So, we replaced it by missing and then imputed the value



## Imputation

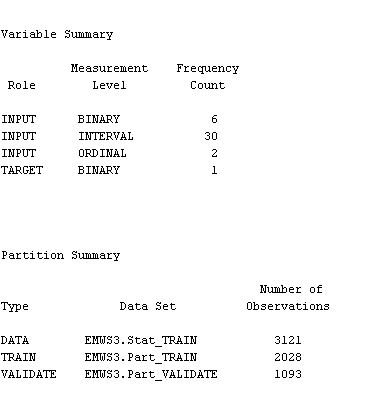
We have used impute node to impute the missing values.

* For Class Variables, we decided to impute data with Tree surrogate imputation method
* For Interval variables, we imputed data with Median

## Dataset Partition

Training: 65% of the data was allocated for training

Validation: 35% of the data was allocated for validation



# Part II. Building Decision Trees

## Automatically pruned decision tree

**Pruning tree:**

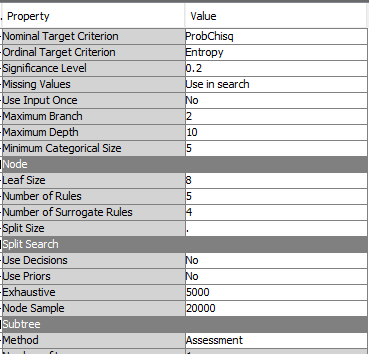
* Maximum Depth: 10
* Leaf Size: 8
* Number of Surrogate rules: 4

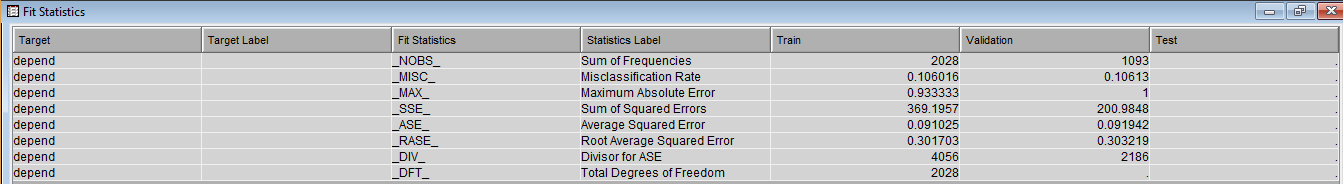
**Results:**

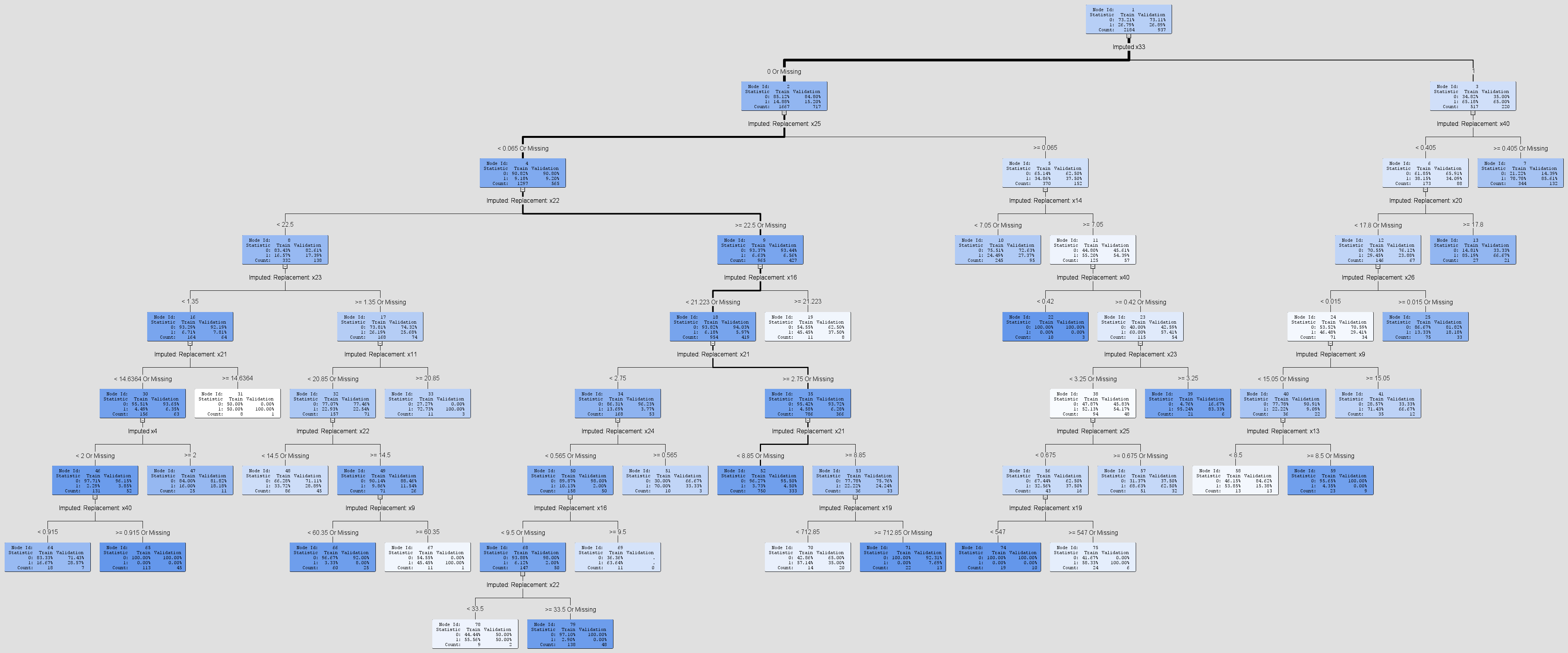
*Misclassification Rate:*

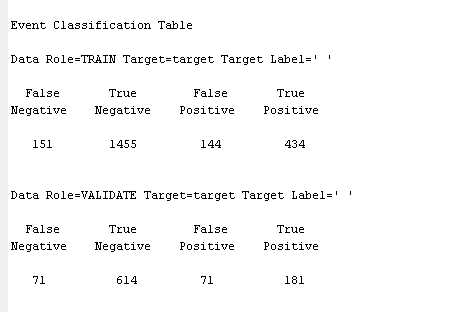
***Train: 0.106016***

***Validation: 0.10613***









## Decision Tree using Log worth (Interactive decision tree)

**Interactive decision tree:**

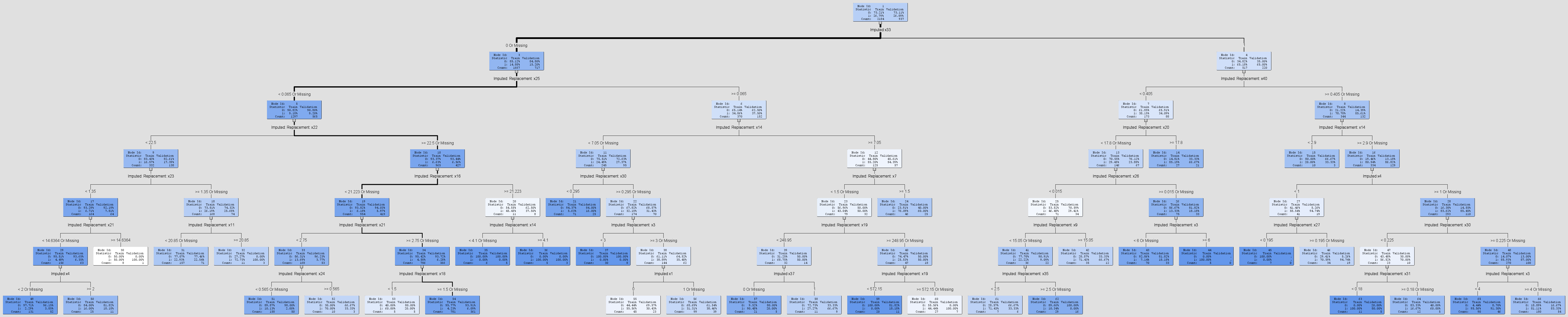
On the basis of log worth decision values pruned tree explicitly and ran the model with same other pruning parameters as the above decision tree model.

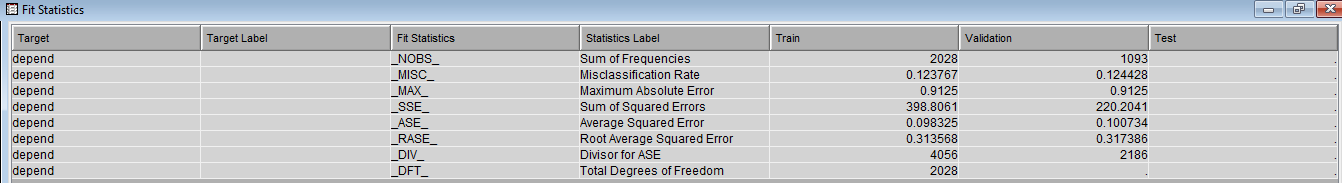
**Results:**

*Misclassification Rate:*

***Train: 0.123767***

***Validation: 0.124428***

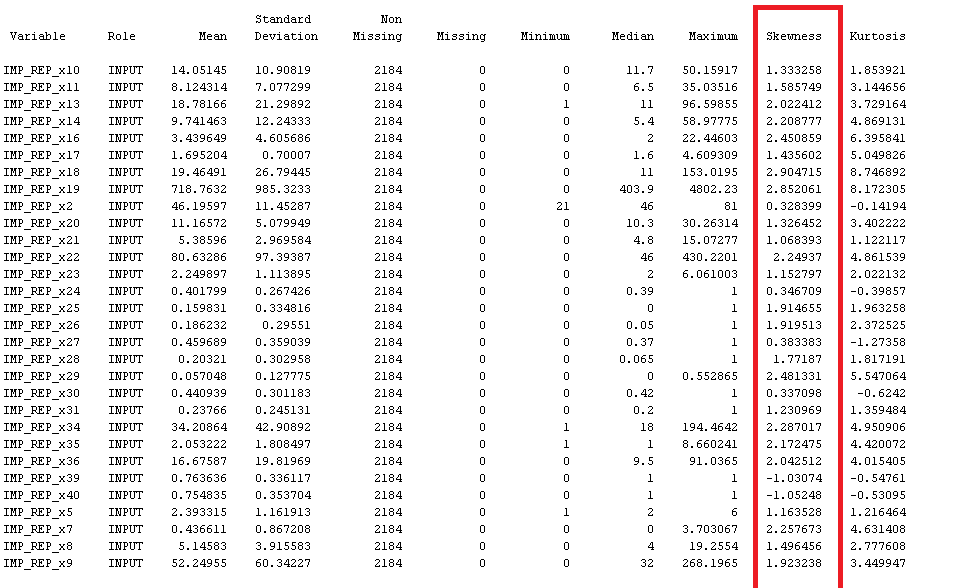




# Part III. Building Neural Networks and a Regression Model

## Transform Variables

Next step in data pre-processing is transforming the input variables to reduce the skewness in data. We have used transform variable node to reduce the standard deviation and skewness of the input variables. We used the log10 transformation on the positively skewed interval variables and dummy indicators for class variables.



## Regression Model

Since our target variable is binary, we have used logistic regression.

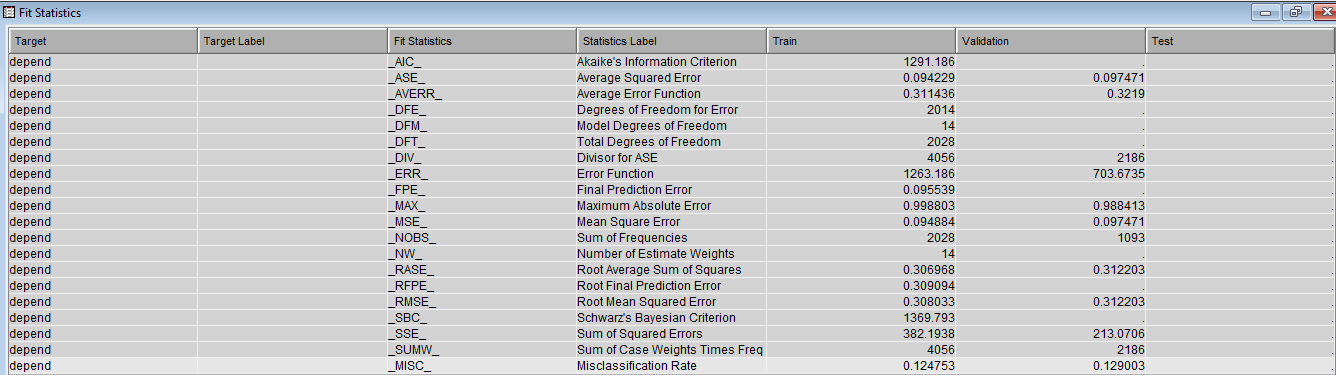
Model Selection: Stepwise

**Results:**

*Misclassification Rate:*

***Train: 0.126531***

***Validation: 0.125178***



## Neural Network

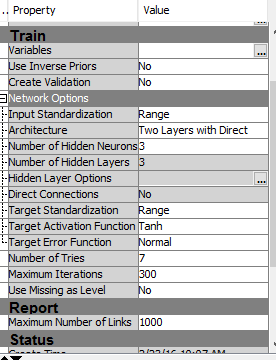
Developed HP Neural Network with the following configurations

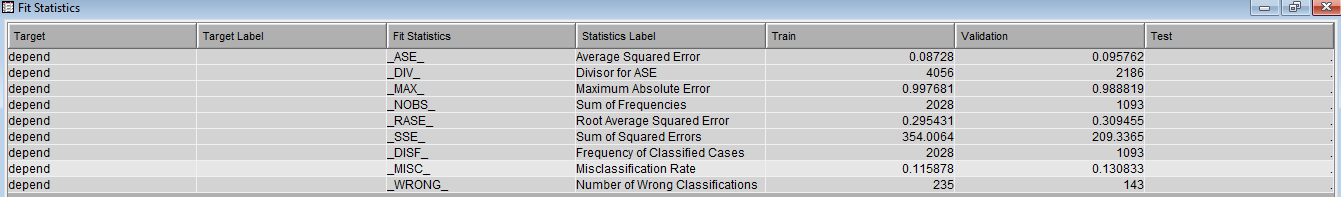
**Results:**

*Misclassification Rate:*

***Train: 0.115878***

***Validation: 0.130833***

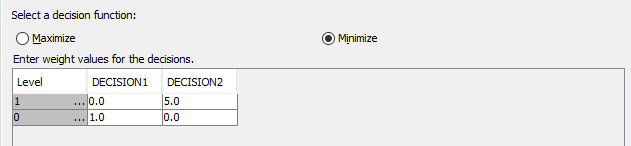




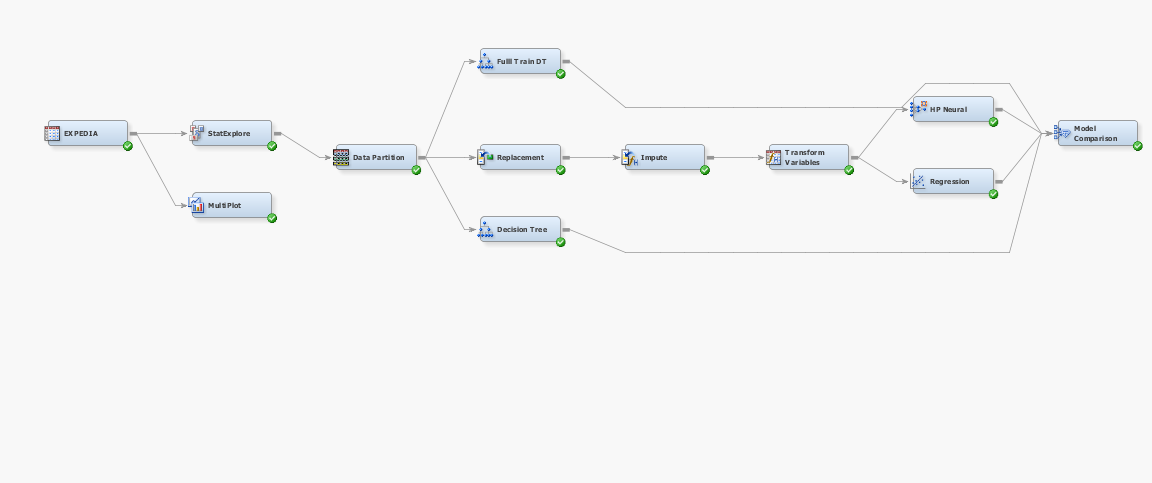


# Part IV. Model Comparison and Champion Model Selection

Using a cost of 5 for misclassifying 1 as 0 and a cost of 1 for misclassifying 0 as 1.Decision Function and Weights:



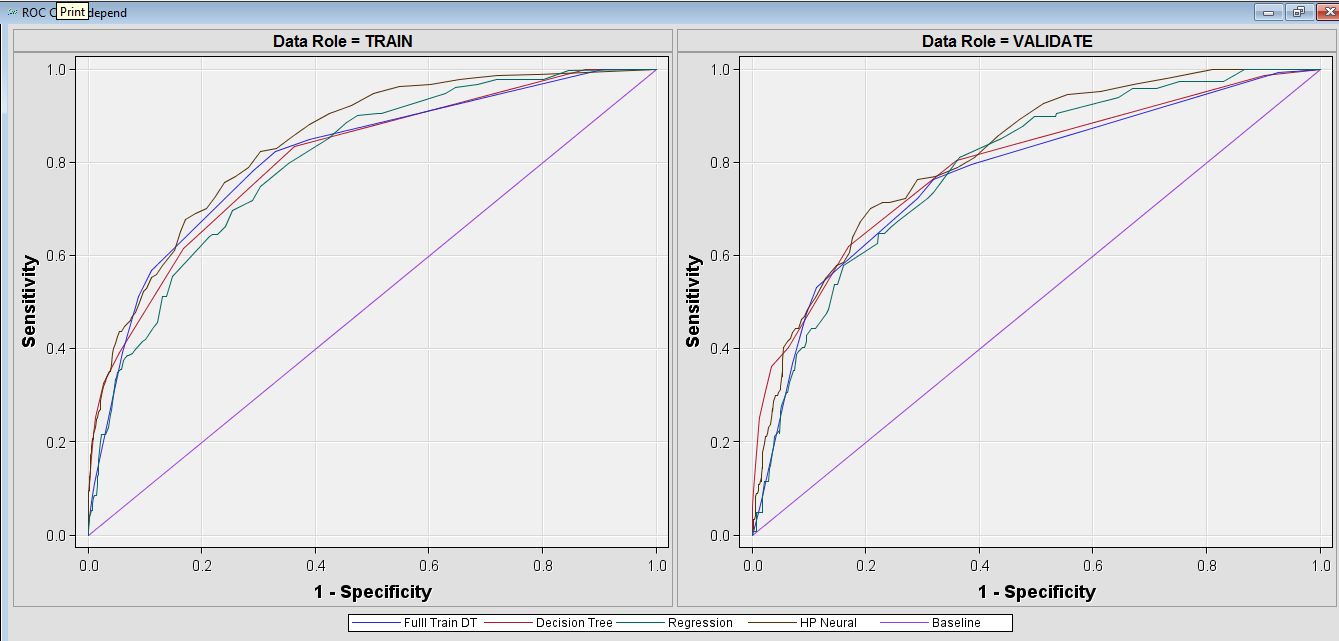
**Models:**



**Fit Statistics (Model Comparison):**

# 

**ROC Curve:**

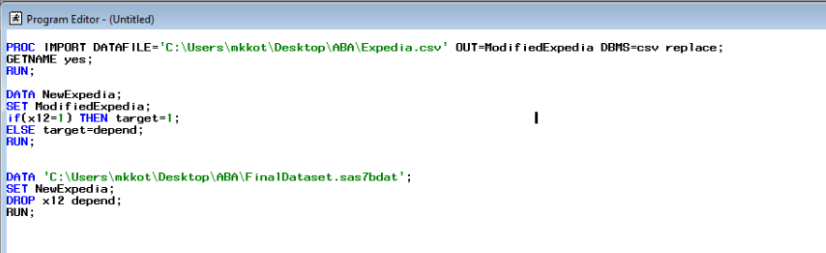


**Champion Model:** Decision Tree

# Part V. Improving Model Performance

Leveraging X12 variable significance

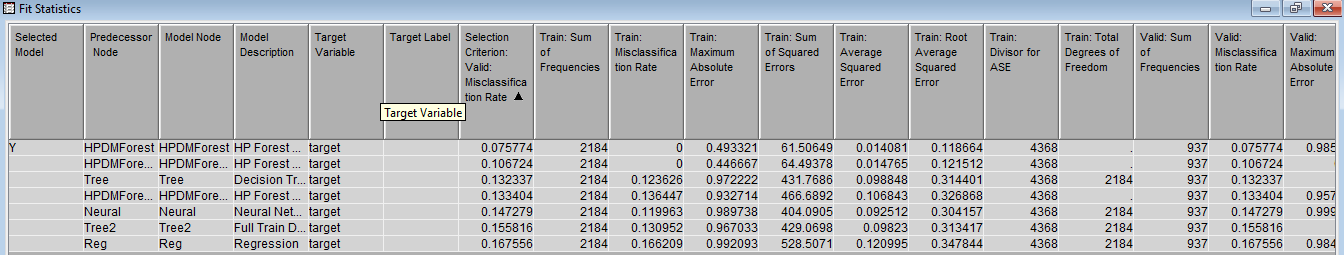
To optimize the model performance, we can use X12 to combine with depend and come with a new variable target that would represent the combined result as whether the user is going to book in the session or not.



Random Forest

Decision Tree being the champion model, we decided to use ensemble method i.e. Random Forest with decision trees using inbuilt bootstrap aggregation and feature selection. We have used various variations of HP Random Forest i.e. HP Random Forest with Transform Variables, HP Forest with PCA, HP Forest with Variable Selection.

* HP Forest with Transformation comes out to be the best model.



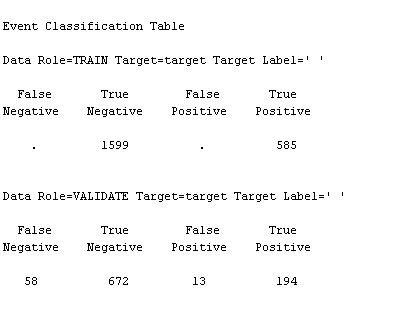
**Results:**

*Misclassification Rate:*

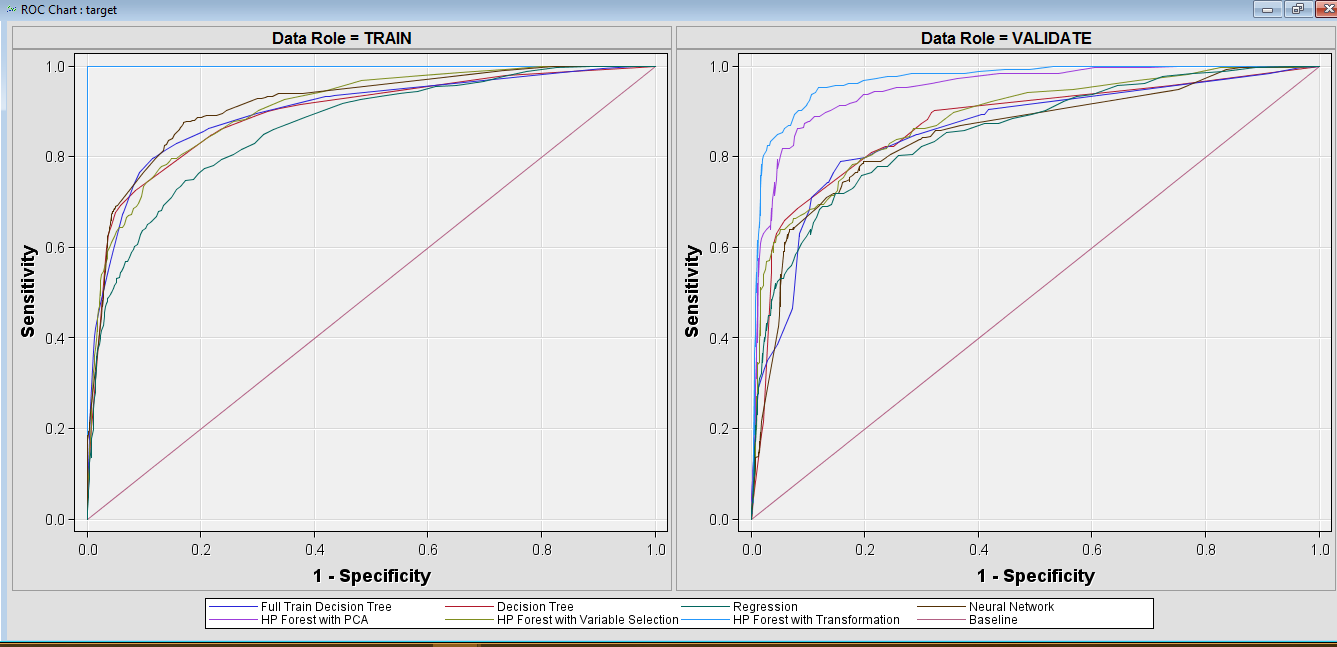
***Train: 0***

***Validation: 0.075774***

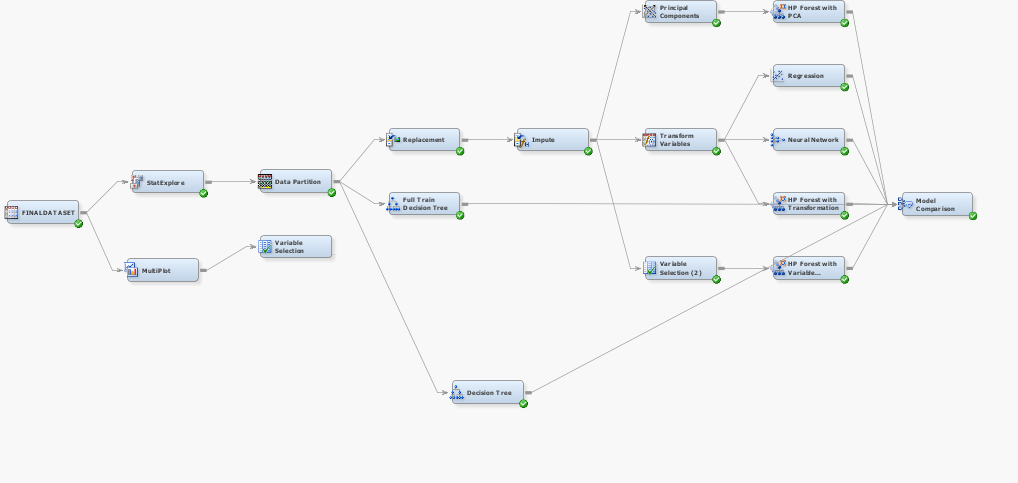
**Confusion Matrix:**



**ROC Curve:**



**Model:**



# Managerial Implications and Learning

# Managerial Implications

# In course of the analysis and findings we can recommend the below managerial insights:

1. Targeting potential customers based on their online behaviour

* Customers browsing data (Cookie Information) and his past activity on other websites
* Designing tactics based on the user information that is received
* We can leverage the following variables to support our insights, X33(if the user has booked at any site up to this point) and X25(% of total bookings are to this site).

1. Using the variable which signifies if the user has spent substantial time on our website and hasn’t proceeded with the booking, we can make use of this information and use it in remarketing, that is if the user is trying to book the ticket on competitor’s website, we can make provision to flash our tickets price (with a discounted amount) and try to navigate the customer back to our website.
2. Also, going one step further, we can consider Age and Gender factors combined with the time the user spends on the site, if we have predicted the user is a potential customer we can come up **Demographic** **targeting**. In Demographic targeting we can segregate people based on variable importance such as age and gender and target them with likely discounts and coupons and make sure they book with our site.

Learnings

1. Stat Explorer Node helped to look for missing class and interval values, skewness of data, outliers and inconsistent data
2. To deal with noisy data, it is recommended to use Replacement technique to keep the data limits in 3 sigma limit.
3. Tree Surrogate is a good technique to deal with Class variables when doing Data imputation.
4. Full decision trees are generated in Random Forest, so it is always better to reduce the data skewness and use the variable selection node.
5. By changing the maximum number of trees in Random Forest node configuration helps us to overcome overfitting.