1. **Executive Summary**

Fatality Analysis Reporting System (FARS) was created in the United States by the National Highway Traffic Safety Administration (NHTSA) to provide an overall measure of highway safety, to help suggest solutions, and to help provide an objective basis to evaluate the effectiveness of motor vehicle safety standards and highway safety programs.

It collects data for analysis of traffic safety crashes to identify problems, and evaluate countermeasures leading to reducing injuries and property damage resulting from motor vehicle crashes. We have chosen this dataset to analyze and provide key insights and general trends on leading causes of a traffic accident fatality.

We can provide recommendations and help officials make an informed decision on new rules and regulations for traffic fatalities.

This dataset can be used to model various variables and factors such as weather conditions, type of road, light conditions, time of accident, drunk driving, speeding and other many other causes against the number of deaths.

With this kind of information in mind, we can easily visualize the trends and causes of a traffic fatality.

1. **Project Background**
   1. **Project Objective-**

Road safety is one of the major subjects within the transport policy of the United States of America. With every passing year, there has been a certain surge in the number of traffic fatalities due to various factors, despite the advancements in the technologies and stringent rules and regulations to prevent accidents. The primary objective of our project is to determine the critical factors that lead to automobile fatalities across different states in the US and also study the effect of speeding and drunk driving and determine which one authorities should focus on.

* 1. **Project Assumption-**

Traffic Fatalities, Vehicle speeding, and alcohol addiction are commonly viewed in the United States as problems that arise out of human behaviors; However, police and practitioners have learned that the circumstances in which people are caught in accidents heavily affects the mortality rate due to the traffic accidents. Influences on death include: Vehicle Speed, Whether the driver was drunk or not.

We also assume that the influences on death due to the above mentioned reasons remain almost the same throughout the year. We also assume that the weather conditions are remaining same throughout the same.

1. Data Source:

We used the second-hand data set “2015 Traffic Fatalities” from <https://www.kaggle.com/nhtsa/2015-traffic-fatalities> . This data set comprises of multiple excel files and we have utilized 5 excel files for our analysis as per the problem definition – vindecode, vehicle, accident, person and vision which contain 48923 rows approximately including the header line.

**Dataset Description-**

The data set used in this analysis is a second-hand dataset obtained from Kaggle. These data of 2015 traffic fatalities from the U.S. government can be used to predict fatalities across different states in the US, based on various predictors.

The data set consist a total of 50056 records with 37 parameters. Below is the description of the parameters included in the data set.

**Data description for factors affecting Death by automobile:**

|  |  |  |
| --- | --- | --- |
| **Sl No.** | **Predictors** | **Data Description** |
| 1 | STATE | This data element identifies the state in which the crash occurred |
| 2 | ST\_CASE | This data element is the unique case number assigned to each crash |
| 3 | VEH\_NO | This data element is the consecutive number assigned to each vehicle in the case |
| 4 | NCICMAKE | The vehicle make generally contains what the general public usually considers to be a vehicle brand name |
| 5 | VINYEAR | The marketing year defined by the OEM within which the vehicle was produced |
| 6 | VEHTYPE | A Polk assigned code that defines the type of a vehicle represented by a specific VIN |
| 7 | BODYSTYL | A Polk assigned code that describes the body style of the vehicle |
| 8 | DOORS | The number of doors the vehicle has |
| 9 | DISPLCI | (Displacement CID) displacement in cubic inches |
| 10 | CYLNDRS | Contains a code that represents the number of cylinders a vehicle's combustion engine can have. |
| 11 | FUEL | What an internal combustion burns to move a piston in a cylinder |
| 12 | WHLBSH | Contains the distance between the front and rear axles of a vehicle in inches of the base model of the vehicle. |
| 13 | WHLBLG | Contains the longest distance between the front and rear axles of a vehicle in inches for a particular series of that vehicle. |
| 14 | TIRESZ\_F | Describes the size of the front tire. |
| 15 | SHIPWEIGHT | Contains the base weight of the vehicle - The base weight of a vehicle is the empty weight of the base model of the vehicle |
| 16 | MSRP | Contains the base price of the vehicle - BASE PRICE includes only the price for the base model of the vehicle |
| 17 | DRIVETYP | This element describes type of driving configuration for cars and trucks |
| 18 | ABS | A code that describes whether a vehicle has or does not have anti-lock brakes, and what kind of brakes they are |
| 19 | PLNTCNTRY | A code representing the country the plant is in. |
| 20 | DAY | data element records the day of the month on which the crash occurred |
| 21 | MONTH | data element records the month in which the crash occurred |
| 22 | YEAR | data element records the year in which the crash occurred |
| 23 | HOUR | data element records the hour at which the crash occurred |
| 24 | HIT\_RUN | data element identifies whether this vehicle was a contact vehicle in the crash that did not stop to render aid |
| 25 | DEATHS | data element records the number of fatalities that occurred in this vehicle |
| 26 | DR\_DRINK | data element records whether the driver was drinking |
| 27 | AGE | data element identifies this person’s age at the time of the crash, in years, with respect to their last birthday |
| 28 | SEX | data element identifies the sex of this person involved in the crash. |
| 29 | PBPTYPE | data element describes the role of this person involved in the crash |
| 30 | ROUTE | data element identifies the route signing of the traffic way on which the crash occurred |
| 31 | LGT\_COND | data element records the type/level of light that existed at the time of the crash as indicated in the case material |
| 32 | WEATHER1 | data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material - coded data element |
| 33 | WEATHER2 | data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material - coded data element |
| 34 | WEATHER | data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material - derived from WEATHER1 and WEATHER2 |
| 35 | FATALS | data element records the number of fatally injured persons in the crash |
| 36 | DRUNK\_DR | data element records the number of drunk drivers involved in the crash |
| 37 | DEATH\_Y\_N | Driver Survived or Died |

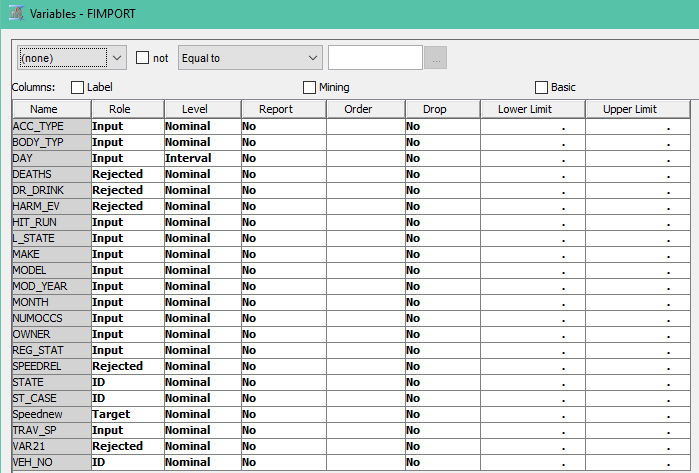
**Data Description for Speeding**

Dependent Variable

|  |  |
| --- | --- |
| **DAY\_WEEK** | This data element records the day of the week on which the crash occurred |
| **FUNC\_SYS** | 01 (Interstate), 02 (Principal Arterial – Other Freeways and Expressways), 03 (Principal Arterial – Other), 04 (Minor Arterial), 05 (Major Collector), 06 (Minor Collector), 07 (Local), 96 (Trafficway Not in State Inventory), 98 (Not Reported), and 99 (Unknown |
| **LGT\_COND** | This data element records the type/level of light that existed at the time of the crash as indicated in the case material |
| **MONTH** | This data element records the month in which the crash occurred. |
| **ROUTE** | This data element identifies the route signing of the trafficway on which the crash occurred |
| **RUR\_URB** | 1 (Rural), 2 (Urban), 6 (Trafficway Not in State Inventory), 8 (Not Reported) and 9 (Unknown). |
| **WEATHER** | This data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material |
| **rest\_use** | This variable shows whether the driver had any type of restraint on the body (1) restraint was present (2) restraint was absent |
| **COUNTY** | This data element records the location of the unstabilized event with regard to the County. The codes are from the General Services Administration’s (GSA) publication of worldwide Geographic Location Codes |
| **DAY** | This data element records the day of the month on which the crash occurred |

Target Variable

|  |  |
| --- | --- |
| **DRUNK\_DR\_new** | It states if the the driver was drunk (1) or not drunk (0) |



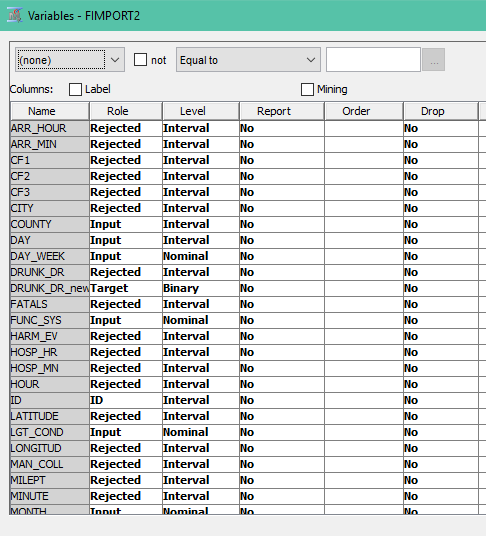
**Data Description for drunk driving:**

Dependant variables

|  |  |
| --- | --- |
| **DAY\_WEEK** | This data element records the day of the week on which the crash occurred |
| **FUNC\_SYS** | 01 (Interstate), 02 (Principal Arterial – Other Freeways and Expressways), 03 (Principal Arterial – Other), 04 (Minor Arterial), 05 (Major Collector), 06 (Minor Collector), 07 (Local), 96 (Trafficway Not in State Inventory), 98 (Not Reported), and 99 (Unknown |
| **LGT\_COND** | This data element records the type/level of light that existed at the time of the crash as indicated in the case material |
| **MONTH** | This data element records the month in which the crash occurred. |
| **ROUTE** | This data element identifies the route signing of the traffic way on which the crash occurred |
| **RUR\_URB** | 1 (Rural), 2 (Urban), 6 (Trafficway Not in State Inventory), 8 (Not Reported) and 9 (Unknown). |
| **WEATHER** | This data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material |
| **rest\_use** | This variable shows whether the driver had any type of restraint on the body (1) restraint was present (2) restraint was absent |
| **COUNTY** | This data element records the location of the unstabilized event with regard to the County. The codes are from the General Services Administration’s (GSA) publication of worldwide Geographic Location Codes |
| **DAY** | This data element records the day of the month on which the crash occurred |

Target Variable

|  |  |
| --- | --- |
| **DRUNK\_DR\_new** | It states if the the driver was drunk (1) or not drunk (0) |



1. **BI Model:**

We have followed the below process to find the key insights in the data sets.

Managerial

conclusion

Problem Definition

Data

Reduction

Data

Cleaning

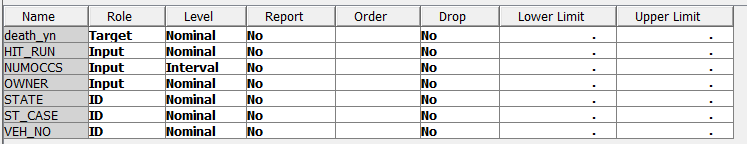
Data Collection

Predictive Modelling

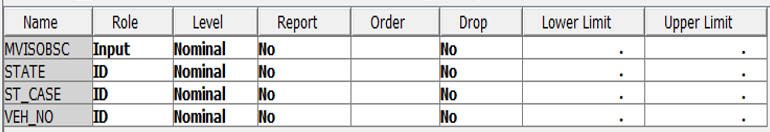
Descriptive Statistics

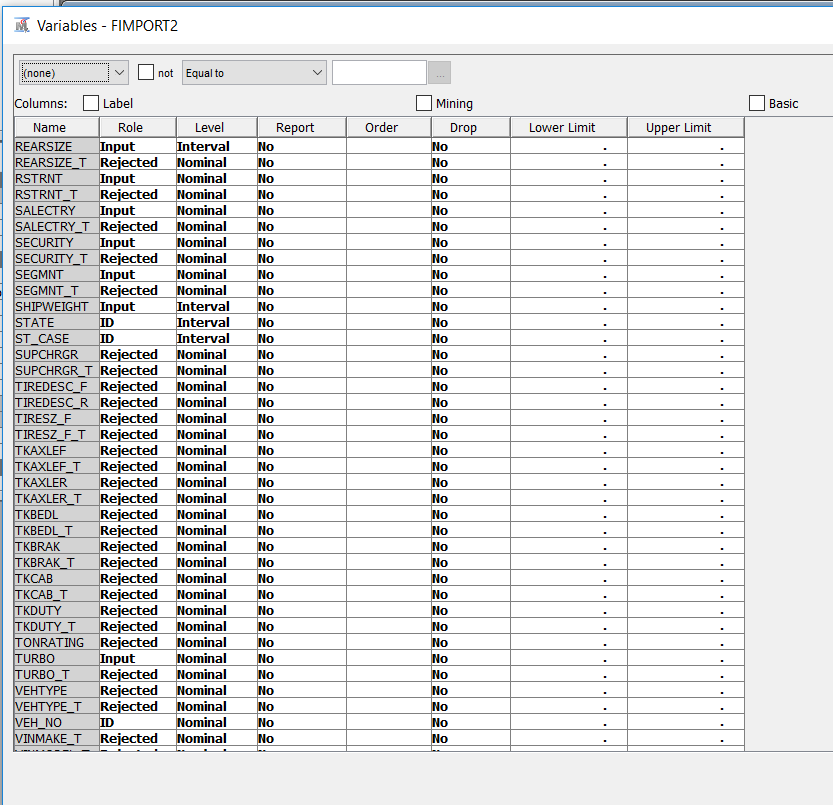
1. Data Preprocessing
   1. **Data Collection-**
      1. **Logistic Regression model:**

We have used 3 ‘File Import’ nodes to merge relevant data from ‘vindecode.csv’,’vehicle2.csv’ and ‘vision.csv’. We have kept the three variables STATE, ST\_CASE and VEH\_NO as ID in the definition and have rejected the variables which we will not be using according to our understanding. Following are the variable descriptions for the data files-

***Vehicle2.csv-*** 

***vision.csv-***



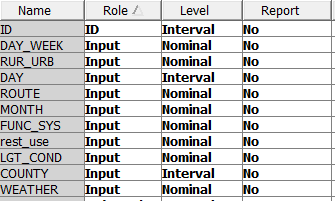
***Vindecode.csv-***   


* + 1. **Decision Tree model**:

For decision tree analysis, we have merged accident, vehicle and person data files. We have kept the three variables STATE, ST\_CASE and VEH\_NO as ID in the definition. We used DRUNK\_DR\_new as the target variable which specify the following:

1 – fatality due to drunk driving

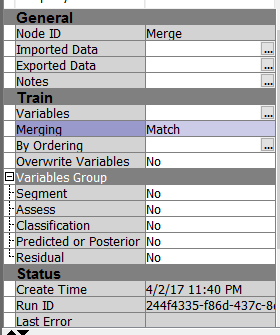
0 – no fatality





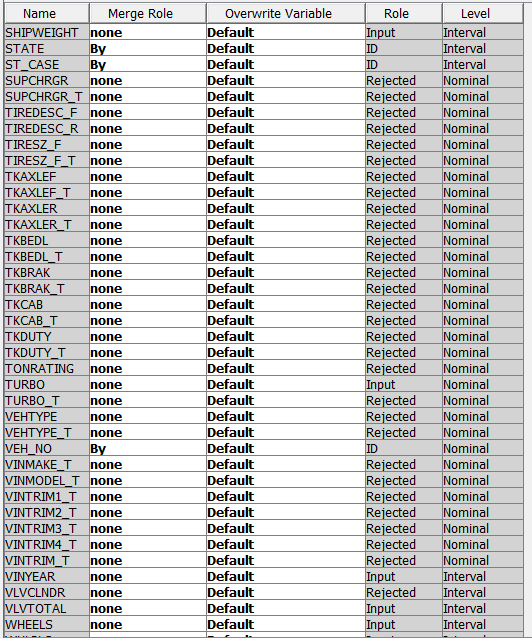
* 1. **Data Cleansing**
     1. **Logistics Regression model**

After getting the data from these three sources, we have merged the data with the help of ‘Merge’ node, we have Merged three datasets using the following setting-

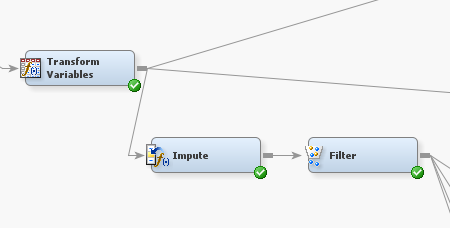


We implemented multiple steps in the code to verify the proper transformation of the raw data at each step in the data cleaning process. After creating the primary key, we ran a proc frequency to ensure that there is no data repetition or redundancy in the dataset.

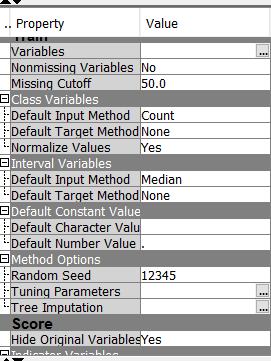
Here, we have selected the Merging as ‘Match’ which specifies the type of merging to perform. By ‘Match’, we combine observations from two or more data sets into a single observation in a new data set according to the values of common variables If One-to-One is selected, combine observations from two or more data sets into a single observation in a new data set. All data sets are sorted by the BY variables. (In our case, the BY variables are STATE, ST\_CASE and VEH\_NO. Following is the snapshot of the Merge node-



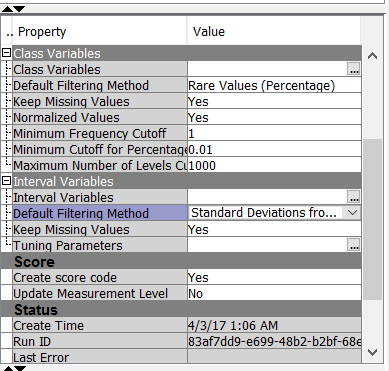
* 1. **Data Reduction (Data Filtering)-**

After the data, has been transformed, we check for missing values in the data and with the help of the ‘Impute’ node in the SAS miner, we will impute the missing data with the specified data. For the categorical values, we have imputed using the ‘Count’ setting whereas for Interval variables, we have used ‘median’ to impute data. We have set the Missing cut-off data to 50.0.

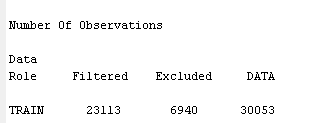
Following is the snapshot of the Impute node settings-

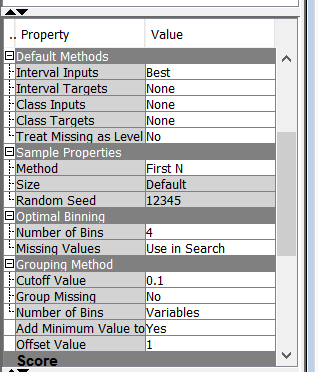


To check the outliers in the data and to handle them, we have used the ‘Filter’ node which will do the Outlier analysis and filter the data. We have set the ‘Default Filtering method as ‘Standard Deviation from the Mean’’. Following is the snapshot of the ‘Filter’ node-

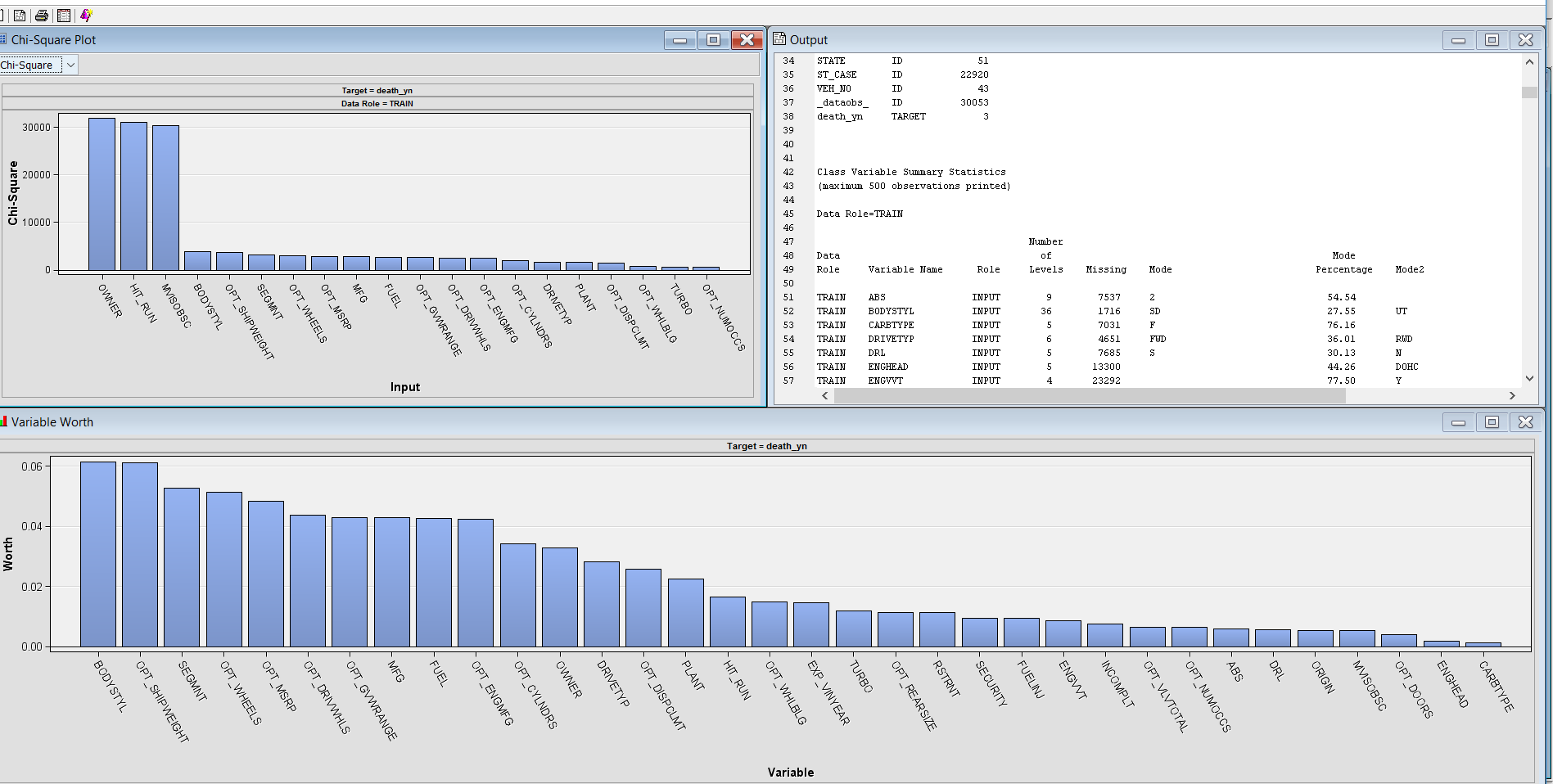


After Filtering the data, we get the following statistics-



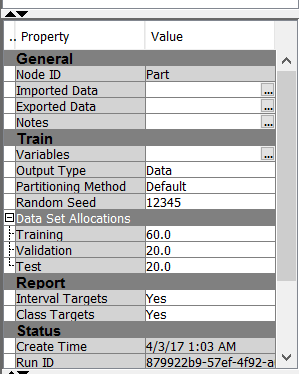
* 1. **Data Transformation-**  
     To remove the skewedness from the data, we used the ‘Data Transformation’ node and set the ‘Optimal Binning’ to 4. We have selected the “Best” interval inputs for Default Methods. Following are settings of the ‘Transform Data’ node.   
     

After running the ‘Transform data’ node, following is the output-

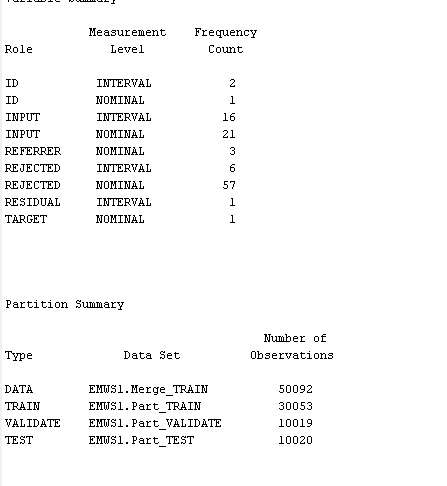


1. **Data Analysis**
   1. **Logistic Regression model**
      1. **Data Partitioning-**

To split and train the data, we have used ‘Data Partition’ node and have divided the data with ‘’Data set allocations” as 60:20:20 for Training, Validation and test data respectively. We have used the same data portioning nodes setting in decision tree as well. These properties define the percentage of input data that is used in each type of mining data set as shown below-

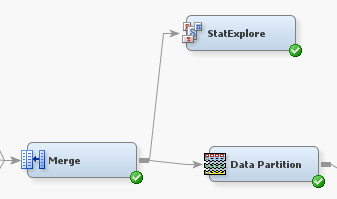


After partitioning the data, following is the summary of the variables and the number of observations as per the dataset allocations-

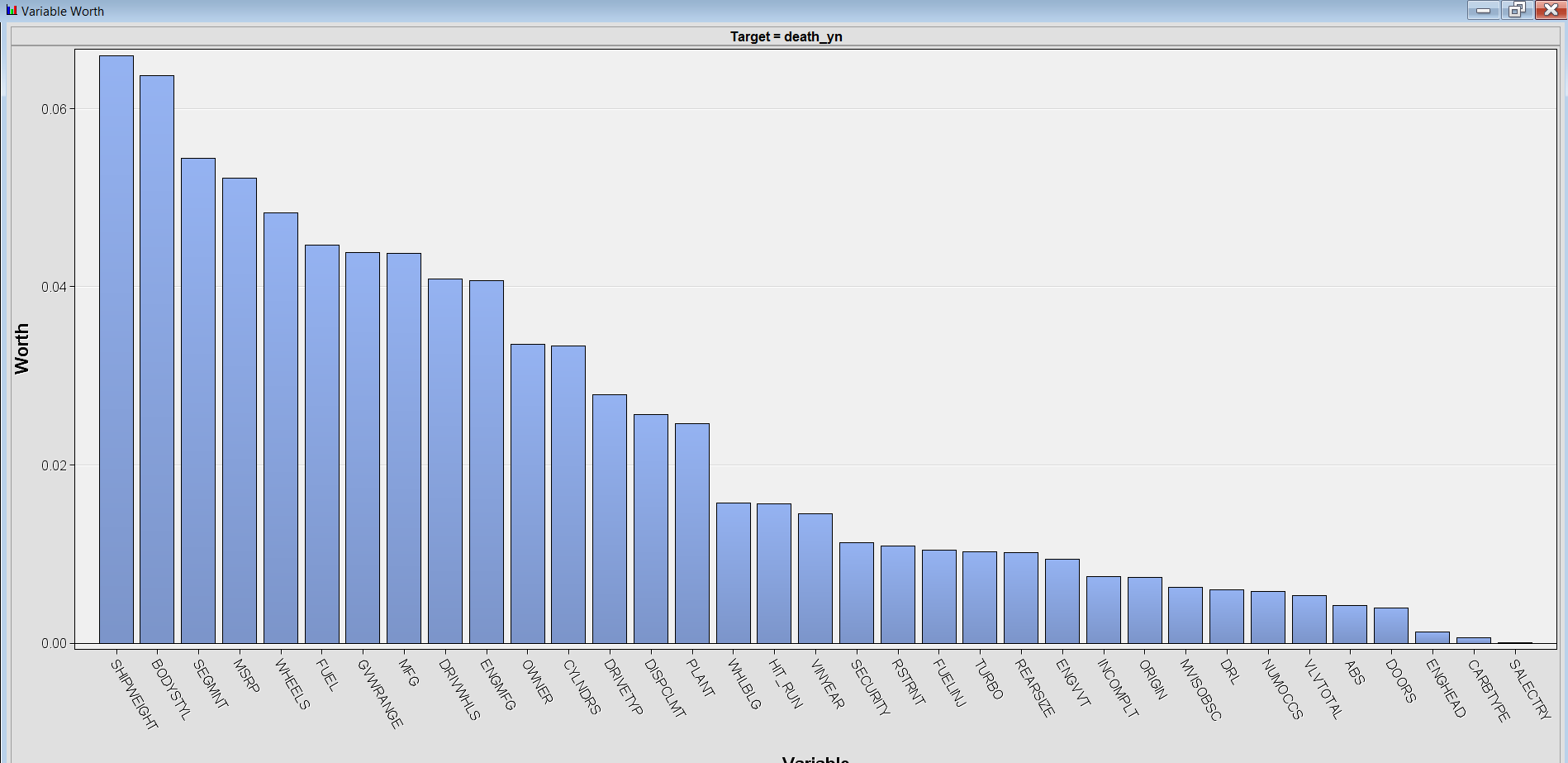


* + 1. **STATEXPLORE**

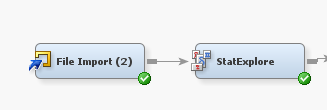
We have used the STATEXPLORE node to see the statistics of our data set



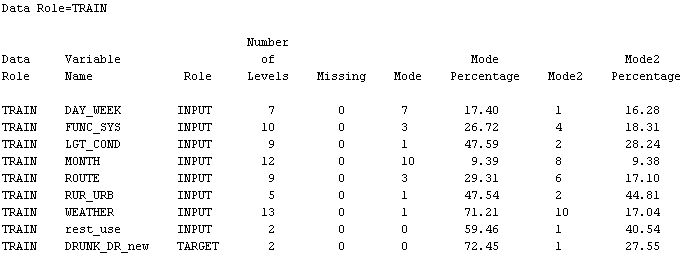
We got the following output after running the STATEXPLORE node which indicates that the ‘shipweight’, ‘bodystyl’, ‘segment’ hold the maximum worth and helps us in predicting the fatalities.



* 1. **Decision Tree model**
     1. **STATEXPLORE**



The below is the output of STATEXPLORE node for the decision tree model:



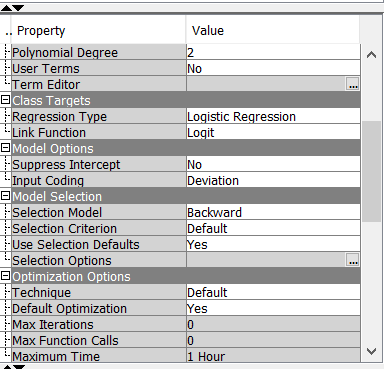
1. **Modelling**
   1. **Merged file (accident+vehicle+person) for factors affecting death by automobile-**
      1. **Regression:**

The target variable is binary which specifies if the person died or not. Hence, we chose the Logistic regression model to analyze the data and gain meaningful insights from the data.

Now, as we have got the filtered data, we will run the Logistic Regression of 3 different types. i.e. Stepwise, Forward and Backward and then we will compare the misclassification rates from each of them. As we chose the Forward, Backward, or Stepwise effect selection method for Model Selection, we can specify a selection criterion to be used to select the final model.

We added the Regression nodes for each of the regression types and set the properties as shown below-   
1. We set the Regression type as ‘Logistic Regression’ and Selection models as follows-

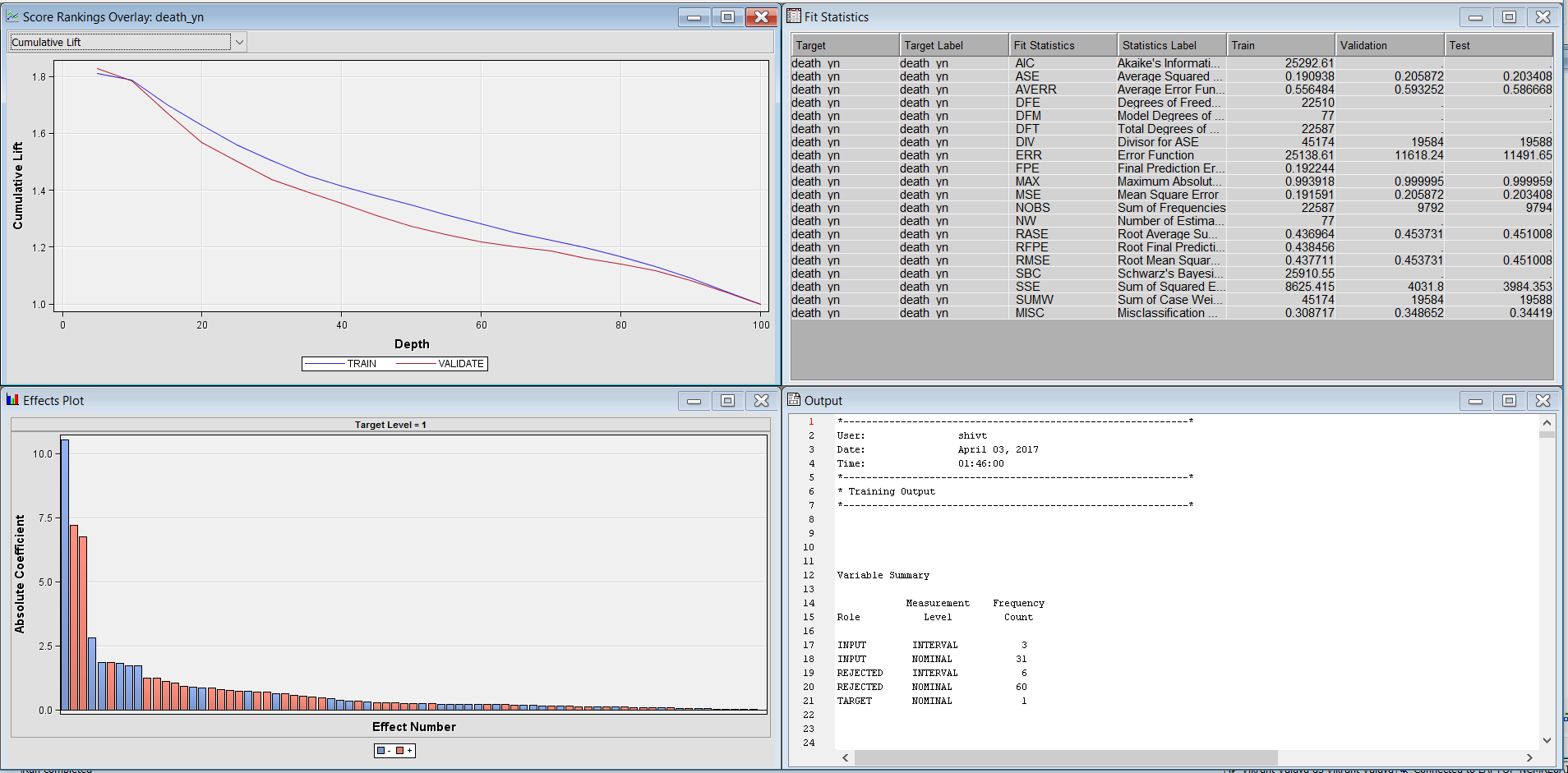
**Backward, Forward and Stepwise selection model-**



**Model Selection:**

After running all the regressions, we have added the Model Comparison node to analyze which model has the best characteristics. According to the result from model comparison, the forward regression

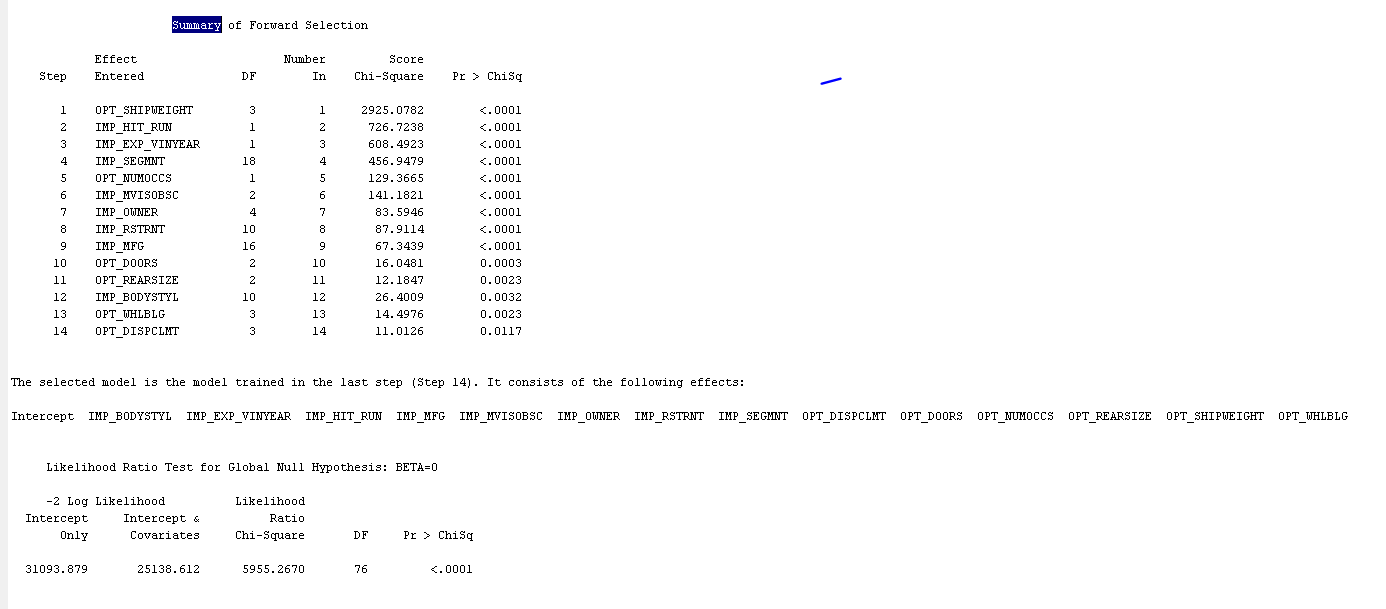
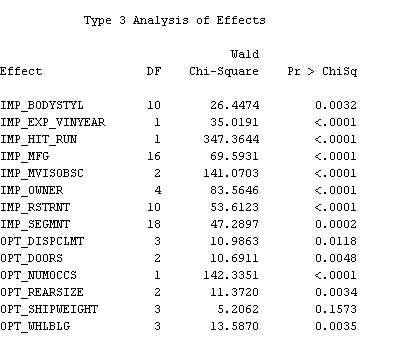
gives the best results and hence, we will be analyzing the same below.



**Interpretation:**

The misclassification rate for this model is 34.86% which means that the model is 65% accurate.

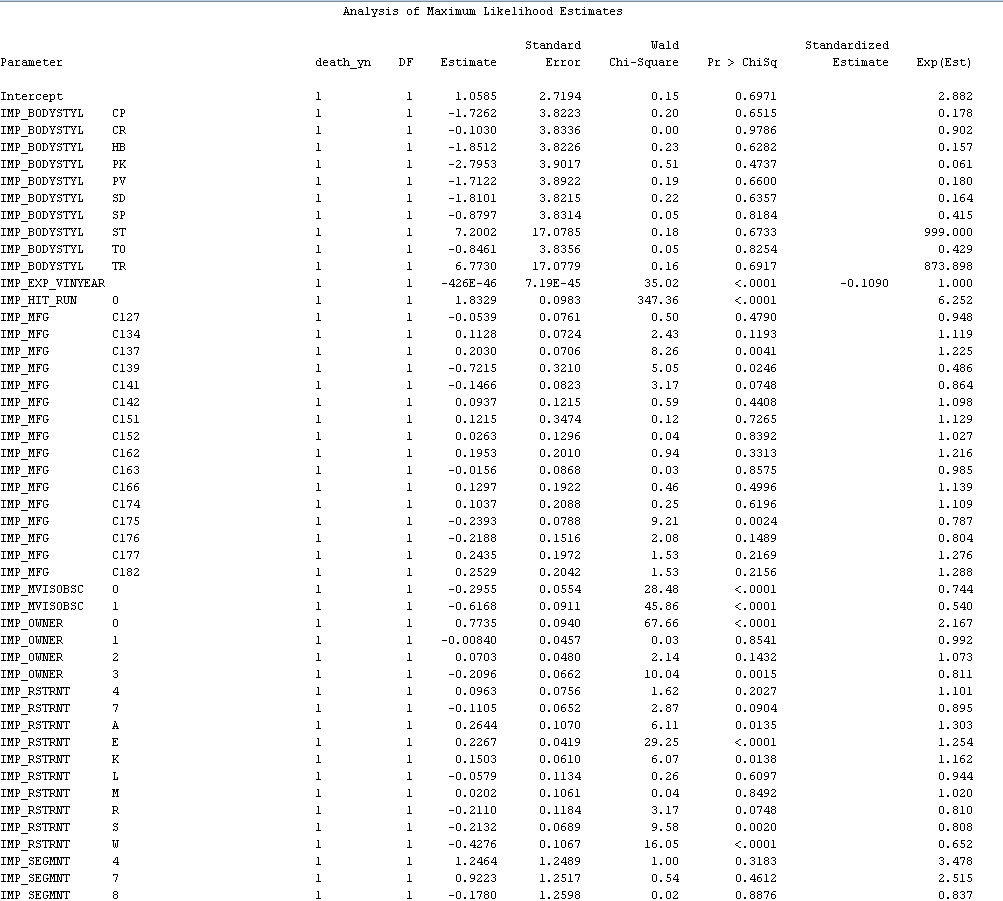
Logistic regression uses maximum likelihood (and not sum of squared errors) to estimate the model parameters. Results show that the model is highly significant based on chi-square test.

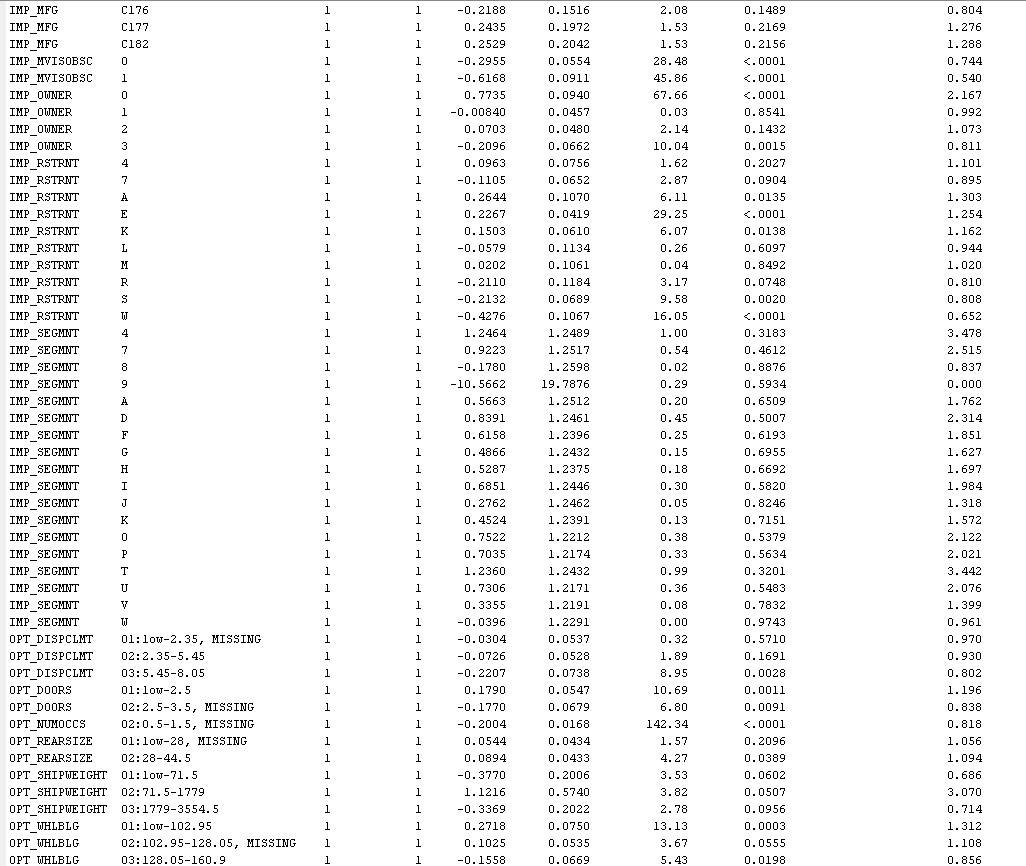
 

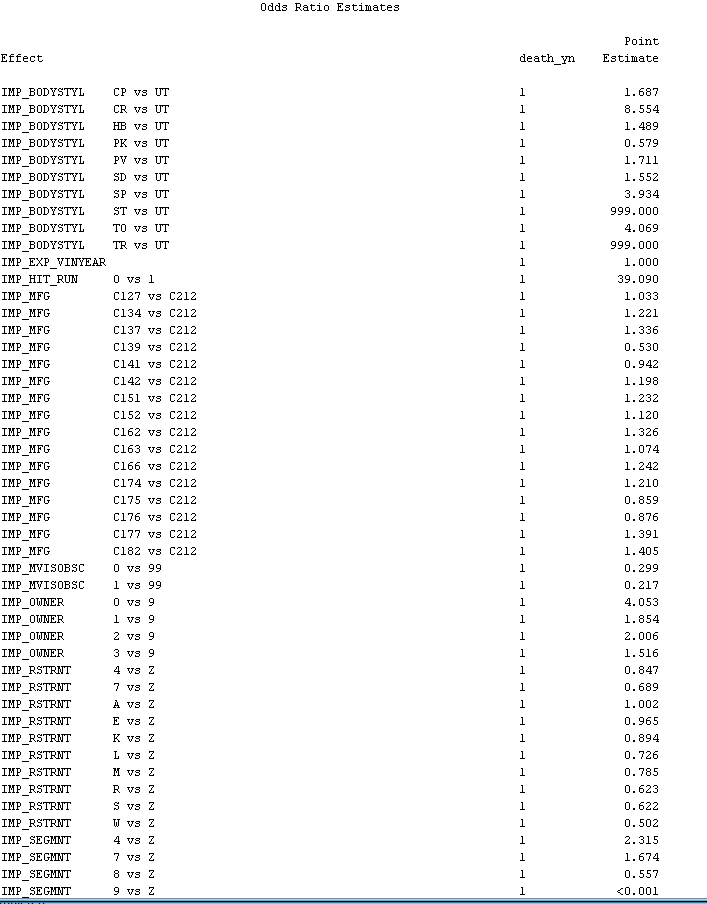
The factors that affect the most and are significant are:

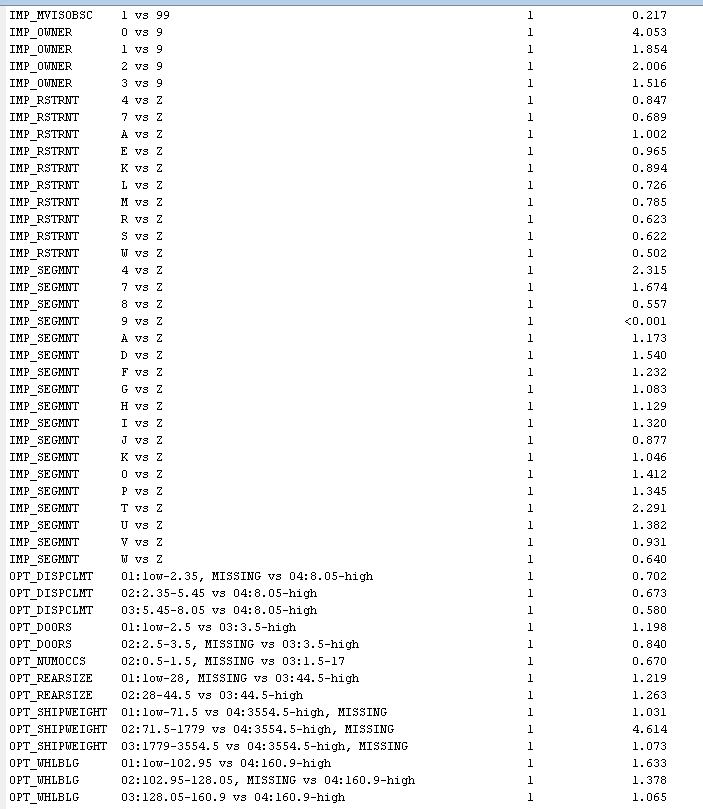
* IMP\_EXP\_VINYEAR (Vehicle Manufacture Year)
* HIT\_RUN(if the accident was hit and run)
* Manufacturer
* Driver Vision Obstruction
* Restraint
* Vehicle Displacement
* Number of doors in the car
* Number of occupant
* Vehicle length

Maximum likelihood estimates of this data are as follows:







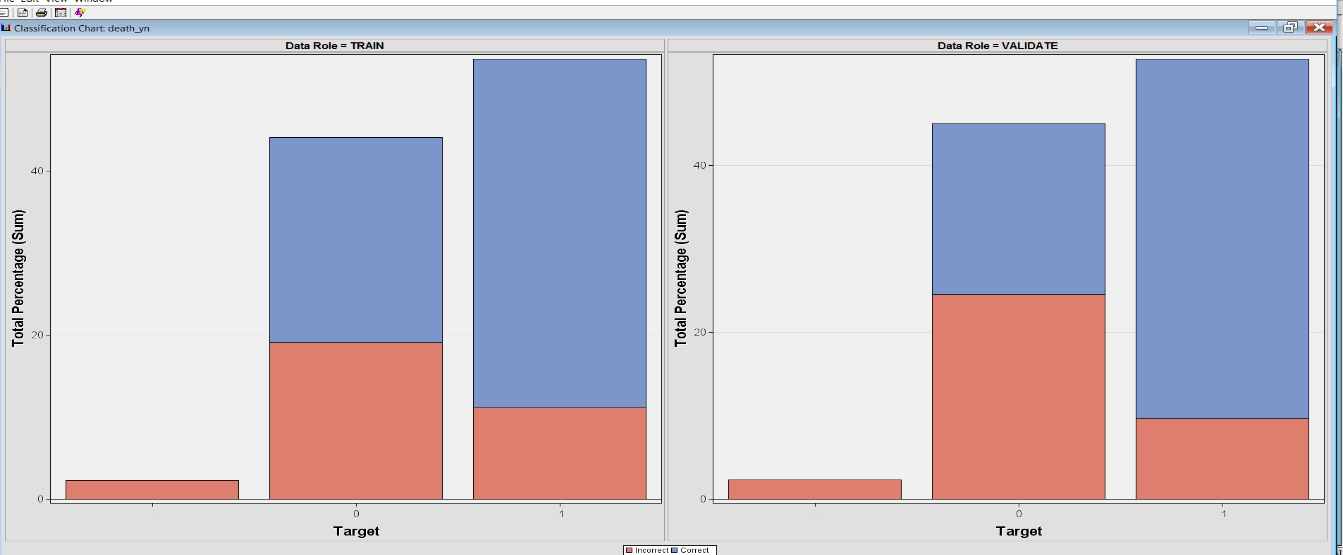


**Estimated Probability:**

**Change in log odds :**

Hence, we can see can interpret the coefficients of the independent variables through the Odds Ratio Estimates table.

For example, if the body style of the car was coupe rather than pickup truck, the ratio estimate that the person died increased by a point estimate of 1.687. Similarly, we can interpret the other significant variables as well.

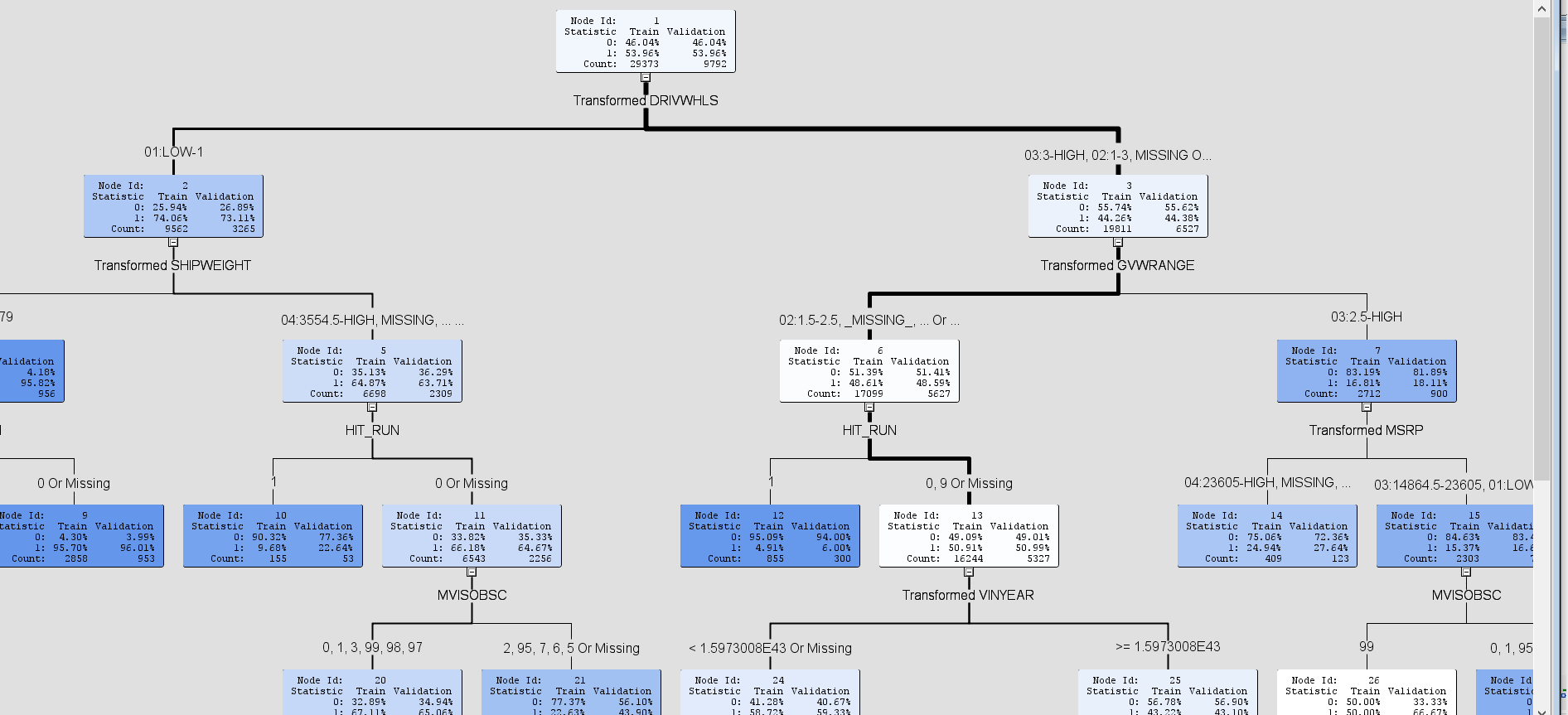


* + 1. **Decision Tree-**

Decision Tree modeling is one of the most widely used predictive modeling technique. It is useful to make sense of data through visualizations. It helps accommodate nonlinear associations when we have input variables with one or more target variables. It is also the best model to handle data with some missing values. We use decision tree modeling to evaluate relationships between data in our dataset.

To run the decision tree model, we have also added a Decision Tree node to compare with the Regression model so that we can select the optimum model.

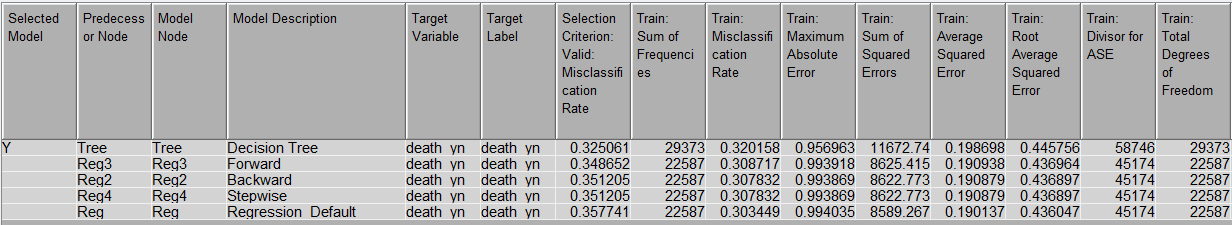
**Following is the output of the Decision Tree-**



The main factor for distinguishing is the DRIVWHLS i.e. Number of wheels driven by the power train. It is divided into Low and High. And, then it is further classified into SHPWEIGHT ( Low and High ) etc. as can be seen in the figure above.

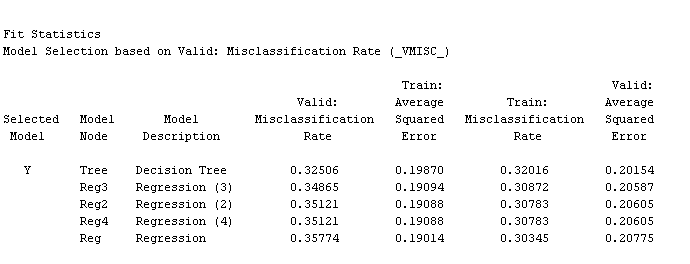
Now, we add a ‘Model Comparison’ node and connect the three Regression nodes and the ‘Decision Tree’ node with it to compare the Misclassification Rates.

Following are the results that we get after running the Model Comparison node-

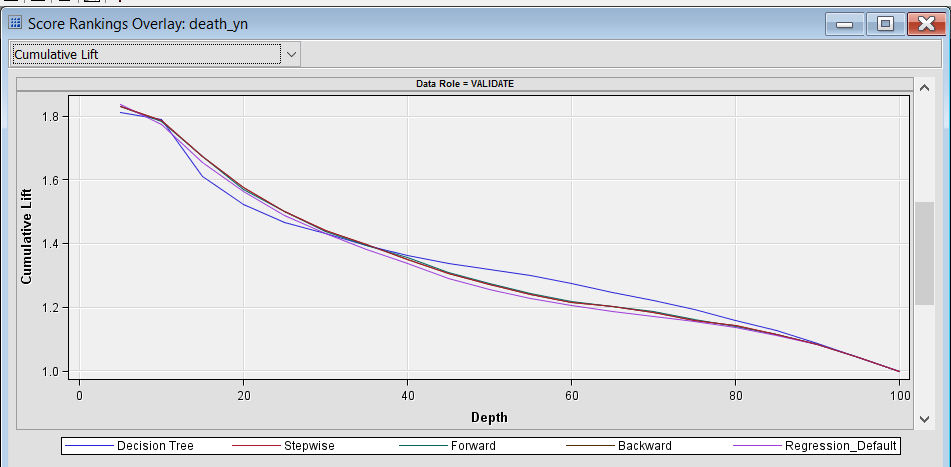


We can see that the Misclassification Rate for the Validation data is minimum (i.e.0.3250 which comes out to be 32.5%) for the Decision Tree wherein the Misclassification rates for Regression Models with Stepwise, Forward and Backward selection criteria are relatively higher.

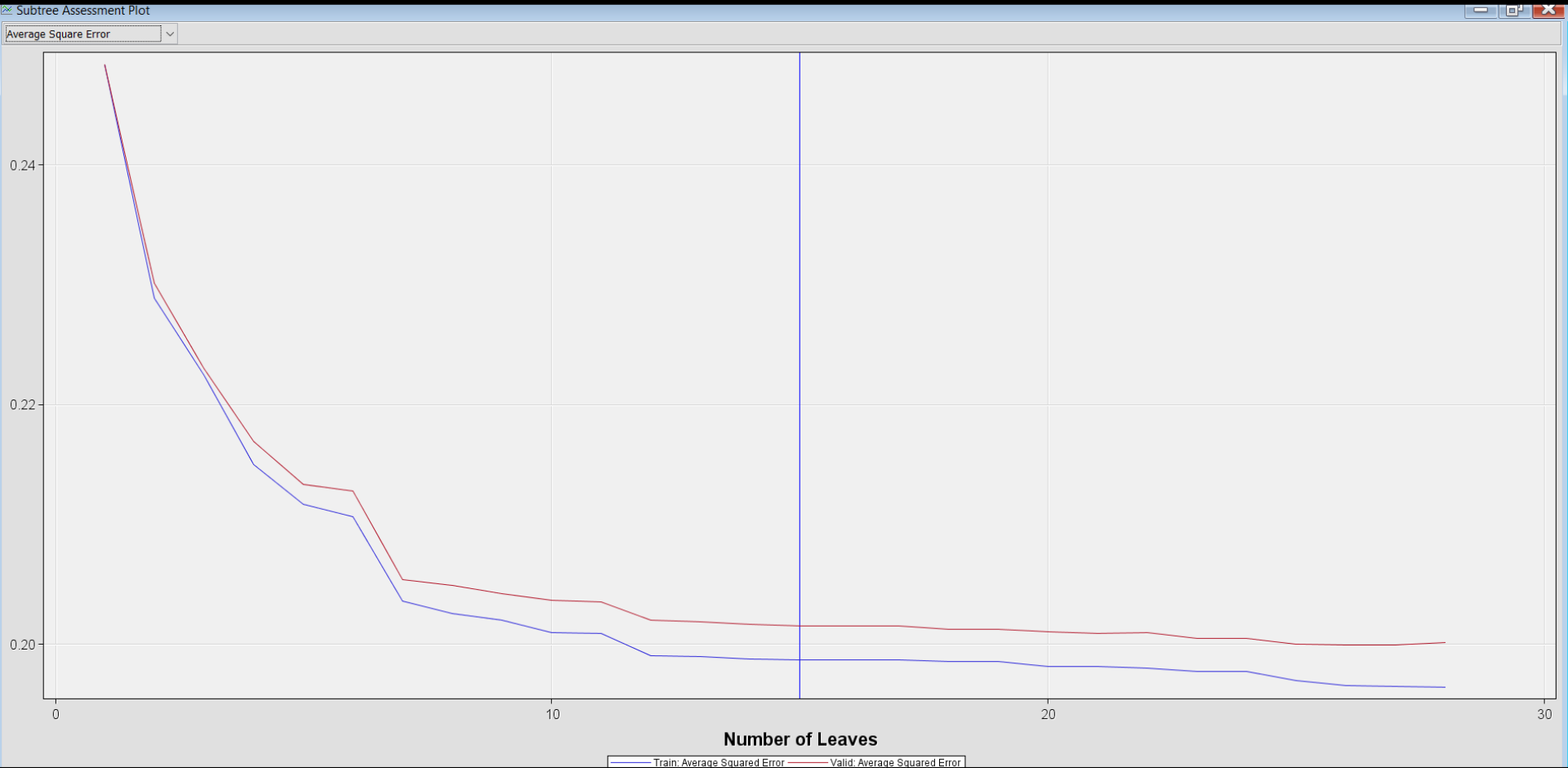
**Following are the Fit statistics-**



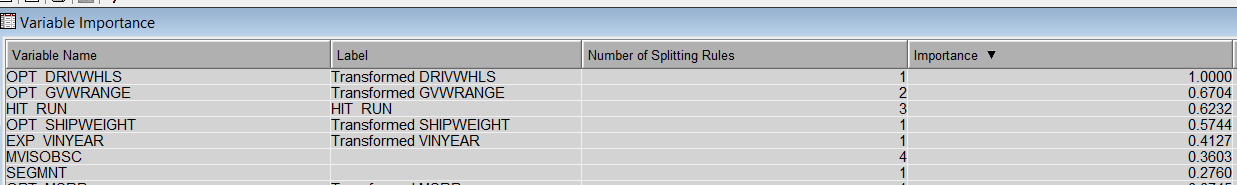
**Following is the Cumulative Lift Graph-**



**Subtree Assessment plot-**

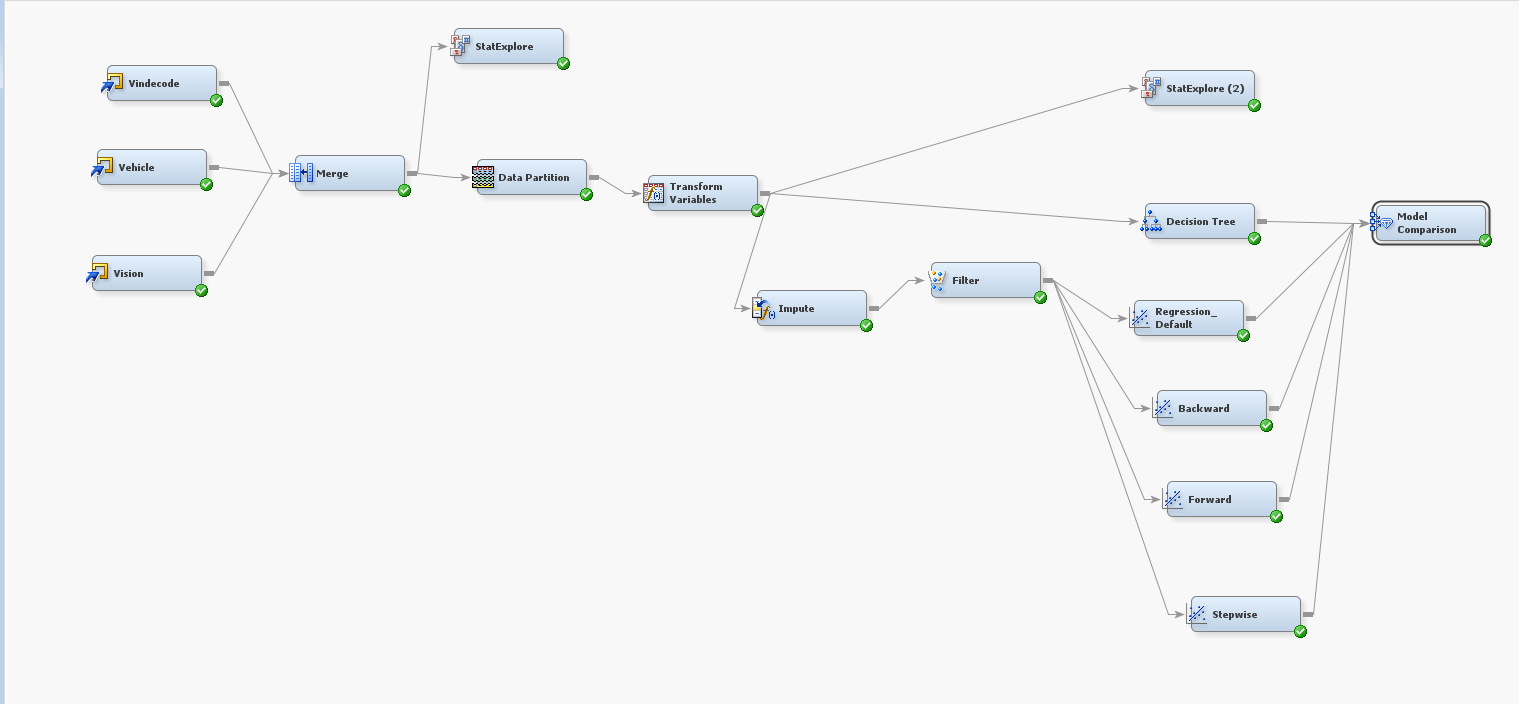


We can see that for Decision Tree (Blue in color), there is significant amount of Lift than the other models as the depth increases.



From the above results , we can say that DRIVWHLS ,GVWRANGE and HIT\_RUN are the most important variables in the decision tree. Including these 3 variables, SHIPWEIGHT and EXP\_VINYEAR are also important.

**Enterprise Miner Diagram for Regression:**



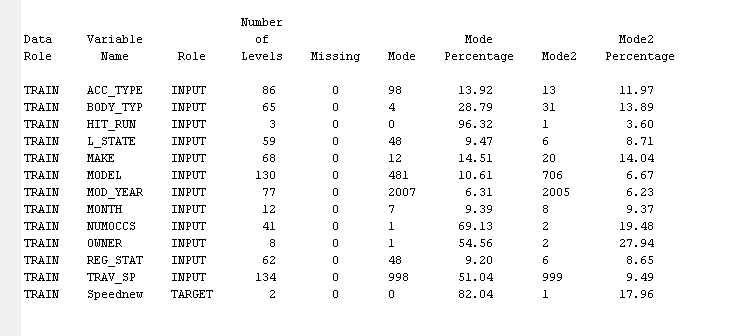
* 1. **Merged file for speeding**

Analyzing and Predicting deaths due to vehicle speeding:

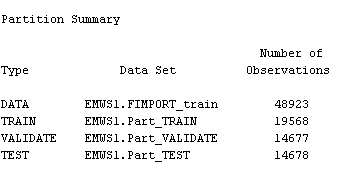
The economic cost to the society of speeding-related crashes is estimated by the NHTSA to be $28 billion per year which has been gradually increasing. Vehicle speeding is a deliberate and calculated behavior where the driver knows the risk but ignores the danger. Through our analysis here we are trying to design a model which can optimally predict the factors which lead to speeding and hence curb speeding related fatalities

Our objective is to analyse if the vehicle was speeding and resulted in death of one or more persons and what factors where prevalent during the automobile accident. We set “speednew” as the target variable and then analysed the classification of all the observations.

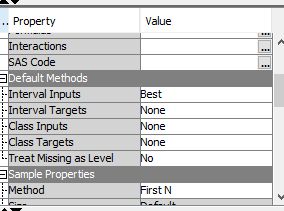
Using the Stat Explore node in SAS Enterprise miner we observed that data is clean and doesn’t have any missing values.

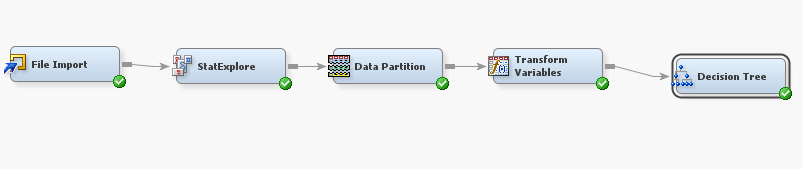


A data partition node was added and the data was split into training (60%), validation (20%) and test (20%)



A transform variable node was also attached with default method of transformation chosen as “best”.



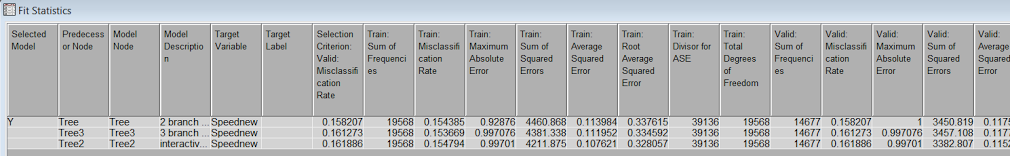


**Decision Tree**

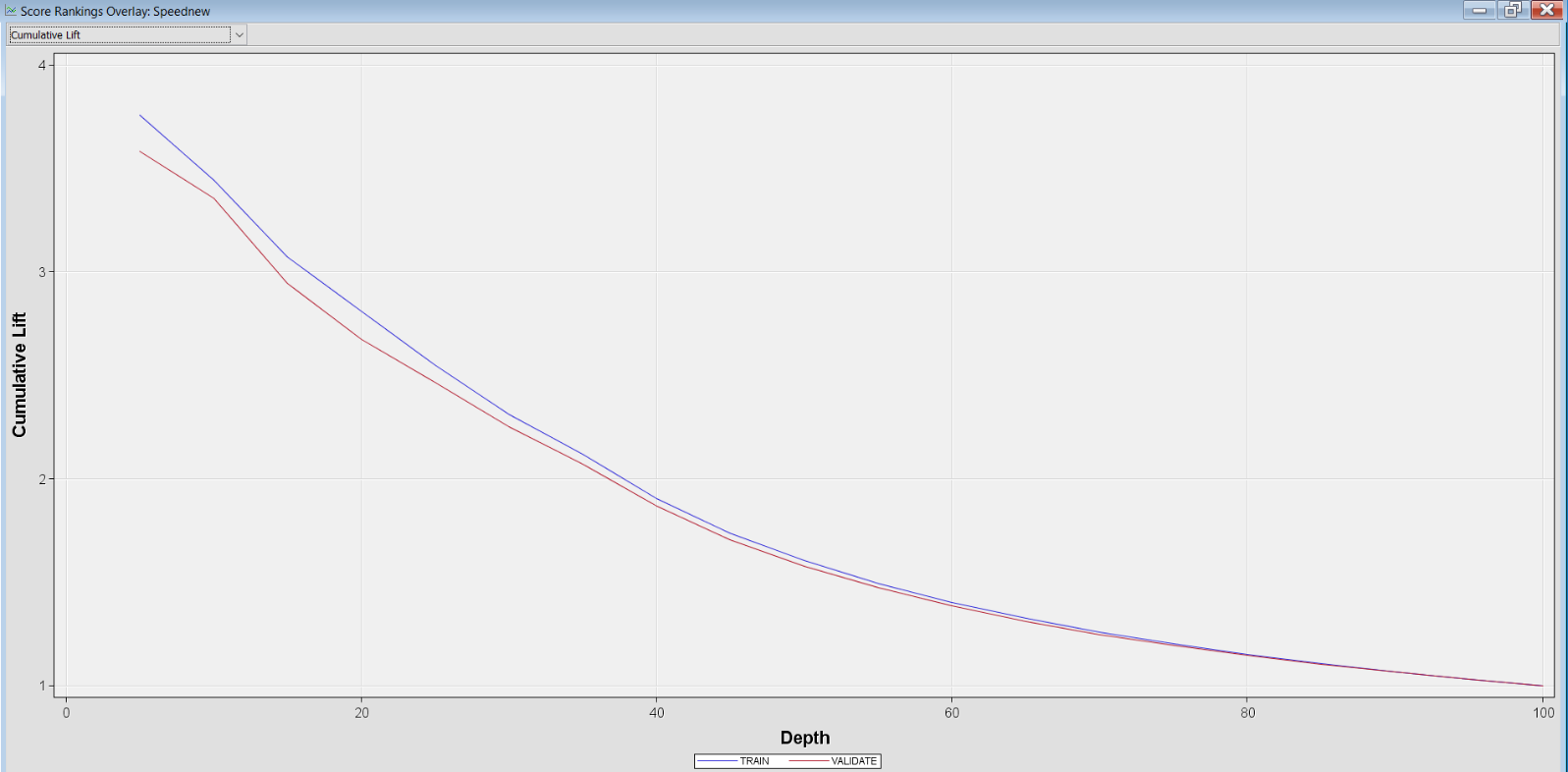
We used three decision trees.

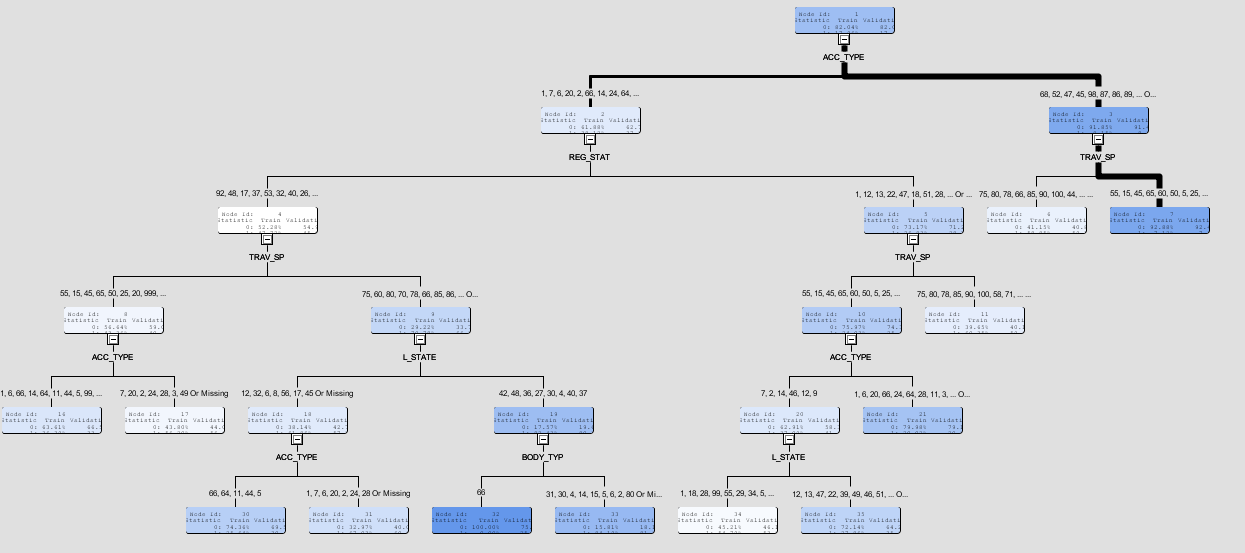
* Two Branch
* Three Branch
* Interactive

We executed all these decision trees and compared through model comparison to see which tree has the best fit. The output of model comparison is as follows:

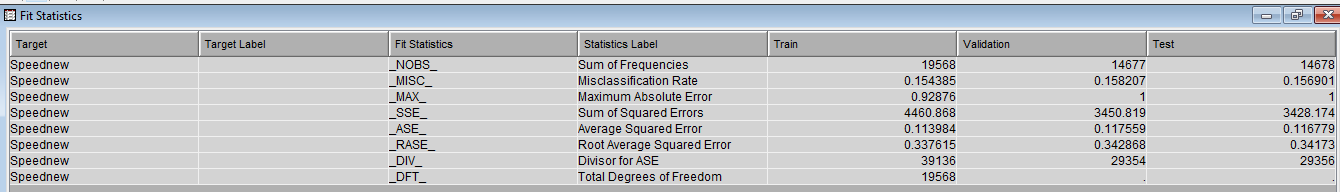


Using the misclassification method for assessment, the two branch decision tree model has best fit. The tree formed is as follows and has 12 leaves





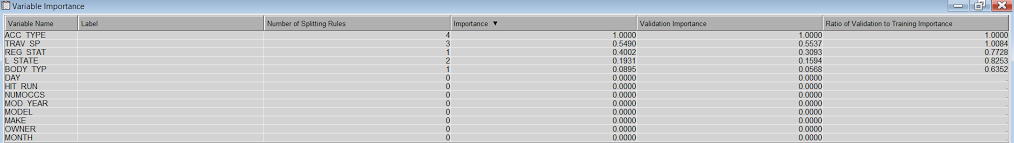
The following table shows the miscalssification rates of the training, validation & test data.



The misclassification rate of the training, validation & test data is 15.43%, 15.82% and 15.69 respectively.

We also observe that the optimal decision tree has very low misclassification rate in validation and test data and is comparable to the training data.

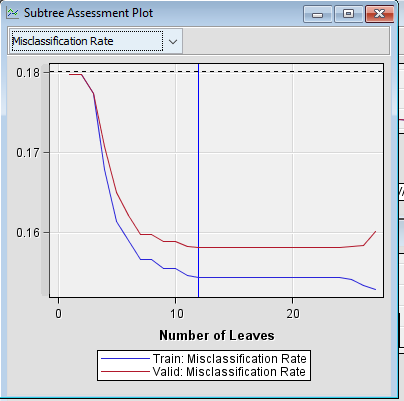
The variable importance statistics show the following result:



Hence the most significant variables for speeding are:

* ACC\_TYPE: First harmful event
* TRAV\_SP: Speed at which the vehicle was travelling
* REG\_STAT: State where the vehicle is registered
* L\_STATE: State where the license of driver was obtained from
* BODY\_TYP: Type of the body of the vehicle

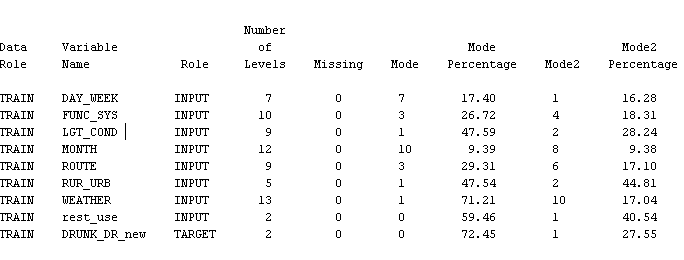
The subtree Assessment plot for the optimal decision tree is:



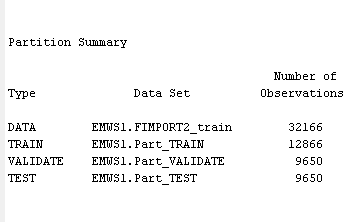
* 1. **Merged file for analyzing drunk driving**

In 2014, 9,967 people were killed in alcohol-impaired driving crashes, accounting for nearly one-third (31%) of all traffic-related deaths in the United States. Our objective is to develop decision rules to predict the factors that where prevalent during an automobile accident and resulted in death of one or more persons when the driver was drunk or not. For this objective, we decided to utilize the decision tree model and logistical regression in the SAS enterprise miner. We set “DRUNK\_DR\_new” as the target variable and then analyzed the classification of all the observations.

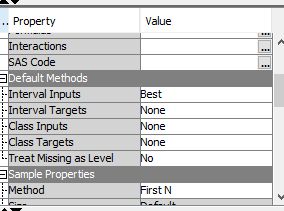
Again, Using the stat Explorer node in SAS Enterprise, we observe that data is clean and doesn’t have any missing values



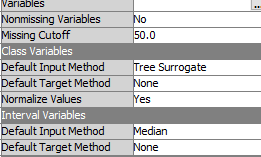
A data partition node is added and the data is split into training (40%), validation (30%) & test (30%)



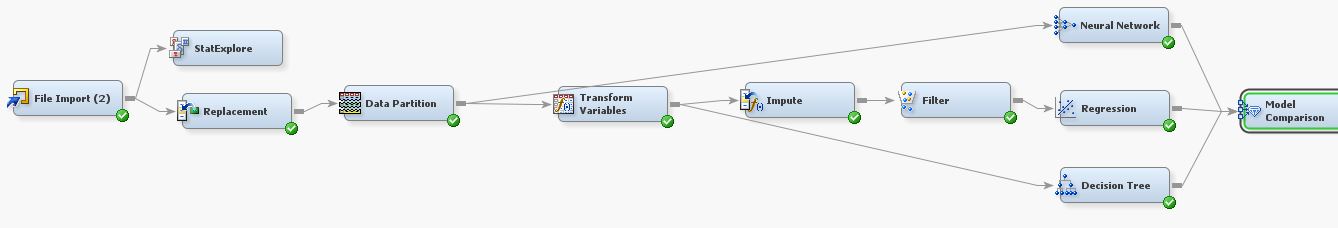
A transform variable node is also attached with default method of transformation chosen as “best”



We added data Impute node before the regression node and set the properties as



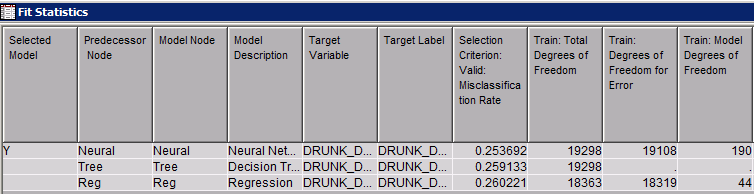
Missing cut off is set as 50 as we don’t want the imputed value to exceed 50% of existing value.

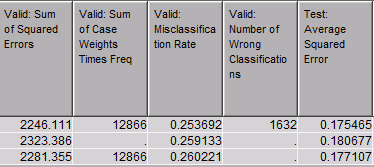


On this dataset we applied the following three data analysis approaches to analyse the best fit:

* Neural Network
* Logistic Regression
* Decision Tree – Optimal

We assesed these model using the model comparison node and obtained the following output:

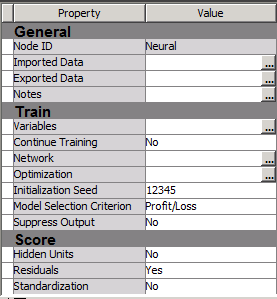


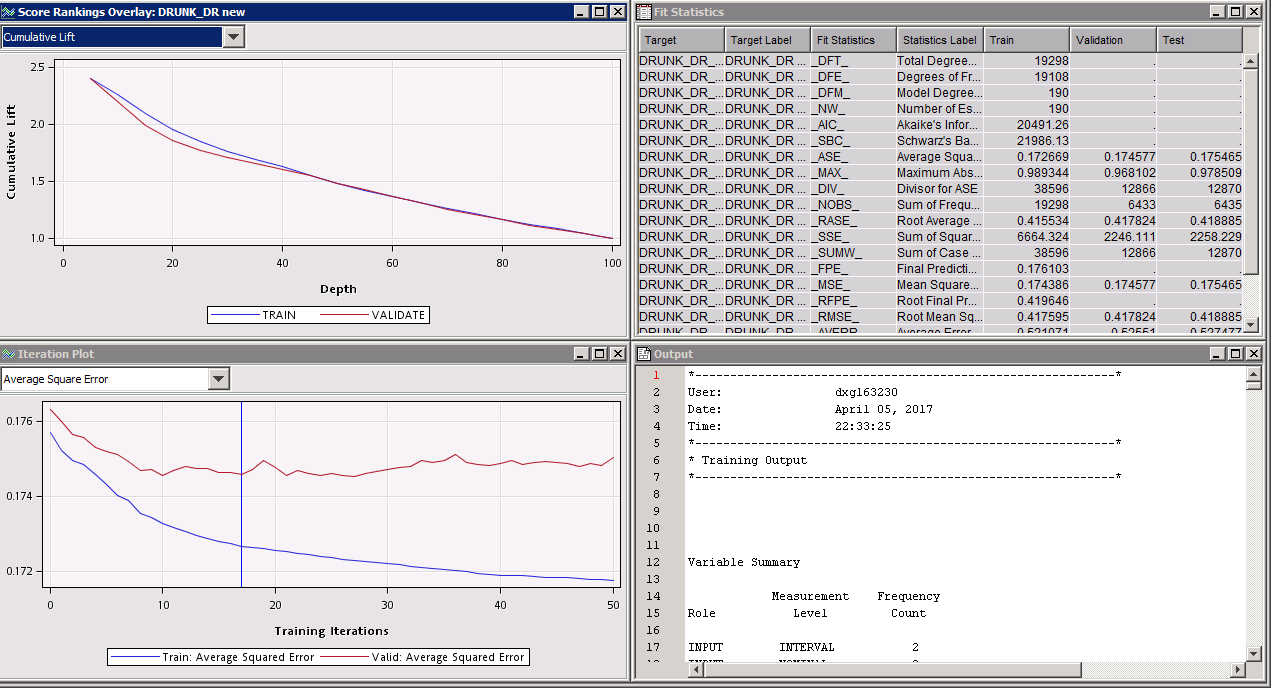


The above statistics show that neural network has the lowest misclassification rate ( 0.2536) and has the best fit among the three models. The second best model is the optimal decision tree.

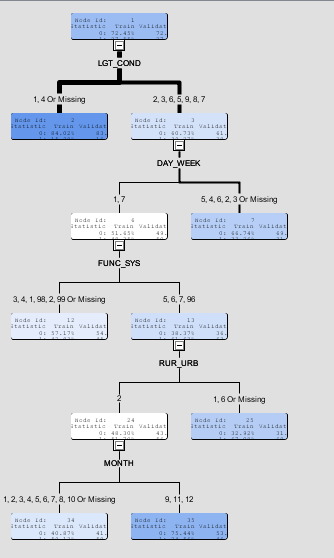
**Neural Network**

Neural network can accommodate a larger variety of nonlinear relationships between the independent and the target variable.

Neural networks have the ability to derive meaningful insights from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex.



**Decision tree**

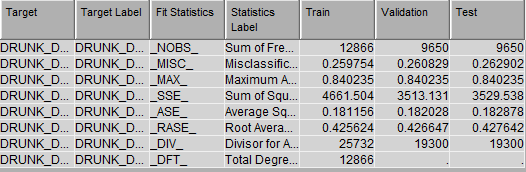


The most important variables for the decision tree are:

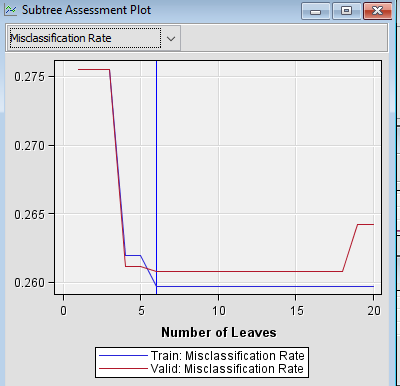
* LGT\_COND : Light conditions
* DAY\_WEEK : Day of the week
* FUNC\_SYS : Type of the road
* RUR\_URB : If the road falls in an urban or rural area
* MONTH : Month of the year

Here again we observe that the misclassification rate is low . valiadation and test data sets are almost comparable to the training data set.

The following table shows the miscalssification rates of the training, validation & test data.

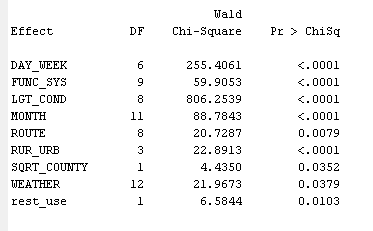


The subtree assesment plot gives the following result which is the optimal decision tree

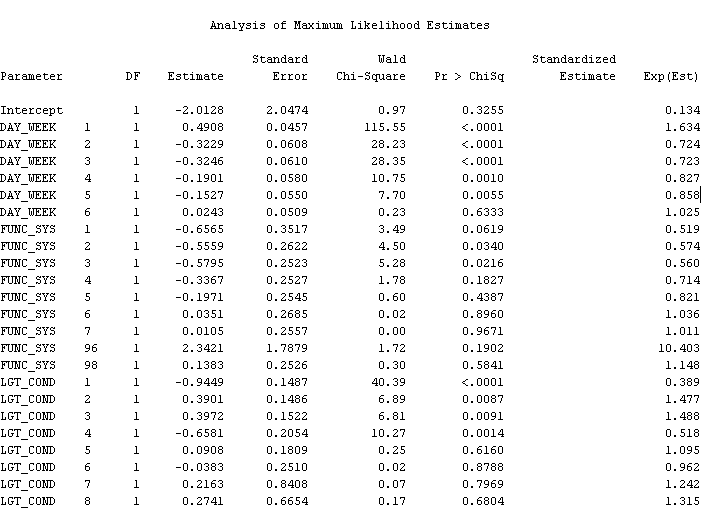


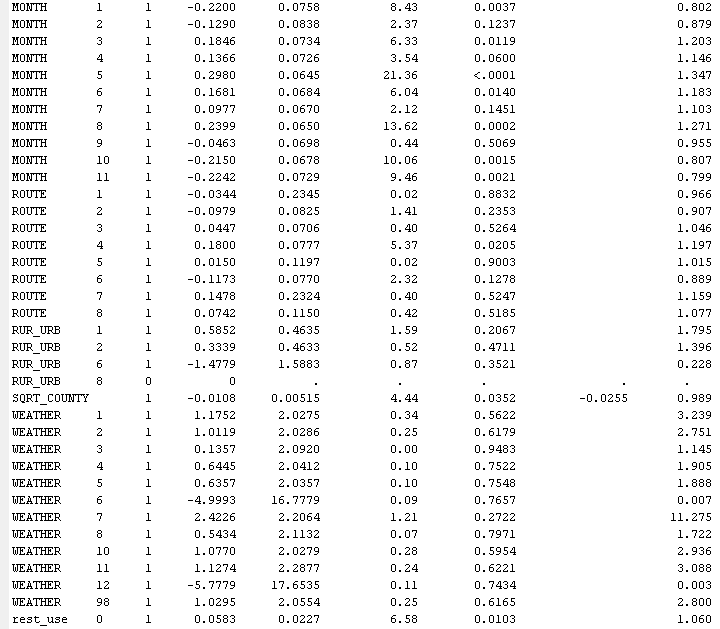
**Logistical regression**

We executed the logistic stepwise regression model and got the model with the significant variables as follows:



**Maximum Likelihood estimates**



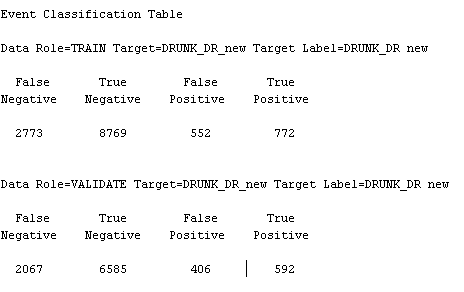


The chart above shows all the variable which was found to be significant as they < 0.05 therefore they carry lot of information about the target variable.

The most significant variables which explain the most prevelent conditions when the automobile fatility occurred due to drunk driving are:

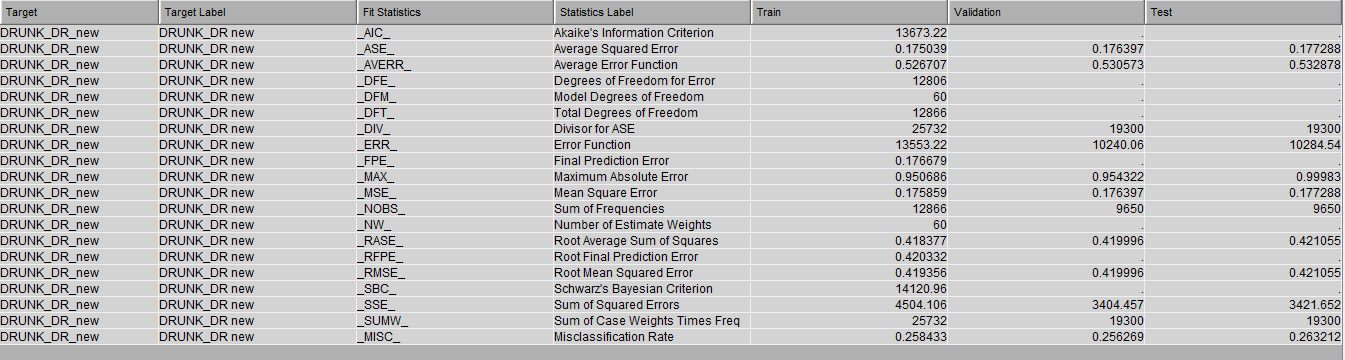
* Week of the accident
* Month of the year
* Light Conditions
* Rural or urban road
* Type of route(Ex: Interstate, National Highways etc.)

**Confusion Matrix:**

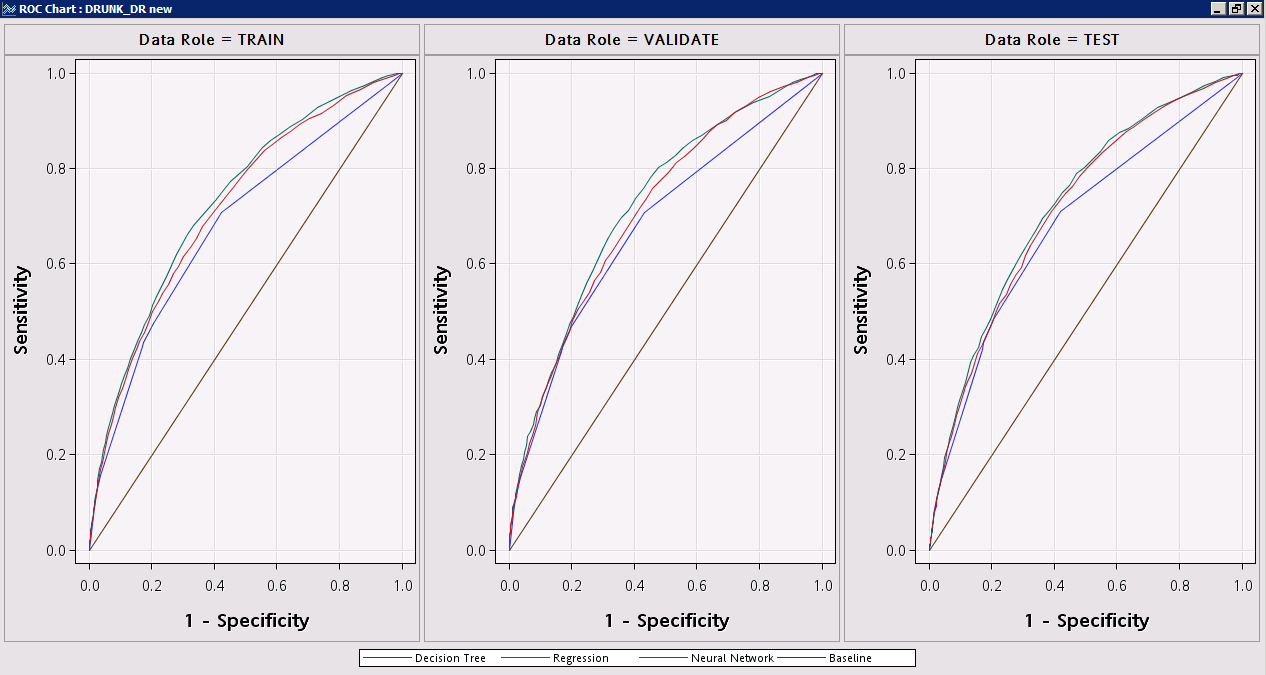


We can observe that the sum of true positive and true negative is much higher than the sum of false positive and false negative. Therefore we can conclude that based on event classification table our model is almost accurate.

**Fit statistics**

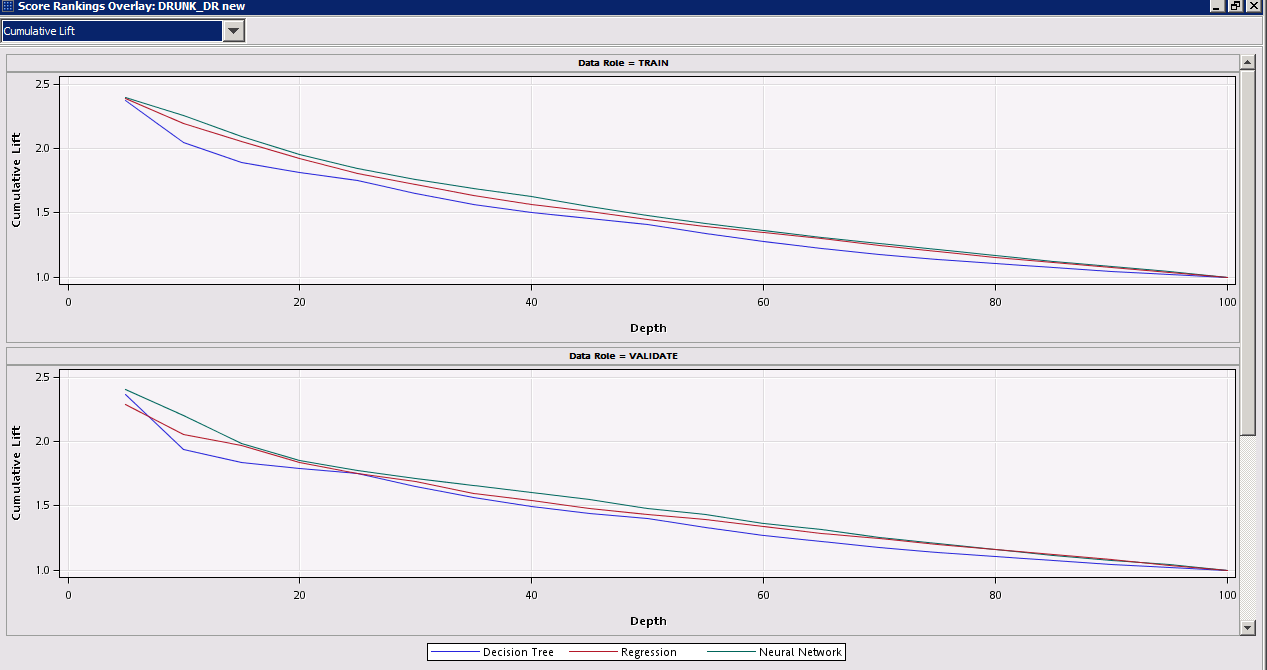


**ROC Curve**



The area under the neural network curve is higher than the area under the decision tree and regression curve. This shows that the quality of the model using neural network is better.

**Cumultive Lift**



We see that the cumulative lift curve for the Neural network is smoother and higher than decision tree and regression, thereby proving that neural network is a better model.

1. **Business Implications and Value**

The Traffic fatalities analysis helps in controlling the factors affecting death rate to the Traffic Authorities and Governments and brings extra-revenue to the governments

* Instituting key registration and increasing penalties on those who break the traffic rules for making the existing law stricter for the public-safety
* Reducing advertising and alcohol sponsorship of events
* Restricting alcohol use or sale on Highways or crowded areas
* Based on our findings, formulating new laws specific to decrease to deal with the traffic fatalities rates

1. **Conclusion**

After analyzing the automobile fatalities (FARS 2015) dataset we can help various authorities to understand the various factors that lead to automobile deaths, the conditions prevalent during speeding and drunk driving and helping them make data driven decisions. This in turn will increase their efficiency and will help them take measures that can reduce automobile fatalities.

We created three models to analyze the FARS 2015 dataset and concluded a number of interesting facts:

1. **Analyzing factors affecting automobile fatalities**  
   We regressed the deaths against a number of factors present during the accidents and leading to a fatality. We determined that bodystyle, manufacturer, car displacement, year of manufacture etc were the leading factors of automobile fatalities.
2. **Decision tree analysis of deaths due to speeding**  
   We created decision tree to analyze factors prevalent during speeding leading to death and found that Speed, registration state of vehicle, license state, body type of the vehicle and type of harmful event were the most significant variables.
3. **Analysis of deaths due to drunk driving**

In this dataset, we created three models to conclude the prevalent factors leading to fatalities due to drunk driving were Light conditions, day of the week, type of the road, type of the locality, month and route.

1. **References**

Galit Shmueli, N. P. (n.d.). Data Mining for Business Intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner. Wiley.

NHTSA. (2015). 2015 Traffic Fatalities. Retrieved from Kaggle: <https://www.kaggle.com/nhtsa/2015-traffic-fatalities>

Peter Christie, J. G. (n.d.). Applied Analytics Using SAS Enterprise Miner .

Shmueli, D. R. (n.d.). Getting Started with Business Analytics: Insightful DecisionMaking. CRC Press.