

# Executive Summary

Road safety is one of the major subjects within the transport policy of the United States of America. Our dataset contains granular descriptions about the traffic violation cases lodged in Montgomery County of Maryland. The majority of traffic violations, such as speeding or ignoring stop signs, are unintentional and they occur due to a lack of concentration rather than because drivers deliberately intend to break the law. It is important to understand the various factors that cause traffic violations so that we can make every effort to prevent them. With the advancement of technology, it is possible to analyze the past data using various data mining techniques and tools available today to understand why violations occur. This is exactly the objective of this project so that the result of the analysis can be used to create precautions and appropriate safety measures/reducing traffic tickets.

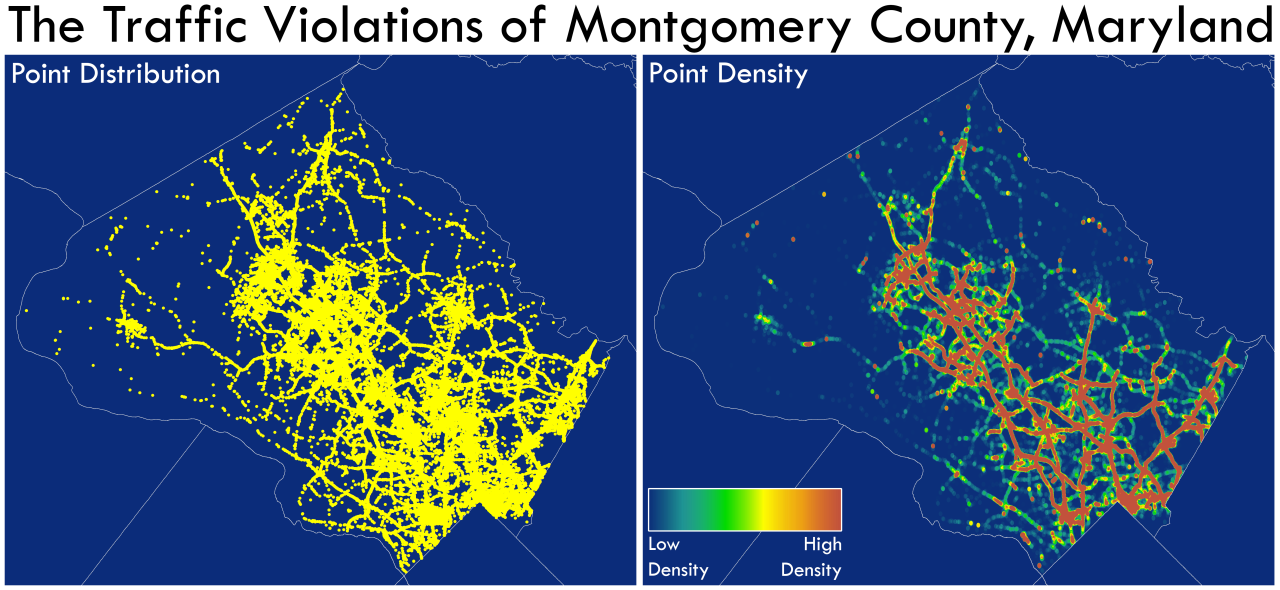
# Project Motivation/ Background

Safety of people is always a concern in every part of the world. The failure of people, equipment, supplies, or surroundings to behave or react as expected cause most of the violations. Violation investigations determine how and why these failures occur. This involves identification and establishing co relation among the situations initiating a traffic violation so that the risk and fatality on road can be reduced.

Another major issue/concern of people is traffic tickets. In the [United States](https://en.wikipedia.org/wiki/United_States), most traffic laws are codified in a variety of state, county and municipal laws or ordinances, with most minor violations classified as [infractions](https://en.wikipedia.org/wiki/Infraction), [civil charges](https://en.wikipedia.org/wiki/Civil_penalty) or [criminal charges](https://en.wikipedia.org/wiki/Criminal_charges). **Traffic tickets** are always an inconvenience, even when they are for **minor traffic violations**. For example, traffic citations for non-criminal offenses including speeding, running a stop sign or following too closely all carry fines and add points to your driving record.

By using the information gained through an investigation and our analysis we would try reducing the number of tickets for minor traffic violation as well.

**Visual Statistics:**



Montgomery County, Maryland, is the most populous county in the state, located adjacent to

Washington D.C. The above is a pair of maps showing point distribution and density. This highlights

major traffic routes, and it’s fun to see the D.C. border (bottom right) defined by a red mass.

# Data Set Description

The data set used in this analysis is a second-hand dataset obtained from data.gov.

These data of traffic violations in Montgomery country of Maryland, from the U.S. government, can be used to predict whether the traffic violation would contribute to an accident, contribute to property damage or personal injury. But for this project we are going to consider the influence of factors to see how they contribute to an accident. We also aim to identify indirect factors that can influence these violations using business intelligence tools and techniques.

The data set consist a total of 825297 records with 35 parameters. Below is the description of the parameters included in the data set.

Below is the description of the parameters included in the data:

|  |  |  |
| --- | --- | --- |
| **S.no** | **Target Variables** | **Values** |
| 1 | Contributed to Accident | If the traffic violation was a contributing factor in an accident. 1 = Yes, 0 = No |

|  |  |  |
| --- | --- | --- |
| **S.no** | **Predictors** | **Values** |
| 1 | Date\_Of\_Stop | Date of the traffic violation |
| 2 | Agency | Agency issuing the traffic violation. (Example: MCP is Montgomery County Police) |
| 3 | SubAgency | Court code representing the district of assignment of the officer. R15 = 1st district, Rockville B15 = 2nd district, Bethesda SS15 = 3rd district, Silver Spring WG15 = 4th district, Wheaton G15 = 5th district, Germantown M15 = 6th district, Gaithersburg / Montgomery Village HQ15 = Headquarters and Special Operations |
| 4 | Description | Text description of the specific charge. |
| 5 | Latitude | Latitude location of the traffic violation. |
| 6 | Longitude | Longitude location of the traffic violation. |
| 7 | Accident | If traffic violation involved an accident. 1 = Yes, 0 = No |
| 8 | Belts | If traffic violation involved a seat belt violation. 1 = Yes, 0 = No |
| 9 | Personal Injury | If traffic violation involved Personal Injury. 1 = Yes, 0 = No |
| 10 | Property Damage | If traffic violation involved Property Damage. 1 = Yes, 0 = No |
| 11 | Fatal | If traffic violation involved a fatality. 1 = Yes, 0 = No |
| 12 | Commercial License | If driver holds a Commercial Driver’s License. 1 = Yes, 0 = No |
| 13 | HAZMAT | If the traffic violation involved hazardous materials. 1 = Yes, 0 = No |
| 14 | Commercial Vehicle | If the vehicle committing the traffic violation is a commercial vehicle. 1= Yes, 0 = No |
| 15 | Alcohol | If the traffic violation included an alcohol related. 1 = Yes, 0 = No |
| 16 | Work Zone | If the traffic violation was in a work zone. 1 = Yes, 0 = No |
| 17 | State | State issuing the vehicle registration |
| 18 | VehicleType | Type of vehicle (Examples: Automobile, Station Wagon, Heavy Duty Truck, etc.) |
| 19 | Year | Year vehicle was made. |
| 20 | Make | Manufacturer of the vehicle (Examples: Ford, Chevy, Honda, Toyota, etc.) |
| 21 | Model | Model of the vehicle. |
| 22 | Color | Color of the vehicle. |
| 23 | Violation Type | Violation type. (Examples: Warning, Citation, SERO) |
| 24 | Charge | Numeric code for the specific charge. |
| 25 | Article | Article of State Law. (TA = Transportation Article, MR = Maryland Rules) |
| 26 | Hour\_Of\_Stop | Interval hour of the traffic violation |
| 27 | Race | Race of the driver. (Example: Asian, Black, White, Other, etc.) |
| 28 | Gender | Gender of the Driver (F=Female, M=Male) |
| 29 | Driver City | City of the driver’s home address |
| 30 | Driver State | State of the driver’s home address |
| 31 | DL State | State issuing the Driver’s license |
| 32 | Arrest Type | Type of arrest (A=Marked, B=Unmarked etc) |
| 33 | Geolocation | Geo-coded location information. |
| 34 | Location | Location of the violation, usually an address or intersection. |

Data Analysis

## BI Model:

Data

Reduction

Data Collection

Managerial

conclusion

Problem Definition

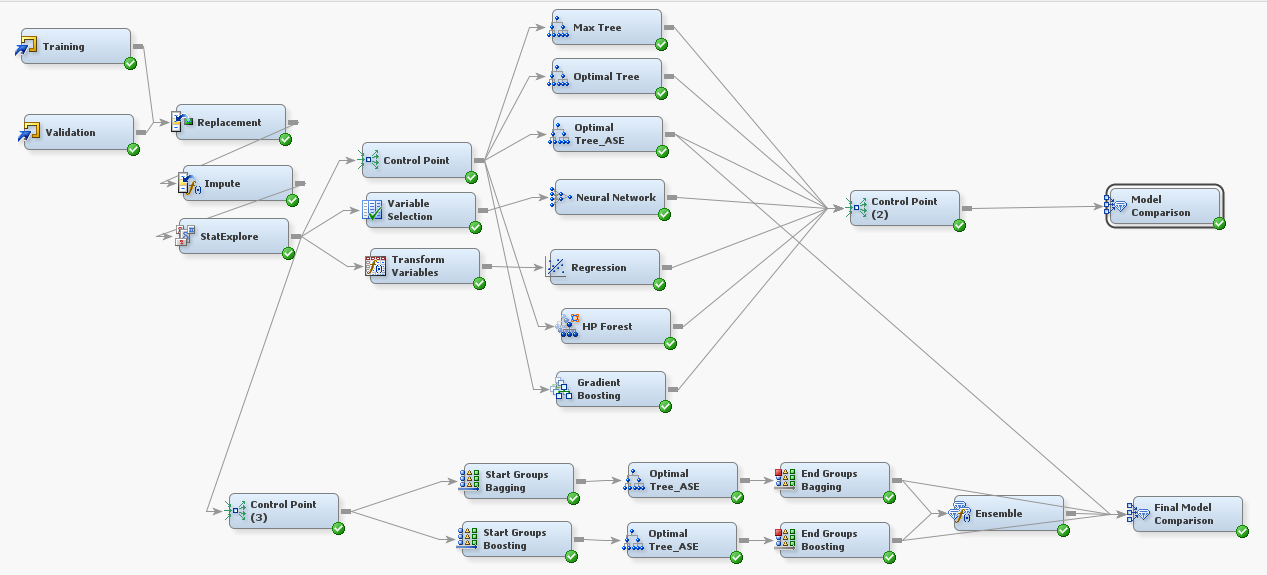
Data

Cleaning

Predictive Modelling

Descriptive Statistics

## Enterprise Miner Diagram:



## Cleaning and Preprocessing of Data

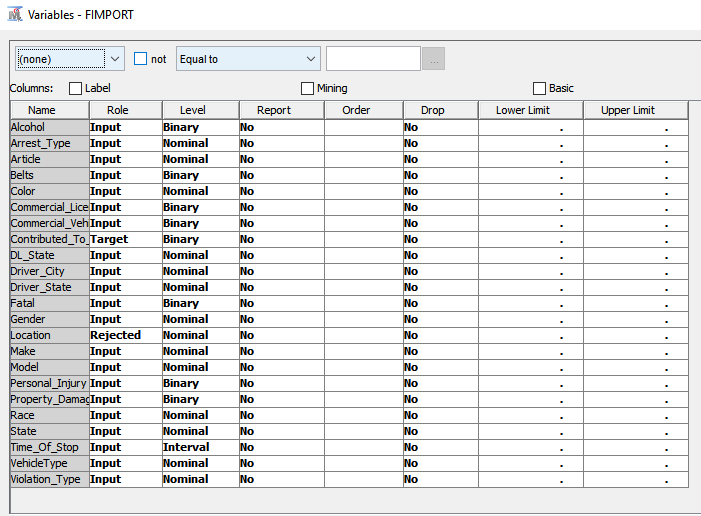
Our dataset is majorly categorical with the target variable values being ‘yes’ or ‘no’. To build a proper model, the observations with target ‘no’ were under-sampled and observations with target ‘yes’ were oversampled in the train data. Certain columns had no variability at all in their values (all values were same throughout the column). Since these columns were uninformative of the target variable, they were removed from the dataset to avoid unnecessary inflation or misclassification in the statistics.

Our original dataset had 80000 observations from which we had to do a drastic data reduction. The train dataset was sampled as mentioned above.

Validation data set was randomly sampled, without any modification, from the original dataset of 800000 records.

The data is imported into SAS Enterprise Miner using ‘File Import’ Node. Below we can see the default role and data type assigned by SAS Enterprise

Miner.



## Target selection and rejecting insignificant variables

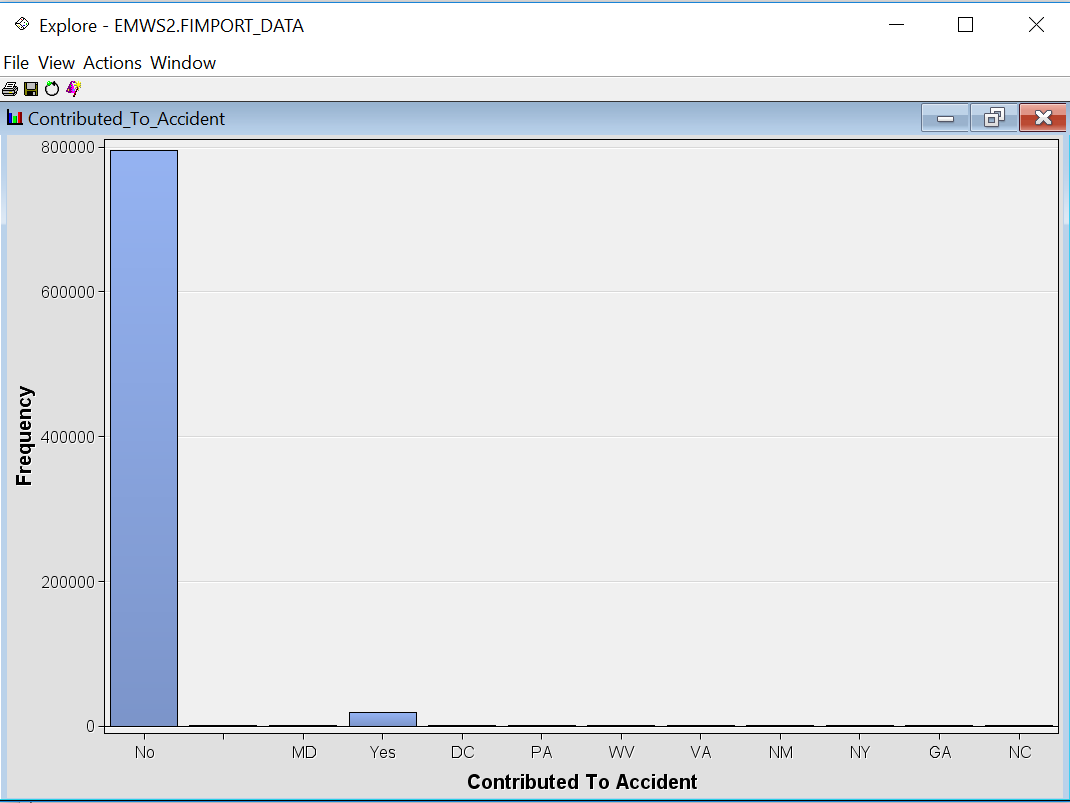
The main purpose of our project is to understand which all factors will contribute to accident, hence we have chosen the following as our dependent variable

* Contributed to Accident

We also found few variables do not have any significance/contribution in our finding/analysis. Hence, rejected the below variable:

* Location

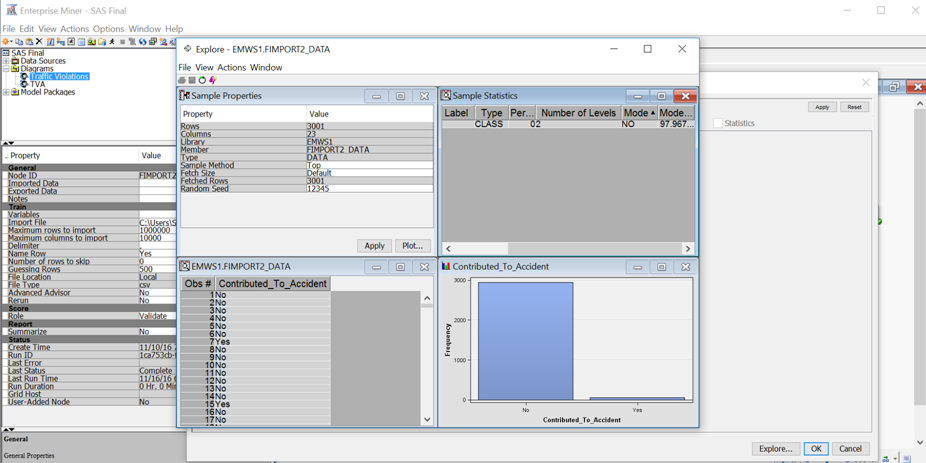
## Remarks and Observation on variables



**Replacement:**

We have extra values as an output for the target variable, where as the expected output is only ‘Yes’

or ‘No’. So we imported the dataset into SAS Base and saved it without impurities.



* Few variables in the data set were not defined properly, for which we used the replacement node.
* We replaced

‘Ye’ to Yes

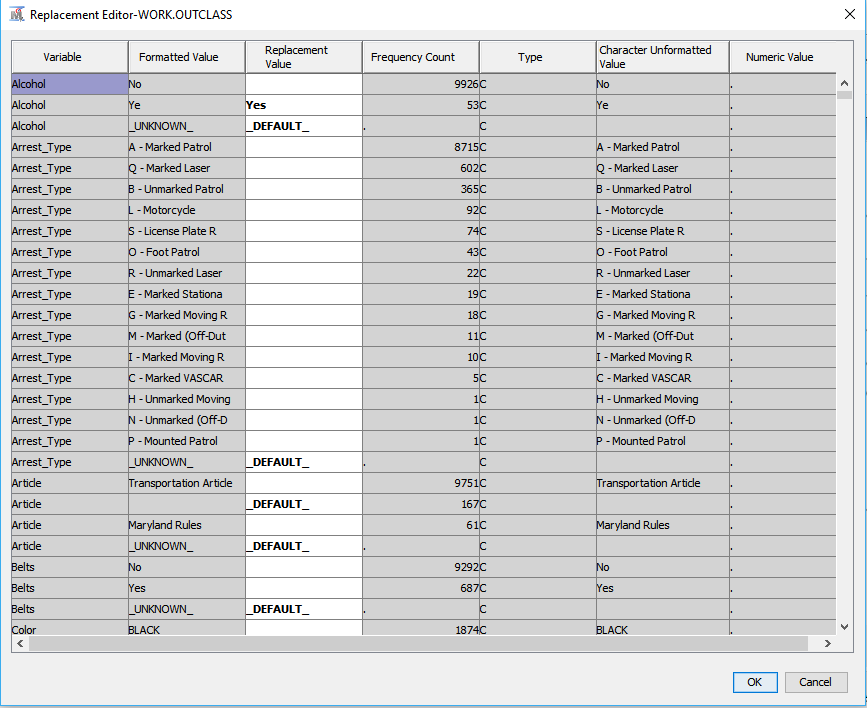
\_UNKOWN\_ to \_DEFAULT\_

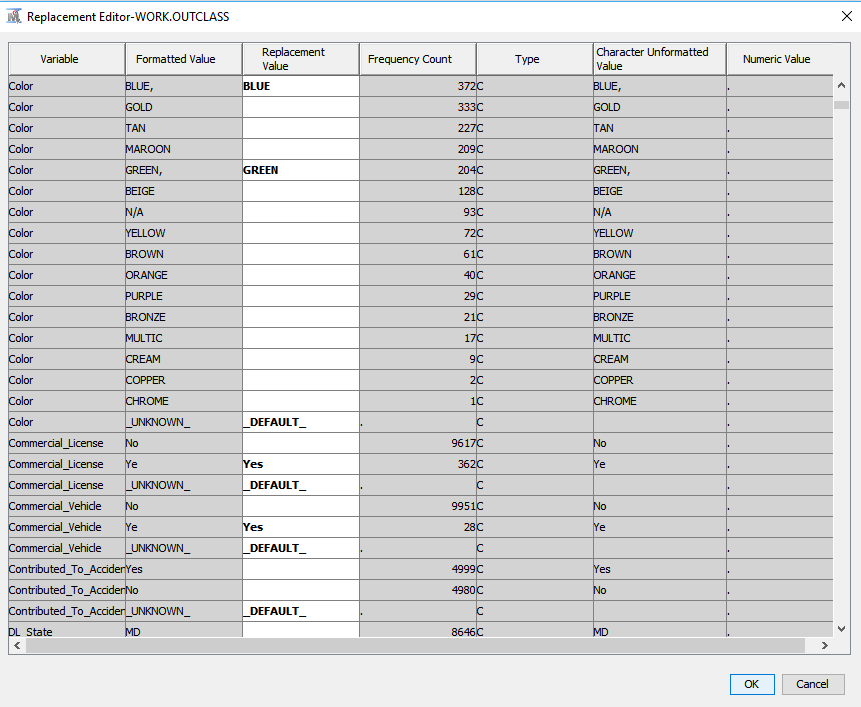
Green, to Green

Blue, to Blue

1. **Replacement Node**: It is used to modify incorrect or improper values for a variable.

Few variables which were replaced are shown below:





# **Impute Node:**

# 

On exploration, it was found that some variables had missing values. We use the Impute node to populate the missing values. Below are the input methods for class and interval variables’ imputation:

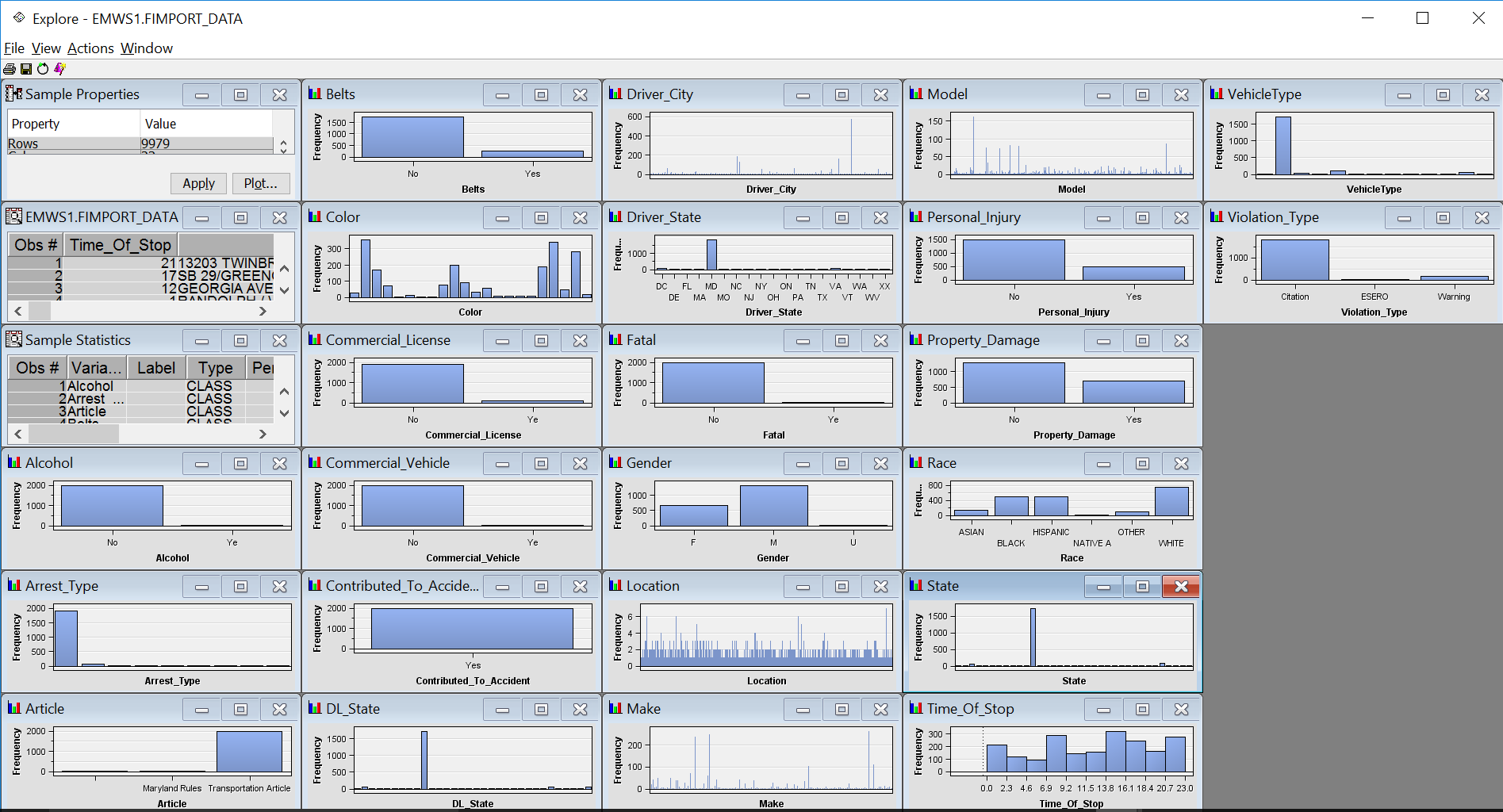
Class Variables: Default Input Method – Count

Interval Variables: Default Input Method – Mean

# 

## Distribution of Input Variable

The figure below shows the distribution of input variables considered for analysis.



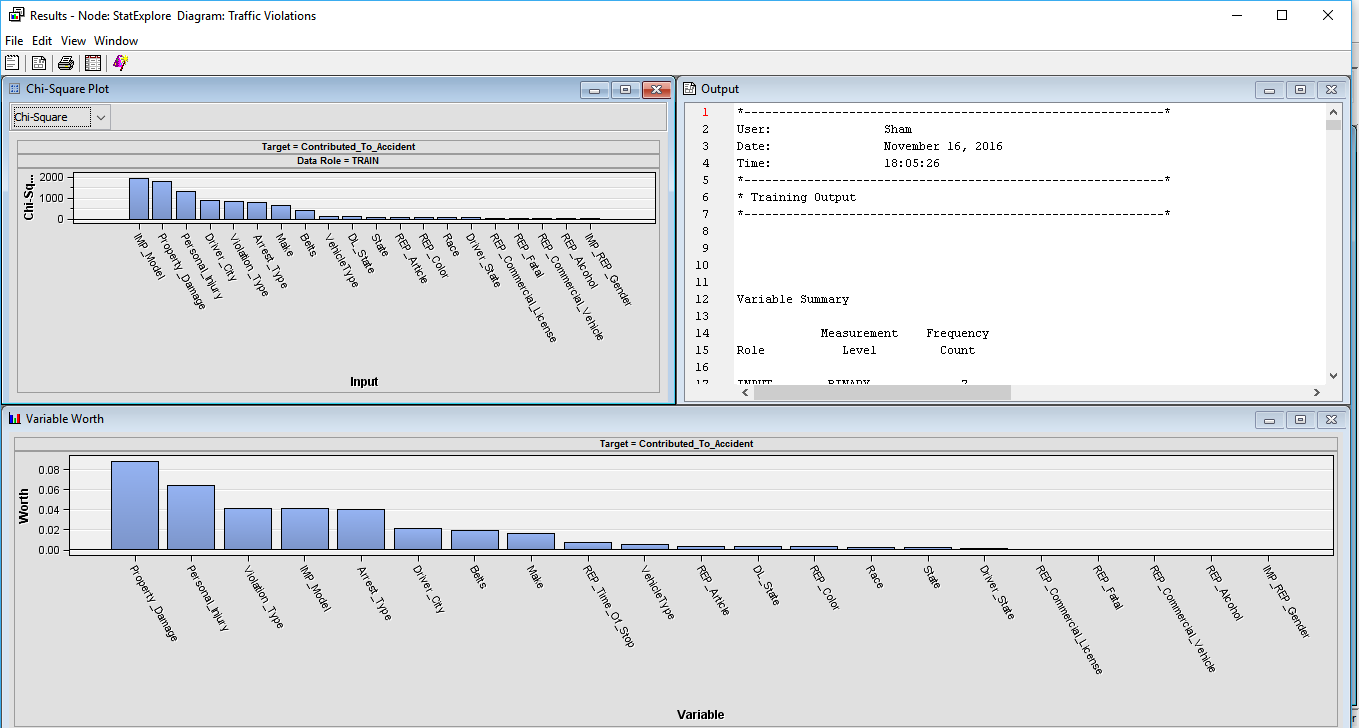
# Data Modeling using SAS Enterprise Miner 9.4

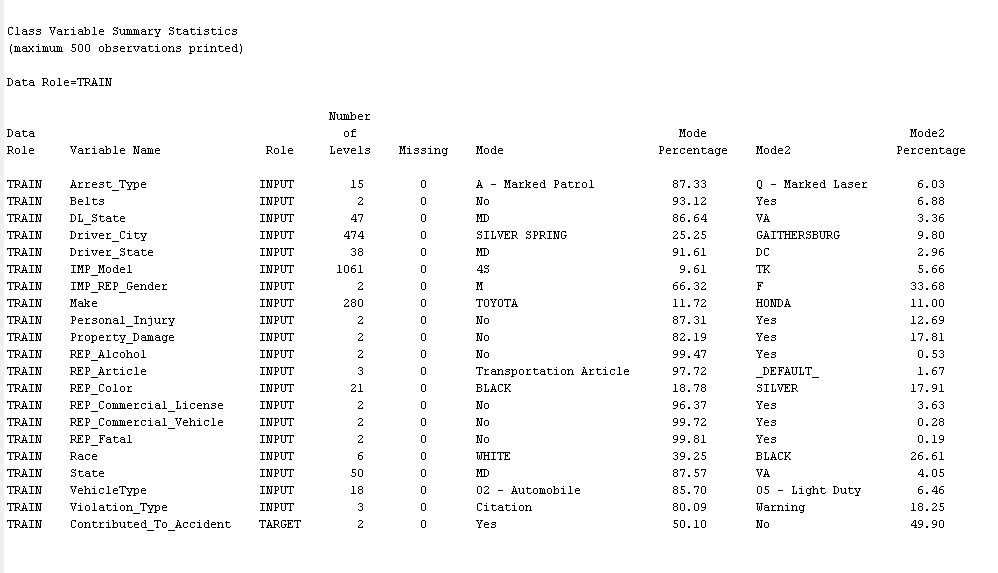
The main objectives of our project are:

* Explore and analyze the perspective input parameters that impacts the severity of traffic violations.
* Explore different data mining techniques to predict the reason of violations and property damages. To uncover the major relationship and patterns between different input variables and the target that are entirely categorical and nominal (absolutely no interval variables), we are applying the following modeling techniques:

## Summary Statistics using StatExplore

We have used the StatExplore node in SAS Enterprise Miner to produce analyze the statistical summary of our data. The summary statistics of the input variables used is shown below:





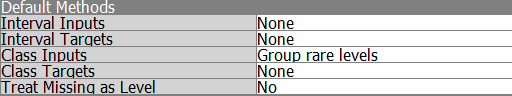
From the above StatExplore result we can confirm we don’t have any missing values in the independent input variables used.

## Transform Variables

It is known that the normally distributed input variables give accurate results. To avoid the distribution effect on the output, the variables that are extremely skewed are transformed. So, the transform variable node is used before logistic regression.

Sometimes, input data is more informative on a scale other than that from which it was originally collected. Variable transformations can be used to stabilize variance, remove nonlinearity, improve additivity, and counter non-normality. Therefore, for many models, transformations of the input data (either dependent or independent variables) can lead to a better model fit. These transformations can be functions of either a single variable or of more than one variable. To use the Transform Variables node is to make variables better suited for logistic regression models.

Group Rare Level transformation method is meant for the transformation of class variables.



Variable selection

If there are too many predictors, it is recommended that we do the variable selection before applying the actual model. Since the dataset is majorly categorical, in variable selection property window, the TARGET MODEL parameter is set to be chi-square. We are using it specifically for the neural network since this model is extremely flexible and bound with the variables.

5.3 Decision Tree

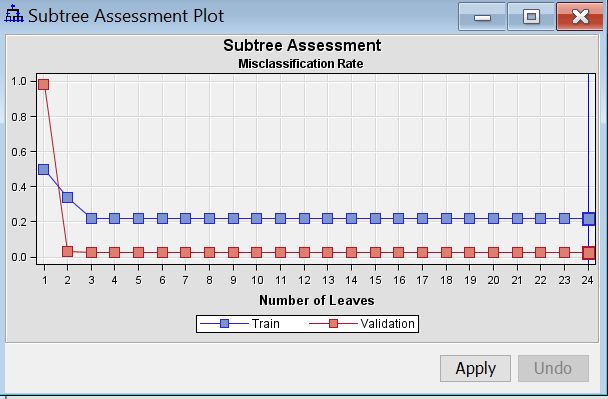
Decision tree provides an excellent introduction to predictive modeling. These models are conceptually easy to understand and they readily accommodate nonlinear associations between input variables and one or more target variables.

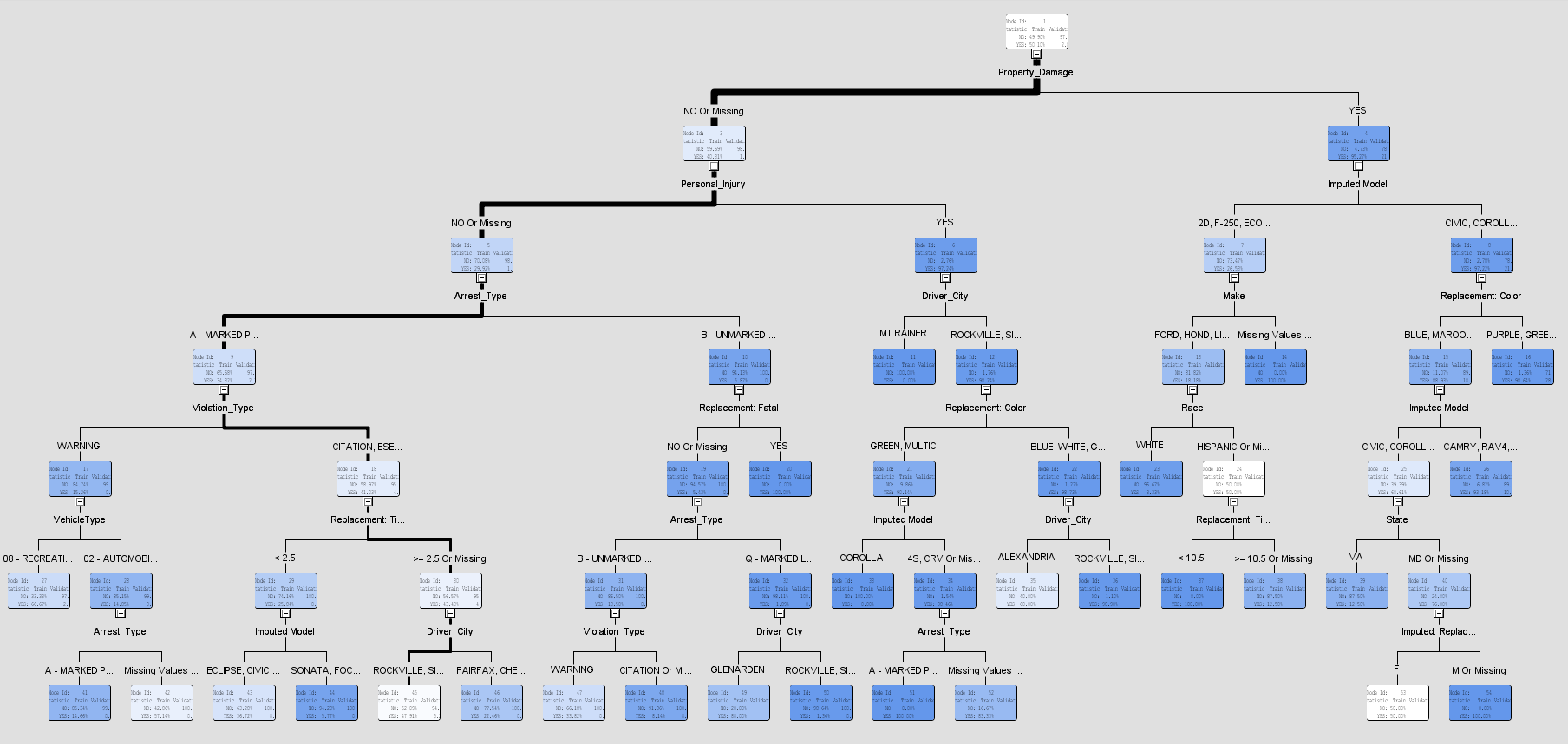
**Max Tree:**

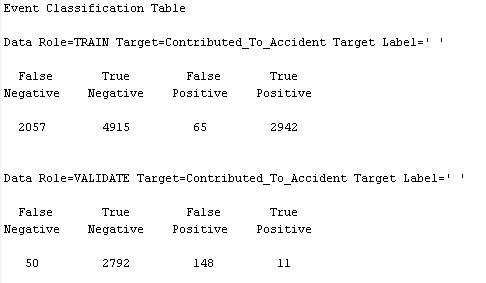
Here, the root node is trained to give the maximum classification tree.

Right click on root node -> Train node

**Subtree Assessment plot of the MAX TREE:**

 No. of. leaves = 24



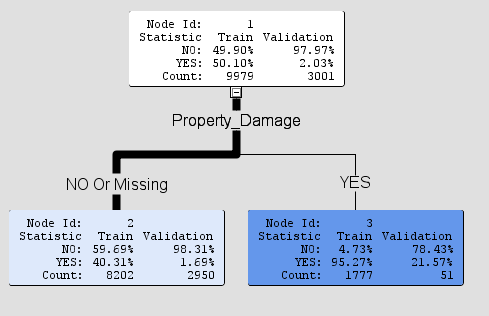


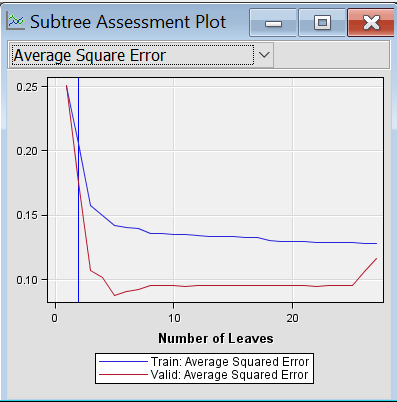
The model is good because the False Negative and False Positive counts are very less when compared to the True Negative and True positive counts respectively.

**Optimal Tree:**

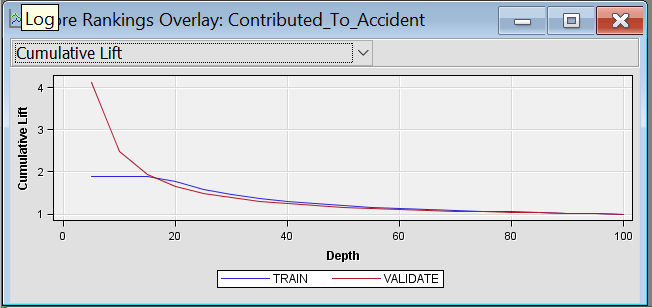
The optimal tree is grown with the ASSESSMENT MEASURE parameter set to “Decision”.

The root node indicates that 50.10 % chances that the accident takes place and 49.90% chances that the accident will not occur.





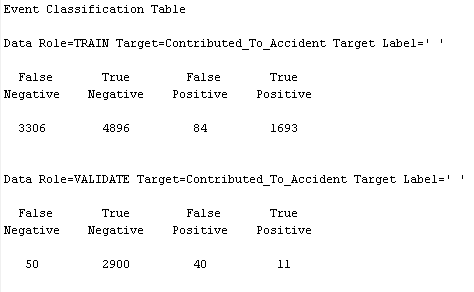
Sub- tree assessment plot of optimal tree indicates that the tree has 3 leaves

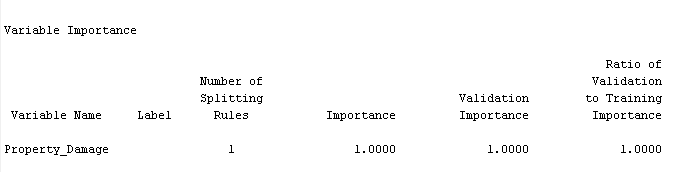


The cumulative lift is above the NO-MODAL-base-lift graph which indicates that the model is very good.

**Confusion Matrix:**

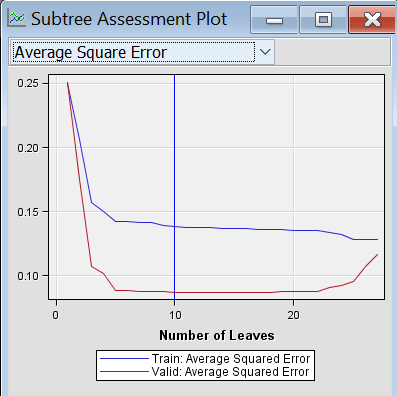
The model is good because the False Negative and False Positive counts are very less when compared to the True Negative and True positive counts respectively.



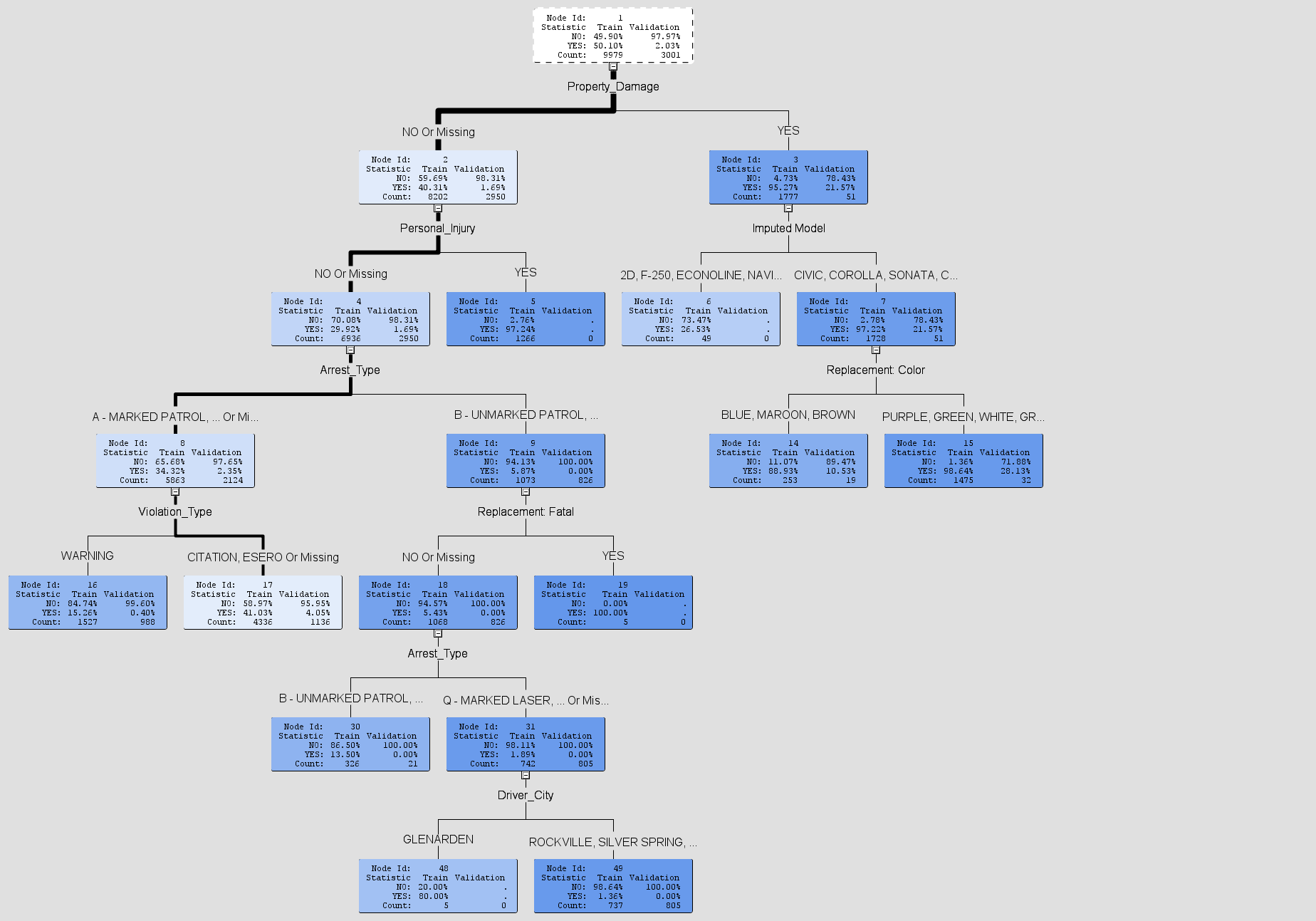


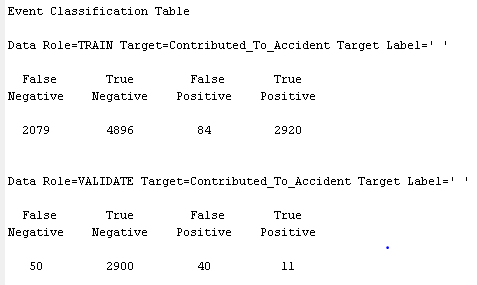
The above represents the variable importance of our data set.

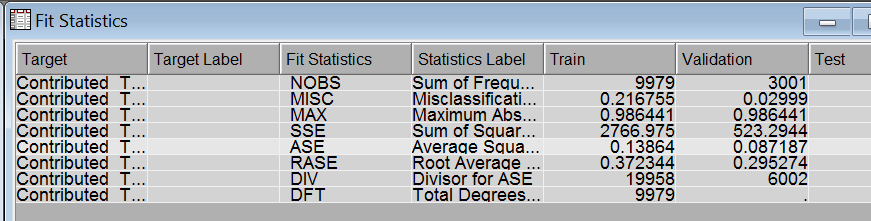
**Optimal Tree ASE:**



No. of. Leaves in the optimal tree ASE = 10

**Tree Interpretation:** If there is no property damage, no personal injury, arrest type is marked patrol and violation type is “citation or esero or missing”, then there is 95.95% chance that the it will not be an accident and 4.05 % chance that it will be an accident.****



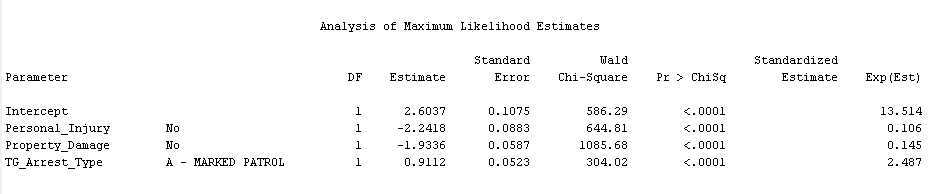


We can notice from the above Fit Statistics that:

* The Misclassification Rate for Train data is 0.216755 and Validation data is 0.02999
* The Average Squared error for Train data is 0.13864 and Validation data: 0.087187

## Logistic Regression

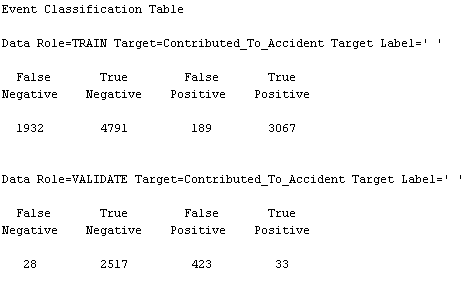
Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by estimating probabilities. Model outcome can be viewed as primary outcome probabilities.



The above variables have p-value <0.0001 which means that they are highly significant carrying a lot of information about the target variable.

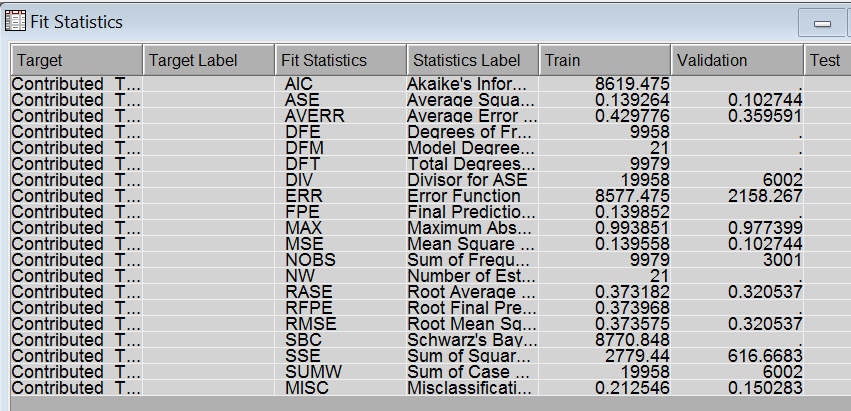
Beta value (Estimate column above): Negative beta value indicates that increase in the variable listed above decreases the odds ratio by a factor of this estimate. Positive beta value indicates that increase in the variable listed above increases the odds ratio by a factor of this estimate.

**Confusion Matrix:**



From above confusion matrix, Sum of True Negative and True Positive is much higher than False Positive and False Negative.

Based on the above result from event classification table of the decision tree model, we analyse that the model is almost accurate.

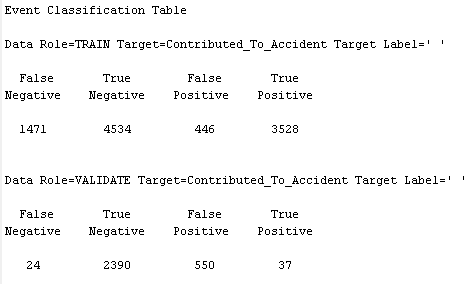


We can notice from the above Fit Statistics that:

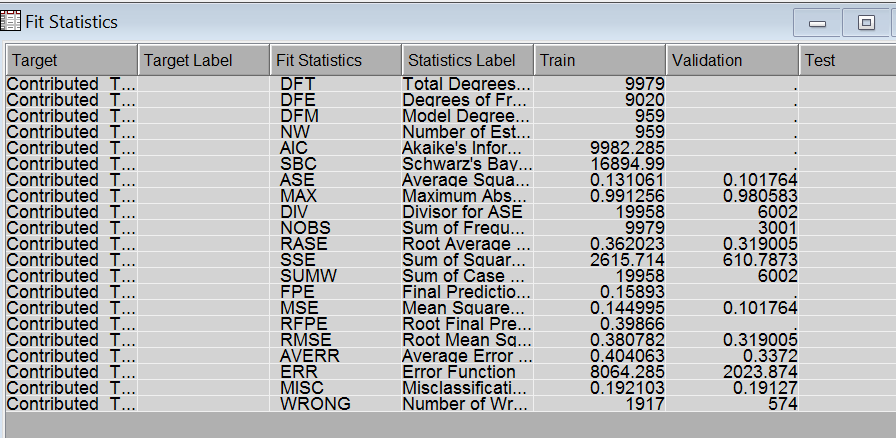
* The Misclassification Rate for Train data is 0.212546 and Validation data is 0.150283
* The Average Squared error for Train data is 0.139264 and Validation data: 0.102744

## Neural Network

Neural networks are a class of parametric models that can accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable than can logistic regression. Building a neural network model involves two main phases. First, you must define the network configuration. You can think of this step as defining the structure of the model that you want to use. Then, you iteratively train the model. SAS Enterprise Miner has two nodes that fit neural network model: Neural Network node and the Auto-Neural node.

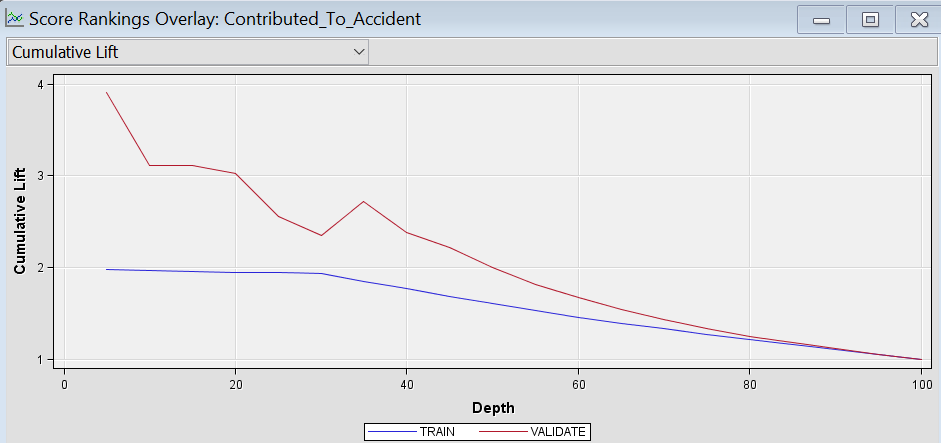


Based on the above result from event classification table of the model, we see that the model is almost good in classification



We can notice from the above Fit Statistics that:

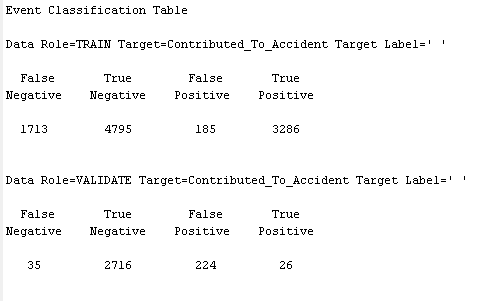
* The Misclassification Rate for Train data is 0.192103 and Validation data is 0.19127
* The Average Squared error for Train data is 0.131061 and Validation data: 0.101764



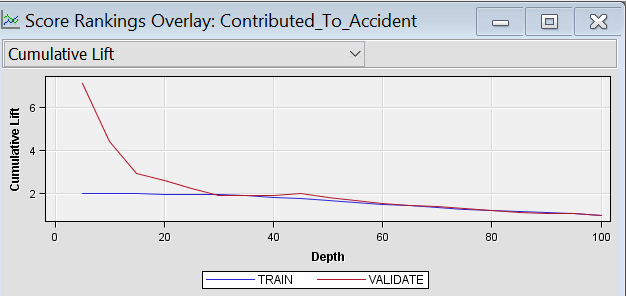
The cumulative lift is above the NO-MODAL-base-lift graph which indicates that the model is very good.

## HP Forest:

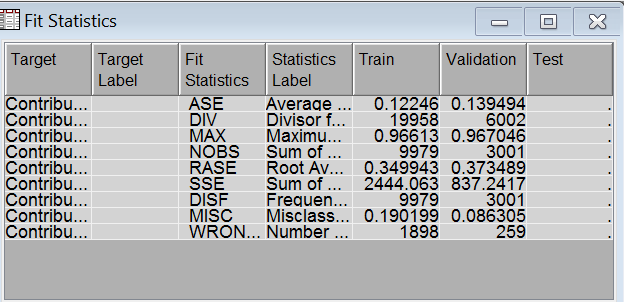
HP Forest is one of the best predictive models for extremely large nominal dataset. It creates several tree forests using random ensemble methodology and combines the predictions by voting for the target variables.



Based on the above result from event classification table of the model, we see that the model is almost good in classification. The false positive and false negative is lesser than the true positive and true negative.



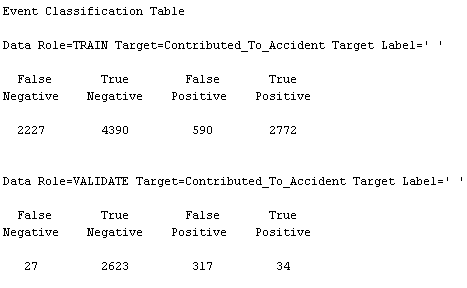
The cumulative lift values are high above base line.



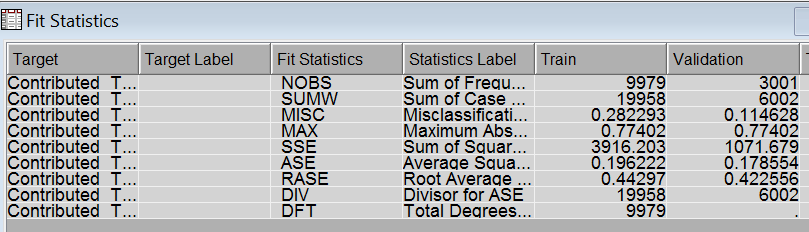
The low misclassification rate and the low average square error showcase that this model is a good model for this categorical dataset.

## Gradient Boosting

Gradient boosting works on iterative sampling and modelling for the predicted target. In each iteration, the data to be trained is the data correctly classified from the previous iteration.



From above confusion matrix, Sum of True Negative and True Positive is much higher than False Positive and False Negative.

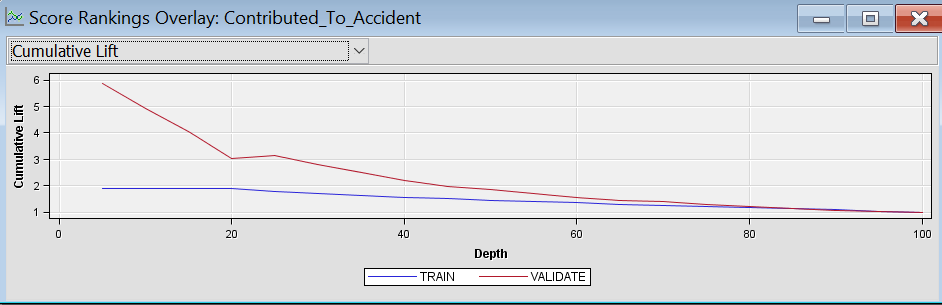


Misclassification rate:

Train: 0.282293, Validation: 0.114628

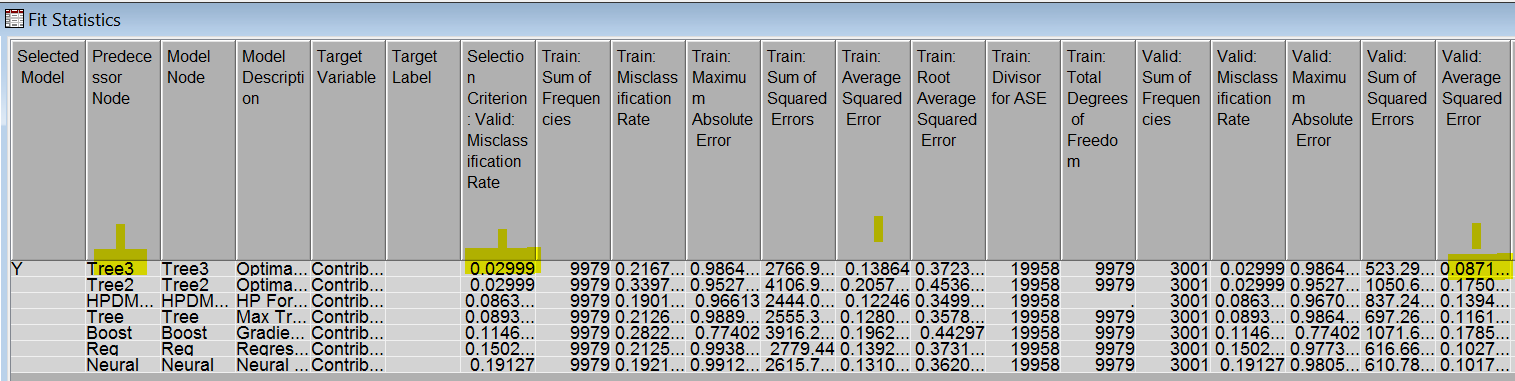
Avg. Square Error:

Train: 0.196222, Validation: 0.178554

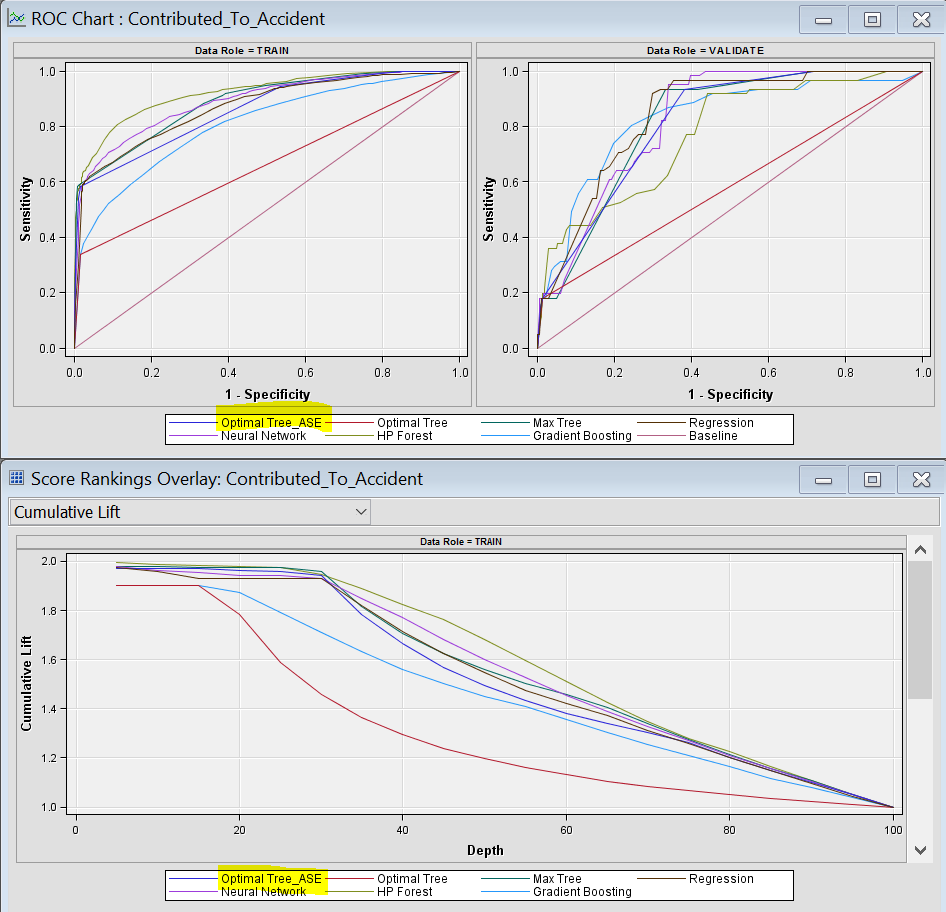


Good Cumulative lift

## Model Comparison



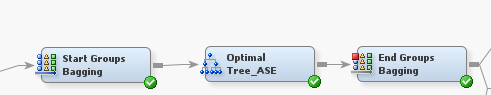
The Optimal ASE decision tree model of all the predictive models seems to be predicting better with low average square error and low misclassification rate. In the graph below, ROC curve of the Optimal ASE decision tree model is high above the baseline. Even the cumulative lift of the Optimal ASE decision tree model is very high above the NO-MODAL baseline. So, we can very well choose this Optimal ASE decision tree model as the best model in predicting the traffic violation accidents.



## 6. Performance Tuning of The Best- Selected Model:

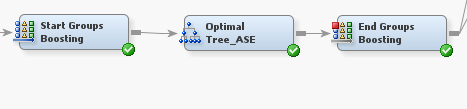
## Bagging:

After identifying the optimal ASE tree to be the best model among the trained simple and ensemble models, the performance of the model was improved by applying the bagging algorithm. The multiple sampling in this algorithm creates several predicted probabilities and concludes with the weighted average results. The start group and the end group nodes are used with the optimal ASE tree.



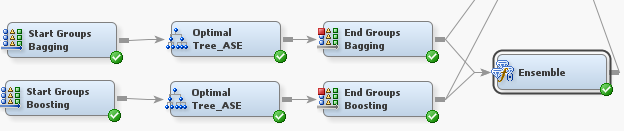
## Boosting:

Like bagging, the boosting uses multiple sampling. But here, the samples are iteratively taken and on each iteration, the misclassification rate is recognized and corrected on its own. The same start group and end groups are used for the previously proven best tree, the optimal ASE tree.



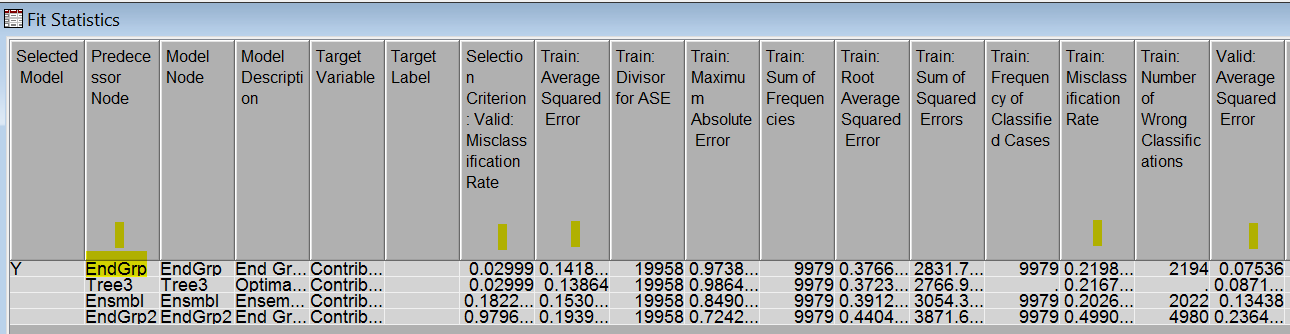
Ensemble:

While bagging and boosting improves the performance of the model, the idea of building a hybrid ensemble model of both bagging and boosting models can drastically give excellent predictive model with least misclassification rate and least average square error.



Final Model Comparison:

A comparison of the simple optimal tree ASE, bagged optimal ASE tree, boosted optimal ASE tree and the ensemble of bag & boost trees rightly showcases that the bagged optimal tree ASE tree is the best performing model with least validation misclassification rate and least average squared error.



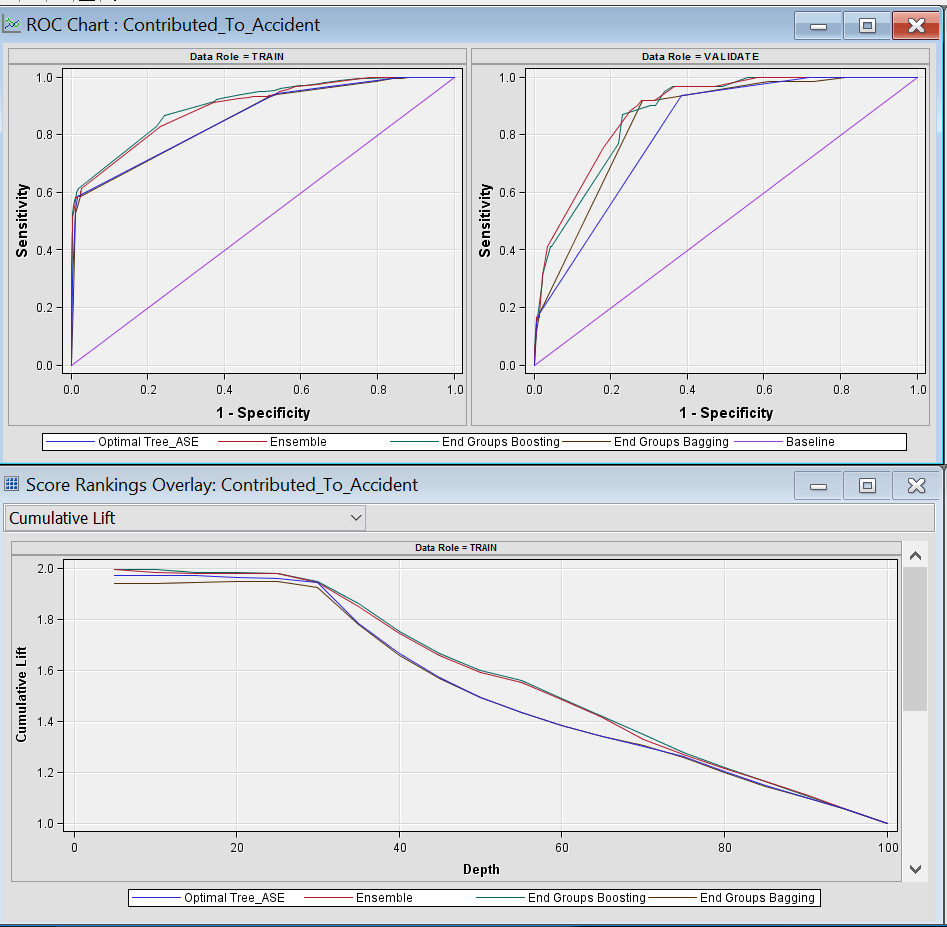
Bagging End Group:

Validation Misclassification rate = 0.02999

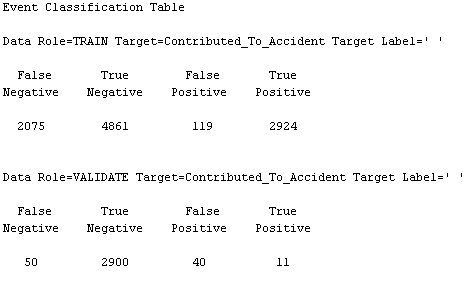
Validation Average Squared Error = 0.07536

The ROC chart between the specificity and the sensitivity clearly states that the bagged optimal tree is very good with the ROC curve well above the baseline.

The cumulative lift plot also clearly states that the bagged optimal tree is very good model with very high cumulative lift , well above the NO-MODEL baseline.



Selected Model: Bagging of the optimal tree



Confusion Matrix Terms:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  |  | "+ve" | "-ve" |
| Actual | "+ve" | a | b |
| "-ve" | c | d |
|  |  |  |  |
|  |  | Predicted | |
|  |  | Accident = yes | Accident = no |
| Actual | Accident = yes | 11 | 50 |
| Accident = no | 40 | 2900 |

**Accuracy:** Proportion of correct predictions

Accuracy = (a+d) / (a+b+c+d)

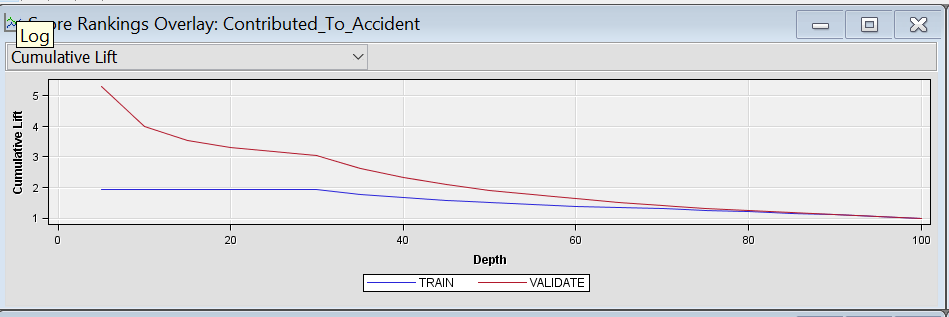
= (11+2900) / (11+50+40+2900)

= 0.970009

**= 97%**

Thus, the selected model not only out performs all other models, but also has very accuracy of 97 % in predicting the contribution to accident.

The Cumulative lift curve is well above the NO-Model baseline with high values of cumulative lift. Hence this model is an excellent model.



## 6. Managerial implications/conclusions

We conclude with the following interesting findings from the model. Also, these findings can be used as a leverage to counter traffic violations which contribute to an accident.

-Predicting the occurrence of an accident and enforcing the preventive measures is a dramatic feet of statistical application

- 18.03% of Accidents occur when Car Model is CIVIC, COROLLA, SONATA, CAMRY and will also result in property damage

- 72% of Accidents occur when the driver did not wear the seat belt, is driving a car of Mitsubishi, Chrysler company and does not have Commercial License

- On a given day, accidents occur almost one and half times more post 4 pm than before it

## 7. References

1.<https://support.sas.com/resources/papers/proceedings14/SAS133-2014.pdf>

2.https://www.sas.com/content/dam/SAS/en\_gb/doc/other1/events/sasforum/slides/day2/I.Brown%20Advanced%20Modelling%20Techniques%20in%20SAS%20EM\_IB.pdf

3. Data Mining for Business Intelligence – Galit Shmeuli

4. MIS 6324 Class Slides