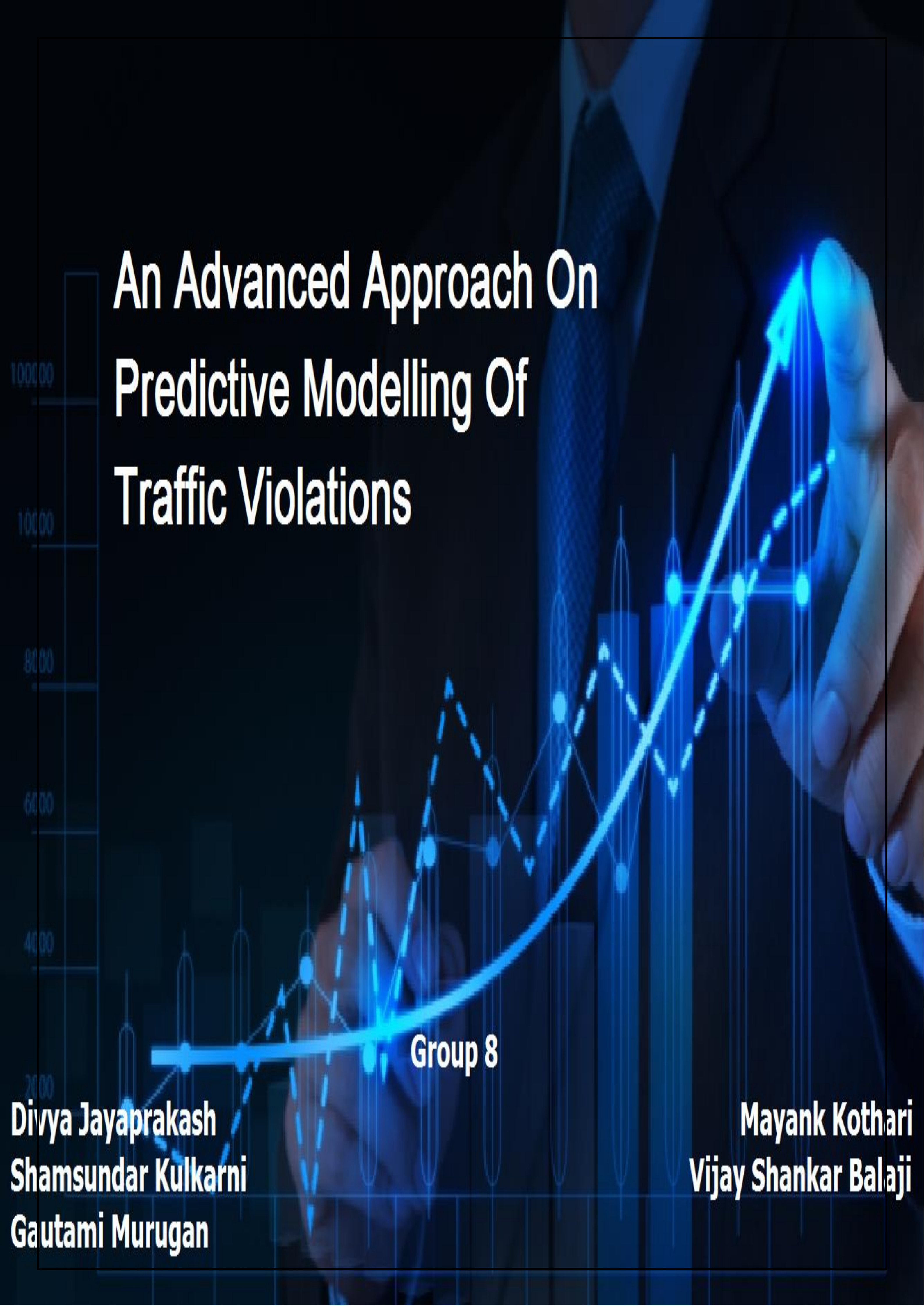


An Advanced Approach On Predictive Modelling Of Traffic Violations

Group 8

Divya Jayaprakash
Shamsundar Kulkarni
Gautami Murugan

Mayank Kothari
Vijay Shankar Balaji

1 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.

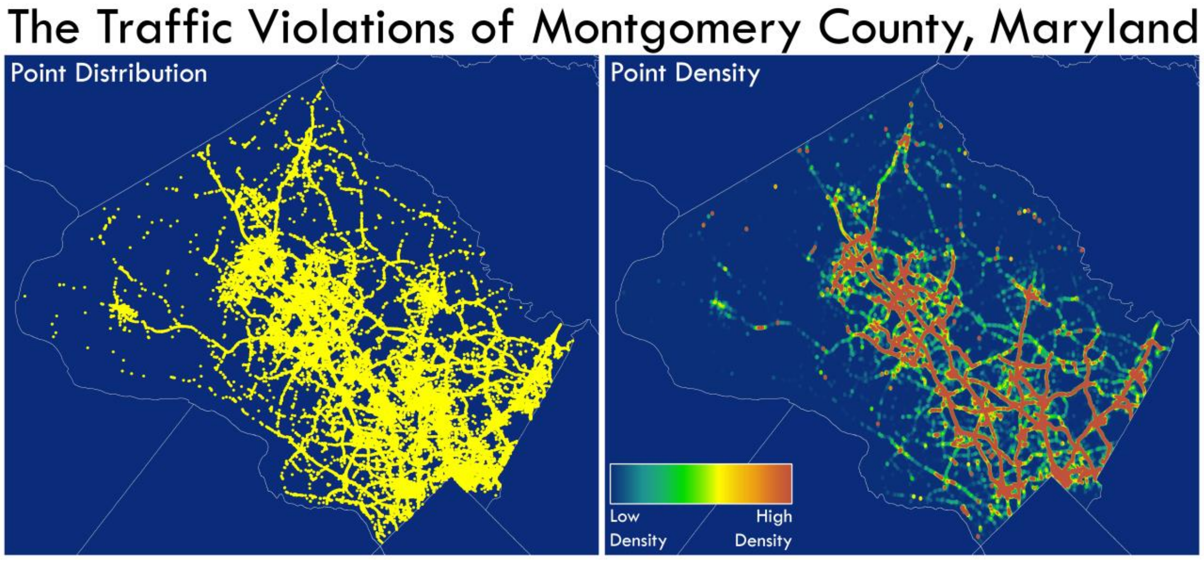
2 Project Motivation/ Background

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, most traffic laws are codified in a variety of state, county and municipal laws or ordinances, with most minor violations classified as infractions, civil charges or 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.

3 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.

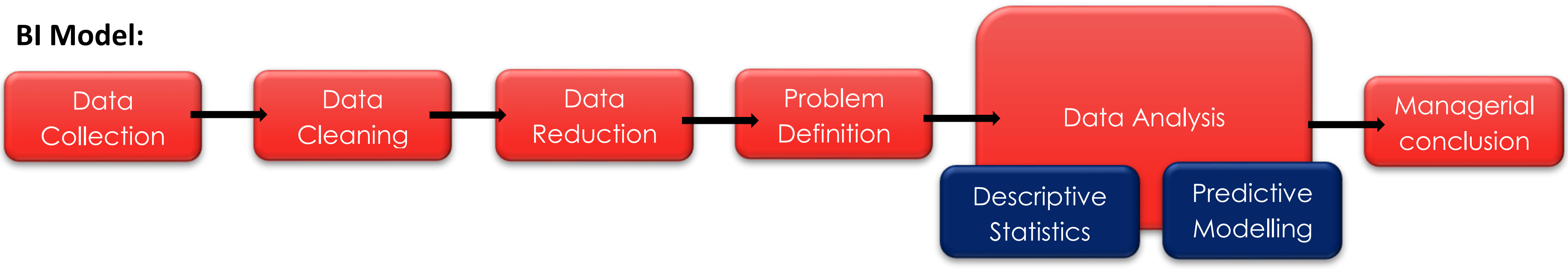
Below is the description of the parameters included in the data:

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

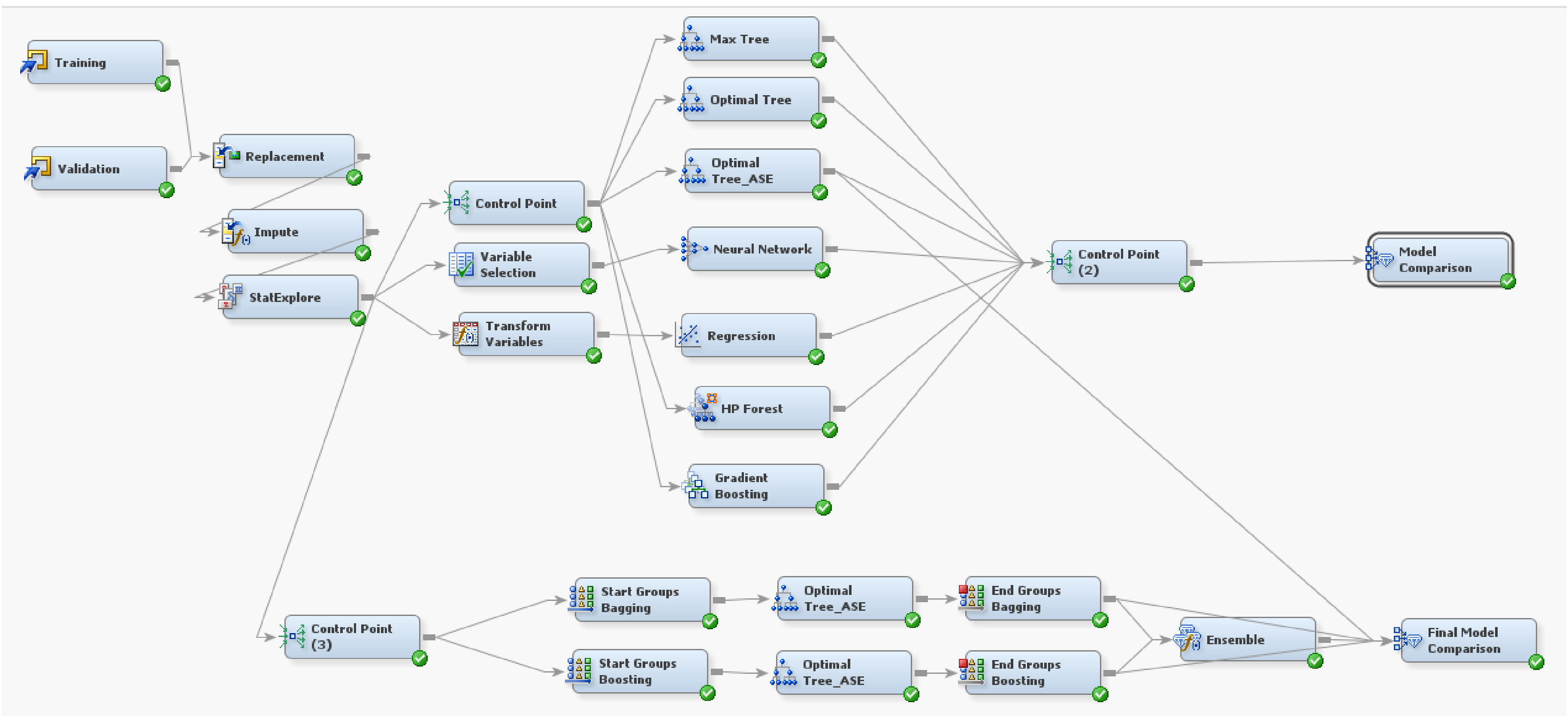
S.no	Predictors	Values
1	Date_Of_Stop	Date of the traffic violation
2	Agency	Agency issuing the traffic violation. (Example: MCP is Montgomery County Police)
3	SubAgency	Court code representing the district of assignment of the officer. R15 = 1st district, Rockville B15 = 2nd district, Bethesda SS15 = 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.

BI Model:



Enterprise Miner Diagram:



4 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.

Variables - FIMPORT

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Alcohol	Input	Binary	No		No	.	.
Arrest_Type	Input	Nominal	No		No	.	.
Article	Input	Nominal	No		No	.	.
Belts	Input	Binary	No		No	.	.
Color	Input	Nominal	No		No	.	.
Commercial_Lice	Input	Binary	No		No	.	.
Commercial_Veh	Input	Binary	No		No	.	.
Contributed_To_	Target	Binary	No		No	.	.
DL_State	Input	Nominal	No		No	.	.
Driver_City	Input	Nominal	No		No	.	.
Driver_State	Input	Nominal	No		No	.	.
Fatal	Input	Binary	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Location	Rejected	Nominal	No		No	.	.
Make	Input	Nominal	No		No	.	.
Model	Input	Nominal	No		No	.	.
Personal_Injury	Input	Binary	No		No	.	.
Property_Damag	Input	Binary	No		No	.	.
Race	Input	Nominal	No		No	.	.
State	Input	Nominal	No		No	.	.
Time_Of_Stop	Input	Interval	No		No	.	.
VehicleType	Input	Nominal	No		No	.	.
Violation_Type	Input	Nominal	No		No	.	.

4.1 Target selection and rejecting insignificant variables

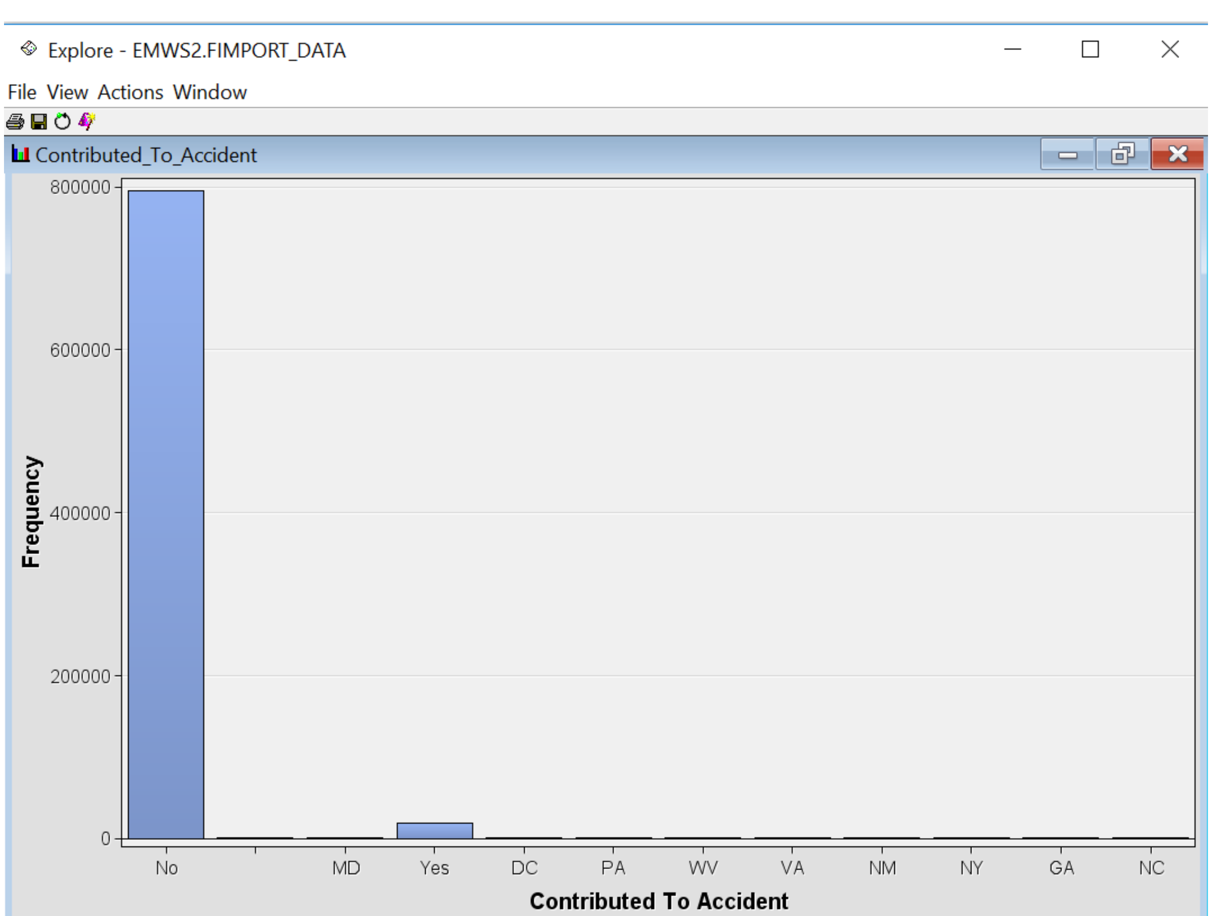
The main purpose of our project is to understand which all factors will contribute to accident, hence we have chosen the following as our dependent variable

- Contributed to Accident

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

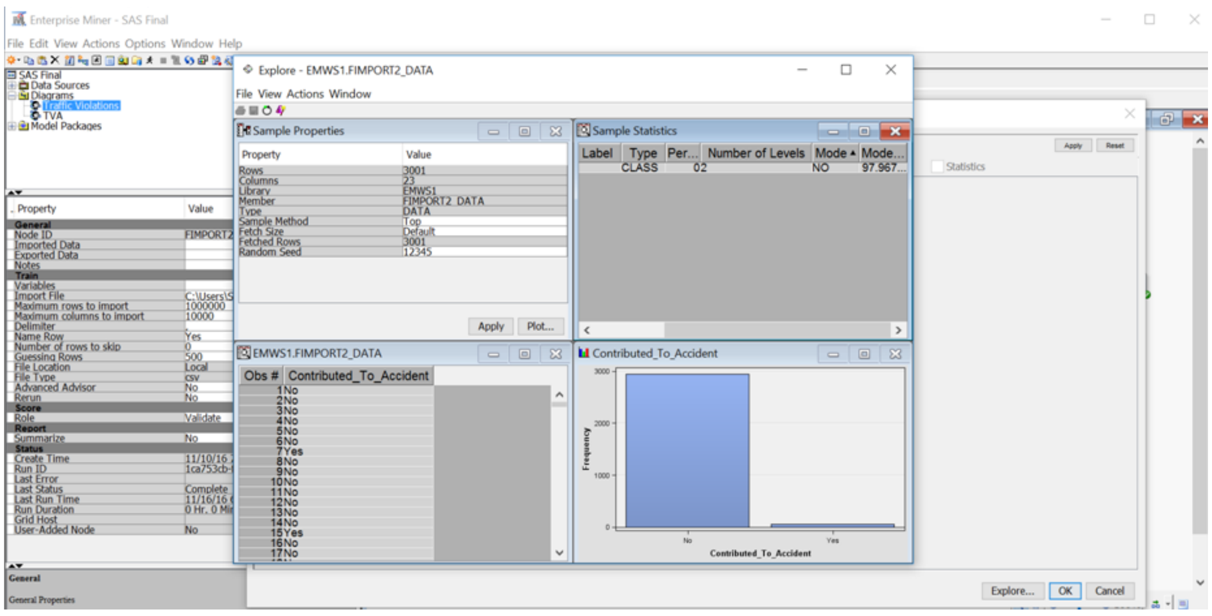
- Location

4.2 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

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

Few variables which were replaced are shown below:

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Alcohol	No		9926	C	No	.
Alcohol	Ye	Yes	53	C	Ye	.
Alcohol	_UNKNOWN_	_DEFAULT_	.	C		.
Arrest_Type	A - Marked Patrol		8715	C	A - Marked Patrol	.
Arrest_Type	Q - Marked Laser		602	C	Q - Marked Laser	.
Arrest_Type	B - Unmarked Patrol		365	C	B - Unmarked Patrol	.
Arrest_Type	L - Motorcycle		92	C	L - Motorcycle	.
Arrest_Type	S - License Plate R		74	C	S - License Plate R	.
Arrest_Type	O - Foot Patrol		43	C	O - Foot Patrol	.
Arrest_Type	R - Unmarked Laser		22	C	R - Unmarked Laser	.
Arrest_Type	E - Marked Stationa		19	C	E - Marked Stationa	.
Arrest_Type	G - Marked Moving R		18	C	G - Marked Moving R	.
Arrest_Type	M - Marked (Off-Dut		11	C	M - Marked (Off-Dut	.
Arrest_Type	I - Marked Moving R		10	C	I - Marked Moving R	.
Arrest_Type	C - Marked VASCAR		5	C	C - Marked VASCAR	.
Arrest_Type	H - Unmarked Moving		1	C	H - Unmarked Moving	.
Arrest_Type	N - Unmarked (Off-D		1	C	N - Unmarked (Off-D	.
Arrest_Type	P - Mounted Patrol		1	C	P - Mounted Patrol	.
Arrest_Type	_UNKNOWN_	_DEFAULT_	.	C		.
Article	Transportation Article		9751	C	Transportation Article	.
Article		_DEFAULT_	167	C		.
Article	Maryland Rules		61	C	Maryland Rules	.
Article	_UNKNOWN_	_DEFAULT_	.	C		.
Belts	No		9292	C	No	.
Belts	Yes		687	C	Yes	.
Belts	_UNKNOWN_	_DEFAULT_	.	C		.
Color	BLACK		1874	C	BLACK	.

OK Cancel

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Color	BLUE,	BLUE	372	C	BLUE,	.
Color	GOLD		333	C	GOLD	.
Color	TAN		227	C	TAN	.
Color	MAROON		209	C	MAROON	.
Color	GREEN,	GREEN	204	C	GREEN,	.
Color	BEIGE		128	C	BEIGE	.
Color	N/A		93	C	N/A	.
Color	YELLOW		72	C	YELLOW	.
Color	BROWN		61	C	BROWN	.
Color	ORANGE		40	C	ORANGE	.
Color	PURPLE		29	C	PURPLE	.
Color	BRONZE		21	C	BRONZE	.
Color	MULTIC		17	C	MULTIC	.
Color	CREAM		9	C	CREAM	.
Color	COPPER		2	C	COPPER	.
Color	CHROME		1	C	CHROME	.
Color	_UNKNOWN_	_DEFAULT_	.	C		.
Commercial_License	No		9617	C	No	.
Commercial_License	Ye	Yes	362	C	Ye	.
Commercial_License	_UNKNOWN_	_DEFAULT_	.	C		.
Commercial_Vehicle	No		9951	C	No	.
Commercial_Vehicle	Ye	Yes	28	C	Ye	.
Commercial_Vehicle	_UNKNOWN_	_DEFAULT_	.	C		.
Contributed_To_Accider	Yes		4999	C	Yes	.
Contributed_To_Accider	No		4980	C	No	.
Contributed_To_Accider	_UNKNOWN_	_DEFAULT_	.	C		.
DL_State	MD		8646	C	MD	.

OK Cancel

b) Impute Node:

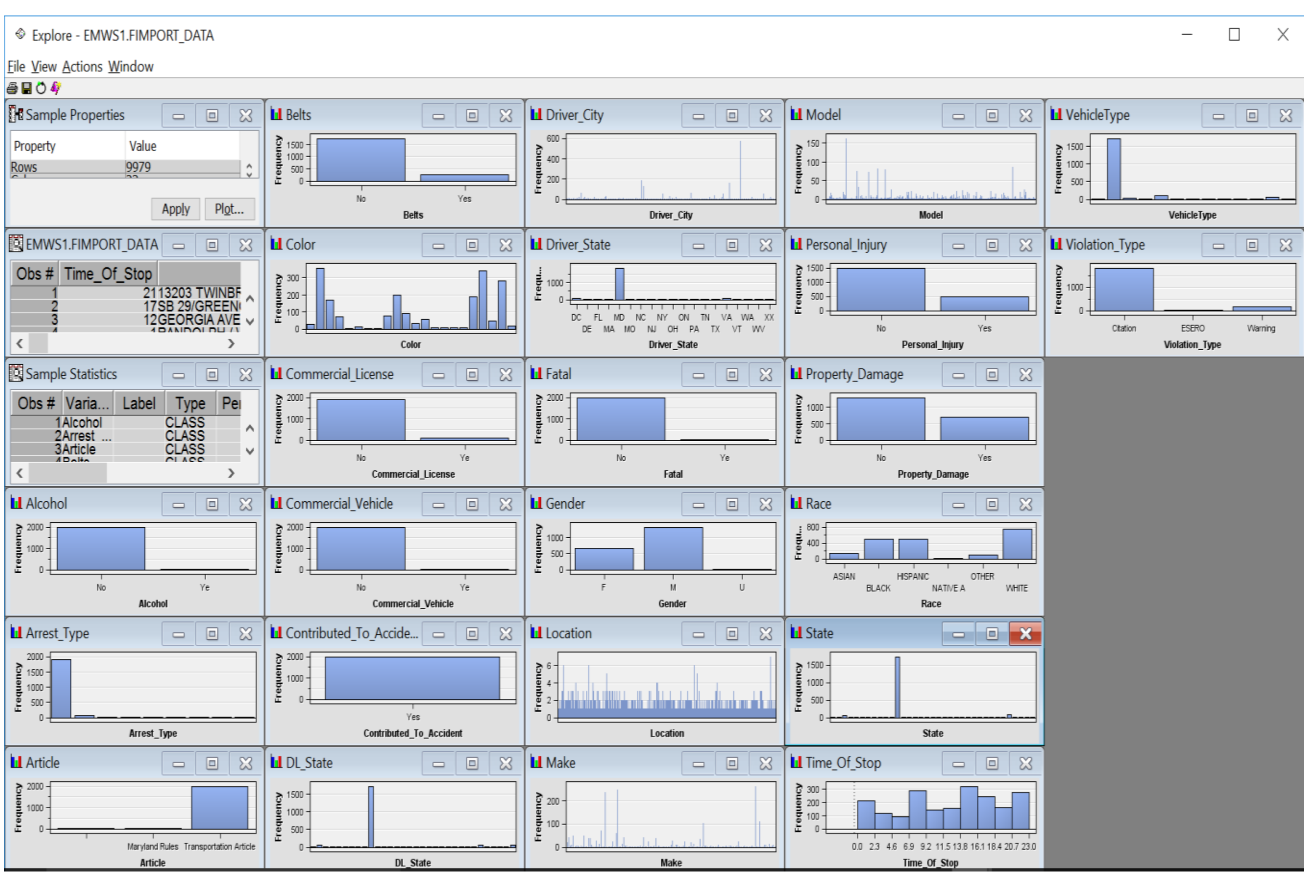
On exploration, it was found that some variables had missing values. We use the Impute node to populate the missing values. Below are the input methods for class and interval variables' imputation:

Class Variables: Default Input Method – Count
Interval Variables: Default Input Method – Mean

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of missing	No
Status	
Create Time	11/10/16 12:20 AM

4.3 Distribution of Input Variable

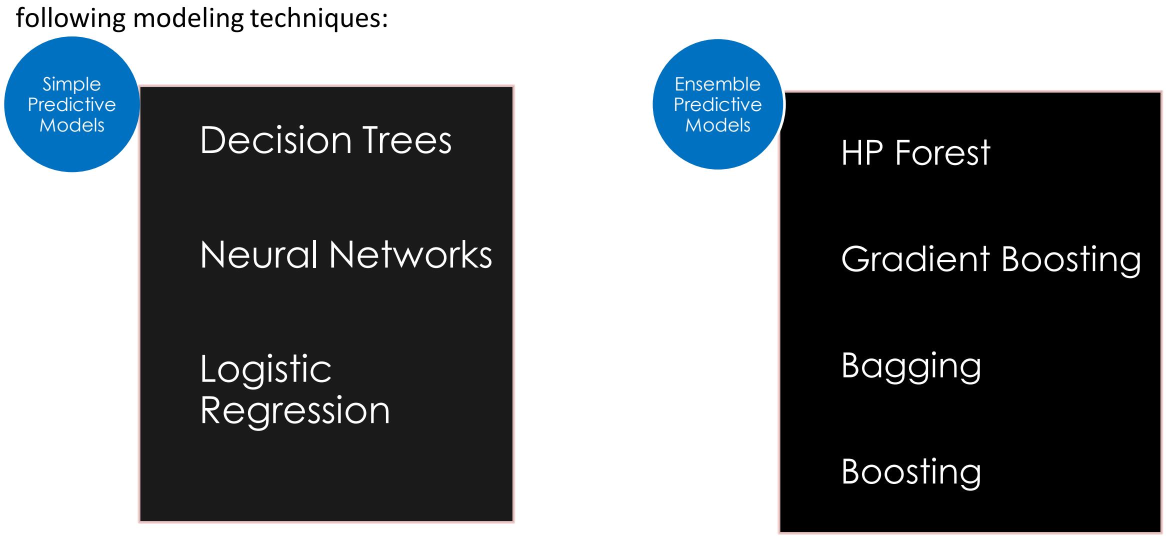
The figure below shows the distribution of input variables considered for analysis.



5 Data Modeling using SAS Enterprise Miner 9.4

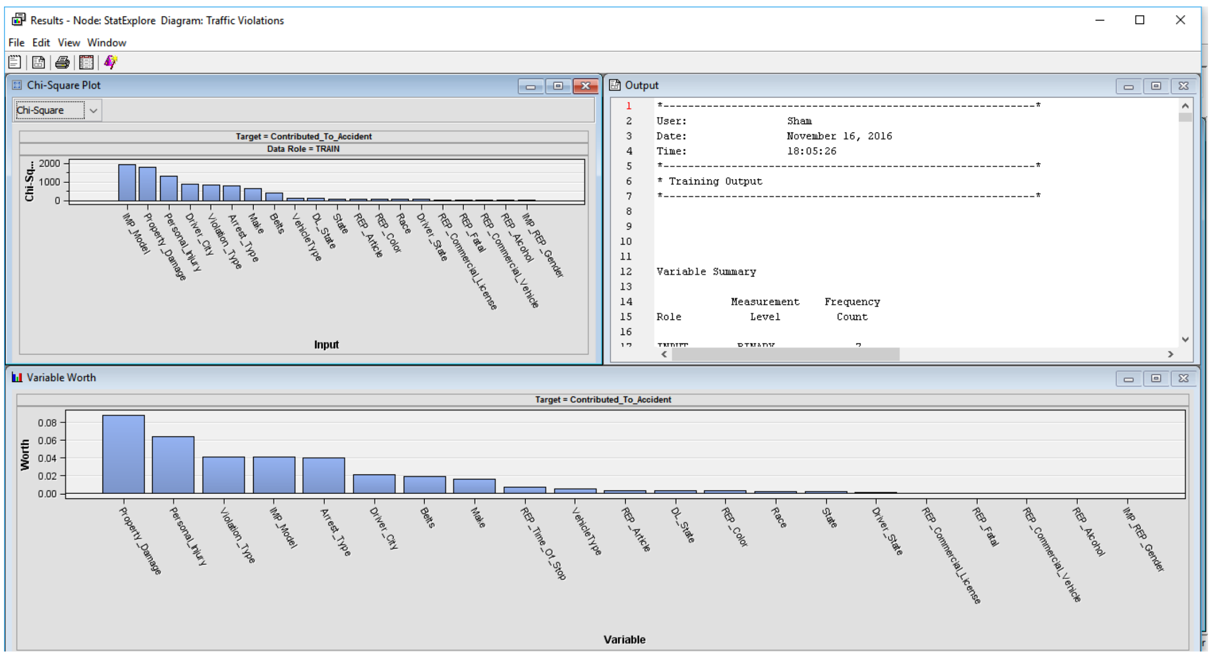
The main objectives of our project are:

- 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:



5.1 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:



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Arrest_Type	INPUT	15	0	A - Marked Patrol	87.33	Q - Marked Laser	6.03
TRAIN	Belts	INPUT	2	0	No	93.12	Yes	6.88
TRAIN	DL_State	INPUT	47	0	MD	86.64	VA	3.36
TRAIN	Driver_City	INPUT	474	0	SILVER SPRING	25.25	GAITHERSBURG	9.80
TRAIN	Driver_State	INPUT	38	0	MD	91.61	DC	2.96
TRAIN	IMP_Model	INPUT	1061	0	4S	9.61	TK	5.66
TRAIN	IMP_REP_Gender	INPUT	2	0	M	66.32	F	33.68
TRAIN	Make	INPUT	280	0	TOYOTA	11.72	HONDA	11.00
TRAIN	Personal_Injury	INPUT	2	0	No	87.31	Yes	12.69
TRAIN	Property_Damage	INPUT	2	0	No	82.19	Yes	17.81
TRAIN	REP_Alcohol	INPUT	2	0	No	99.47	Yes	0.53
TRAIN	REP_Article	INPUT	3	0	Transportation Article	97.72	_DEFAULT_	1.67
TRAIN	REP_Color	INPUT	21	0	BLACK	18.78	SILVER	17.91
TRAIN	REP_Commercial_License	INPUT	2	0	No	96.37	Yes	3.63
TRAIN	REP_Commercial_Vehicle	INPUT	2	0	No	99.72	Yes	0.28
TRAIN	REP_Fatal	INPUT	2	0	No	99.81	Yes	0.19
TRAIN	Race	INPUT	6	0	WHITE	39.25	BLACK	26.61
TRAIN	State	INPUT	50	0	MD	87.57	VA	4.05
TRAIN	VehicleType	INPUT	18	0	02 - Automobile	85.70	05 - Light Duty	6.46
TRAIN	Violation_Type	INPUT	3	0	Citation	80.09	Warning	18.25
TRAIN	Contributed_To_Accident	TARGET	2	0	Yes	50.10	No	49.90

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

5.2 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.

Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	Group rare levels
Class Targets	None
Treat Missing as Level	No

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

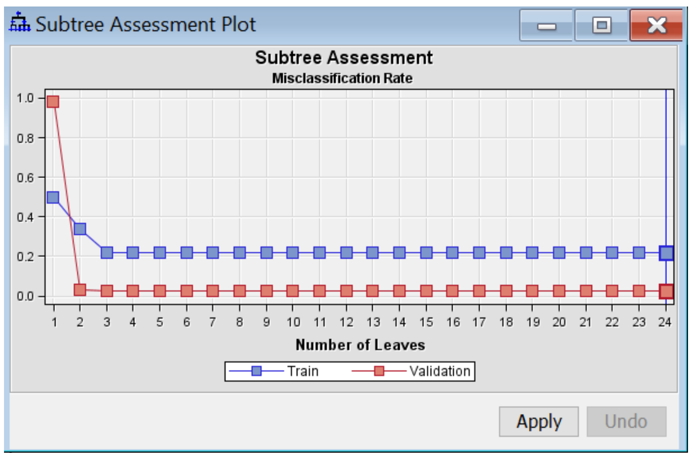
Decision tree provides an excellent introduction to predictive modeling. These models are conceptually easy to understand and they readily accommodate nonlinear associations between input variables and one or more target variables.

Max Tree:

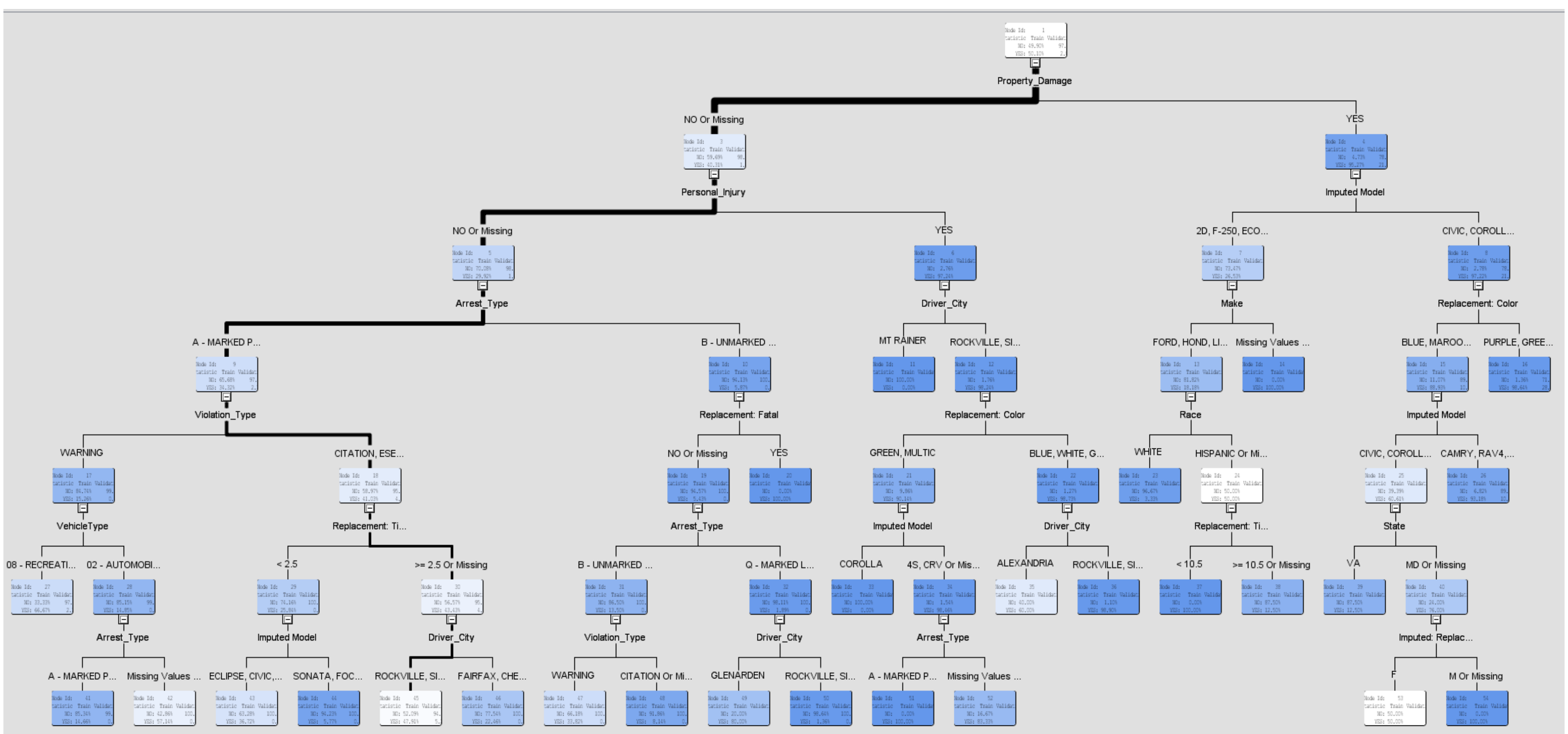
Here, the root node is trained to give the maximum classification tree.

Right click on root node -> Train node

Subtree Assessment plot of the MAX TREE:



No. of. leaves = 24



Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
2057	4915	65	2942

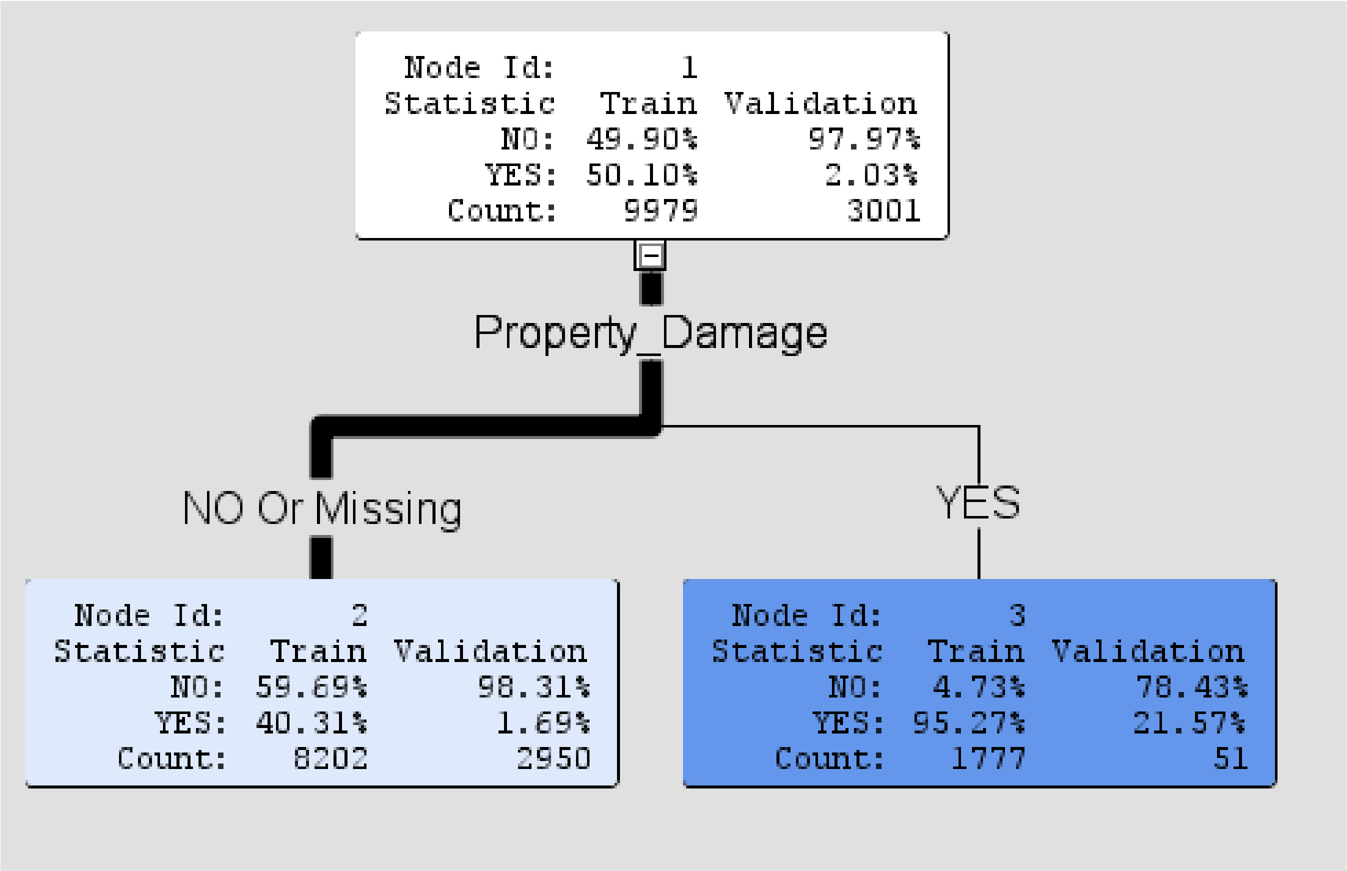
Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

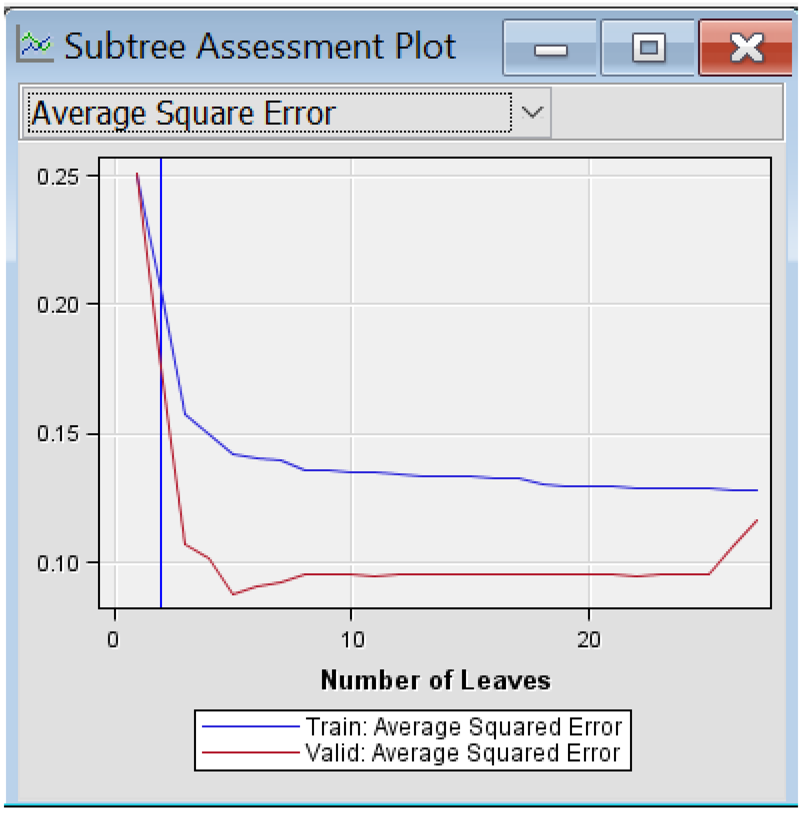
False Negative	True Negative	False Positive	True Positive
50	2792	148	11

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.

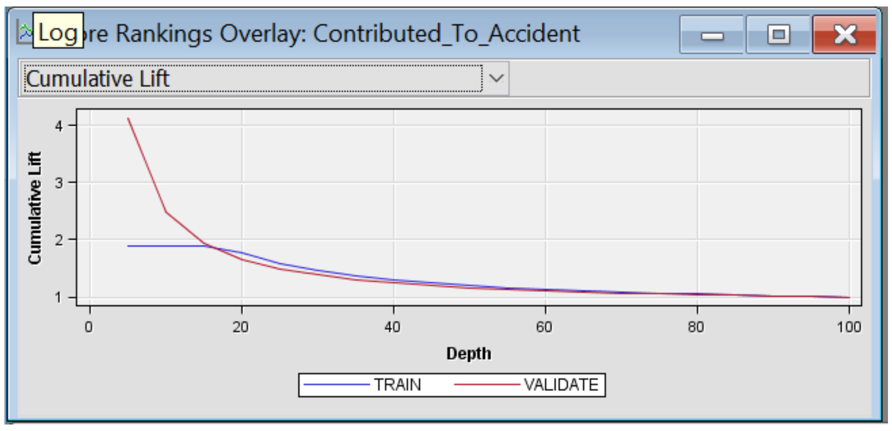
Optimal Tree:

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.

Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
3306	4896	84	1693

Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

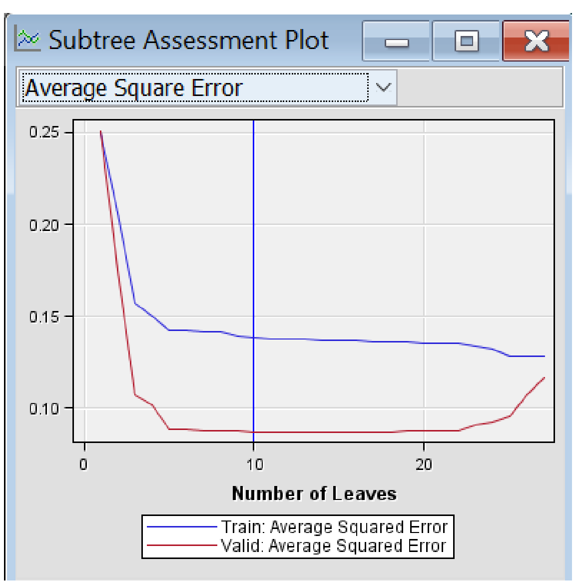
False Negative	True Negative	False Positive	True Positive
50	2900	40	11

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Property_Damage		1	1.0000	1.0000	1.0000

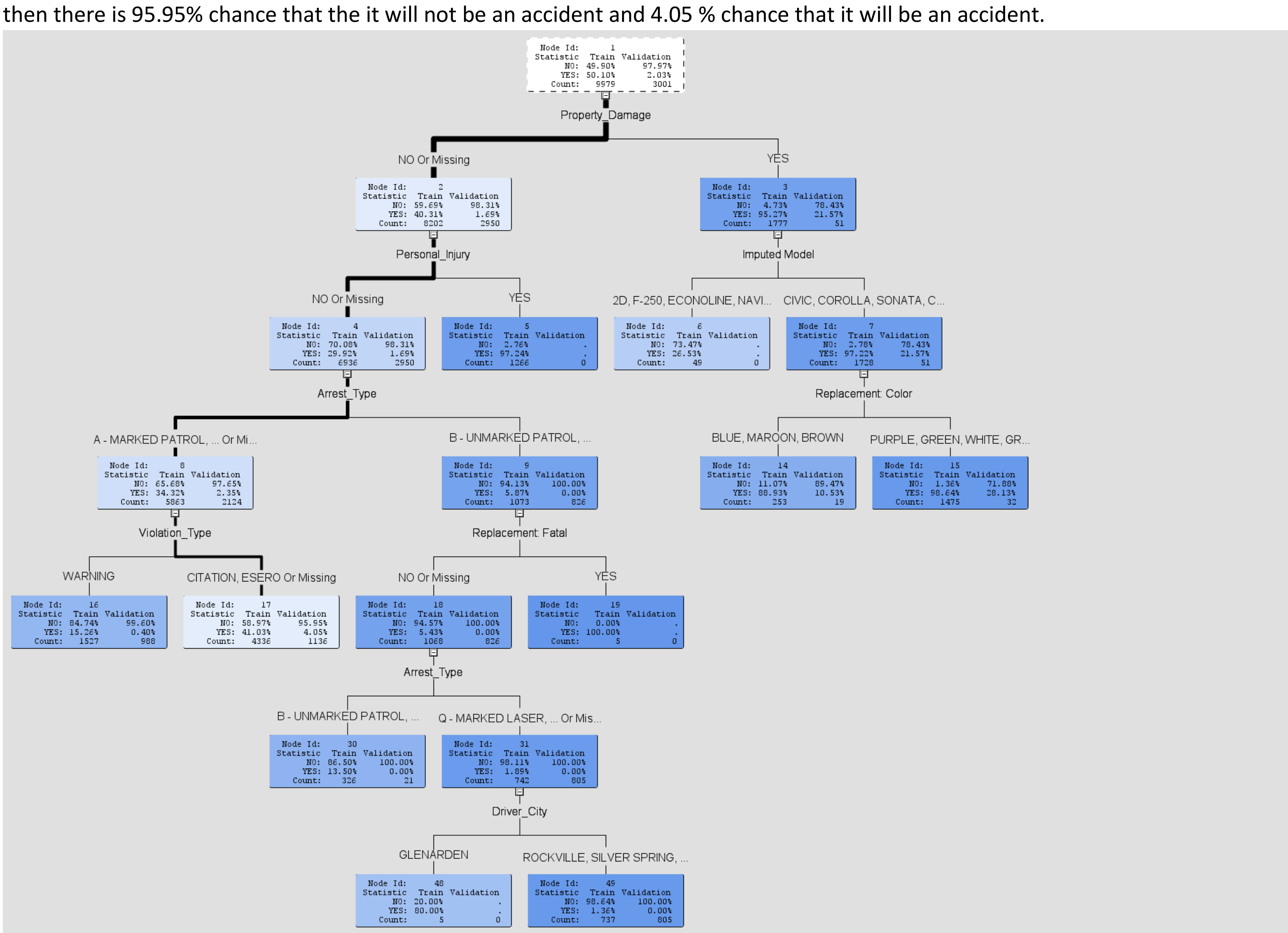
The above represents the variable importance of our data set.

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.



```
Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

  False      True      False      True
Negative   Negative   Positive   Positive

  2079       4896        84       2920


Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

  False      True      False      True
Negative   Negative   Positive   Positive

   50        2900        40        11
```

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Contributed T...		NOBS	Sum of Frequ...	9979	3001	
Contributed T...		MISC	Misclassificati...	0.216755	0.02999	
Contributed T...		MAX	Maximum Abs...	0.986441	0.986441	
Contributed T...		SSE	Sum of Squar...	2766.975	523.2944	
Contributed T...		ASE	Average Squa...	0.13864	0.087187	
Contributed T...		RASE	Root Average ...	0.372344	0.295274	
Contributed T...		DIV	Divisor for ASE	19958	6002	
Contributed T...		DFT	Total Degrees...	9979	.	

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

5.3 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.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	2.6037	0.1075	586.29	<.0001		13.514
Personal_Injury	No	1	-2.2418	0.0883	644.81	<.0001		0.106
Property_Damage	No	1	-1.9336	0.0587	1085.68	<.0001		0.145
TG_Arrest_Type	A - MARKED PATROL	1	0.9112	0.0523	304.02	<.0001		2.487

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

<u>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.</u>

Confusion Matrix:

```
Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

  False      True      False      True
Negative   Negative   Positive   Positive

  1932       4791       189       3067

Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

  False      True      False      True
Negative   Negative   Positive   Positive

   28        2517       423        33
```

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.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Contributed T...		AIC	Akaike's Infor...	8619.475	.	
Contributed T...		ASE	Average Squa...	0.139264	0.102744	
Contributed T...		AVERR	Average Error ...	0.429776	0.359591	
Contributed T...		DFE	Degrees of Fr...	9958	.	
Contributed T...		DFM	Model Degree...	21	.	
Contributed T...		DFT	Total Degrees...	9979	.	
Contributed T...		DIV	Divisor for ASE	19958	6002	
Contributed T...		ERR	Error Function	8577.475	2158.267	
Contributed T...		FPE	Final Predictio...	0.139852	.	
Contributed T...		MAX	Maximum Abs...	0.993851	0.977399	
Contributed T...		MSE	Mean Square ...	0.139558	0.102744	
Contributed T...		NOBS	Sum of Frequ...	9979	3001	
Contributed T...		NW	Number of Est...	21	.	
Contributed T...		RASE	Root Average ...	0.373182	0.320537	
Contributed T...		RFPE	Root Final Pre...	0.373968	.	
Contributed T...		RMSE	Root Mean Sq...	0.373575	0.320537	
Contributed T...		SBC	Schwarz's Bay...	8770.848	.	
Contributed T...		SSE	Sum of Squar...	2779.44	616.6683	
Contributed T...		SUMW	Sum of Case ...	19958	6002	
Contributed T...		MISC	Misclassificati...	0.212546	0.150283	

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

5.4 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.

```
Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

  False       True       False       True
Negative    Negative    Positive    Positive

  1471        4534        446         3528


Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

  False       True       False       True
Negative    Negative    Positive    Positive

   24         2390        550          37
```

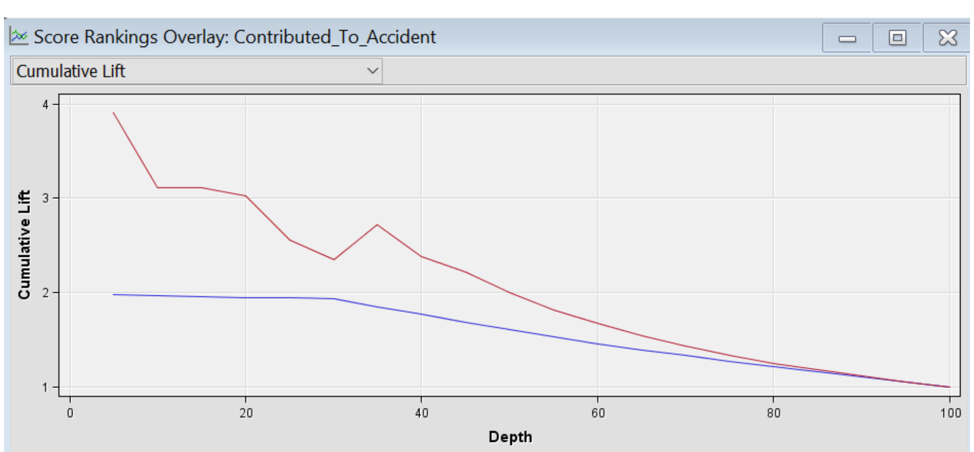
Based on the above result from event classification table of the model, we see that the model is almost good in classification

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Contributed T...		DFT	Total Degrees...	9979	.	
Contributed T...		DFE	Degrees of Fr...	9020	.	
Contributed T...		DFM	Model Degree...	959	.	
Contributed T...		NW	Number of Est...	959	.	
Contributed T...		AIC	Akaike's Infor...	9982.285	.	
Contributed T...		SBC	Schwarz's Bay...	16894.99	.	
Contributed T...		ASE	Average Squa...	0.131061	0.101764	
Contributed T...		MAX	Maximum Abs...	0.991256	0.980583	
Contributed T...		DIV	Divisor for ASE	19958	6002	
Contributed T...		NOBS	Sum of Frequ...	9979	3001	
Contributed T...		RASE	Root Average ...	0.362023	0.319005	
Contributed T...		SSE	Sum of Squar...	2615.714	610.7873	
Contributed T...		SUMW	Sum of Case ...	19958	6002	
Contributed T...		FPE	Final Predictio...	0.15893	.	
Contributed T...		MSE	Mean Square...	0.144995	0.101764	
Contributed T...		RFPE	Root Final Pre...	0.39866	.	
Contributed T...		RMSE	Root Mean Sq...	0.380782	0.319005	
Contributed T...		AVERR	Average Error ...	0.404063	0.3372	
Contributed T...		ERR	Error Function	8064.285	2023.874	
Contributed T...		MISC	Misclassificati...	0.192103	0.19127	
Contributed T...		WRONG	Number of Wr...	1917	574	

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.

5.5 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.

Event Classification Table

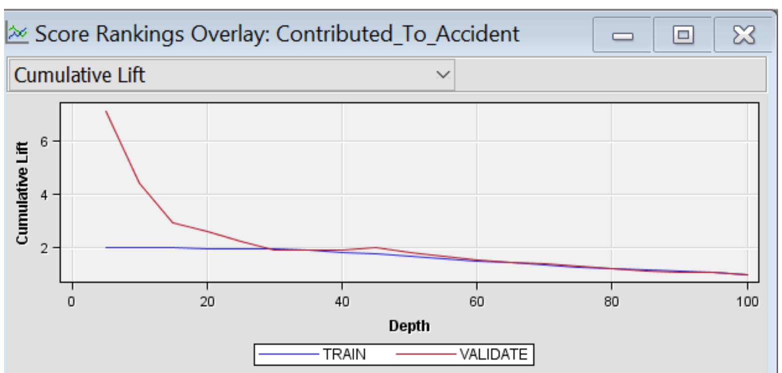
Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
1713	4795	185	3286

Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
35	2716	224	26

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.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Contribu...		ASE	Average ...	0.12246	0.139494	.
Contribu...		DIV	Divisor f...	19958	6002	.
Contribu...		MAX	Maximu...	0.96613	0.967046	.
Contribu...		NOBS	Sum of ...	9979	3001	.
Contribu...		RASE	Root Av...	0.349943	0.373489	.
Contribu...		SSE	Sum of ...	2444.063	837.2417	.
Contribu...		DISF	Frequen...	9979	3001	.
Contribu...		MISC	Misclass...	0.190199	0.086305	.
Contribu...		WRON...	Number ...	1898	259	.

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

5.6 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.

Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
2227	4390	590	2772

Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
27	2623	317	34

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

Fit Statistics

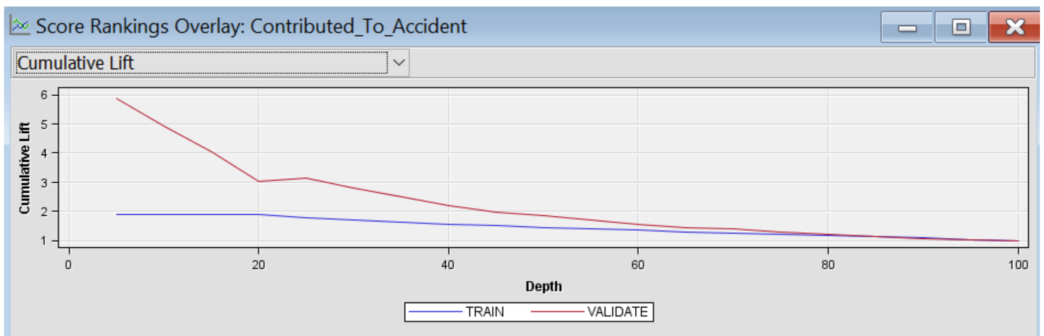
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Contributed T...		NOBS	Sum of Frequ...	9979	3001
Contributed T...		SUMW	Sum of Case ...	19958	6002
Contributed T...		MISC	Misclassificati...	0.282293	0.114628
Contributed T...		MAX	Maximum Abs...	0.77402	0.77402
Contributed T...		SSE	Sum of Squar...	3916.203	1071.679
Contributed T...		ASE	Average Squa...	0.196222	0.178554
Contributed T...		RASE	Root Average ...	0.44297	0.422556
Contributed T...		DIV	Divisor for ASE	19958	6002
Contributed T...		DFT	Total Degrees...	9979	.

Misclassification rate:

Train: 0.282293, Validation: 0.114628

Avg. Square Error:

Train: 0.196222, Validation: 0.178554



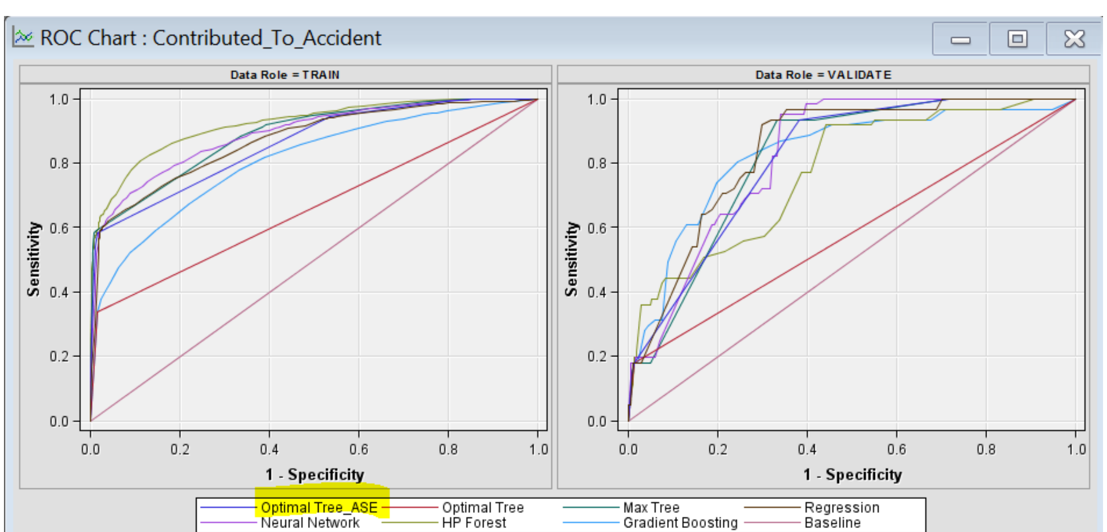
Good Cumulative lift

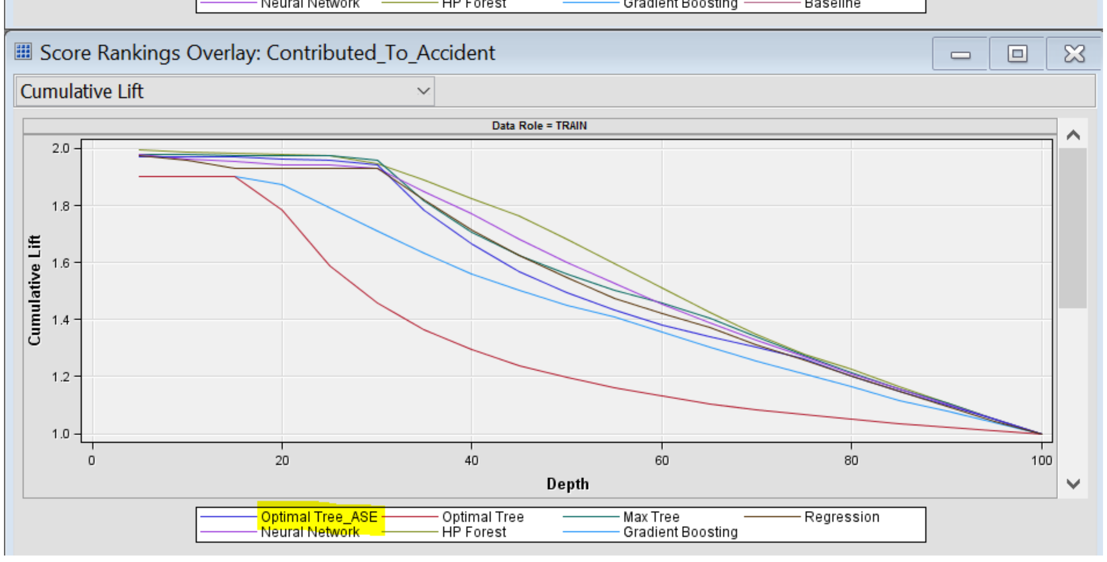
5.7 Model Comparison

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion : Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error
Y	Tree3	Tree3	Optima...	Contrib...		0.02999	9979	0.2167...	0.9864...	2766.9...	0.13864	0.3723...	19958	9979	3001	0.02999	0.9864...	523.29...	0.0871...
	Tree2	Tree2	Optima...	Contrib...		0.02999	9979	0.3397...	0.9527...	4106.9...	0.2057...	0.4536...	19958	9979	3001	0.02999	0.9527...	1050.6...	0.1750...
	HPDM...	HPDM...	HP For...	Contrib...		0.0863...	9979	0.1901...	0.96613	2444.0...	0.12246	0.3499...	19958	.	3001	0.0863...	0.9670...	837.24...	0.1394...
	Tree	Tree	Max Tr...	Contrib...		0.0893...	9979	0.2126...	0.9889...	2555.3...	0.1280...	0.3578...	19958	9979	3001	0.0893...	0.9864...	697.26...	0.1161...
	Boost	Boost	Gradie...	Contrib...		0.1146...	9979	0.2822...	0.77402	3916.2...	0.1962...	0.44297	19958	9979	3001	0.1146...	0.77402	1071.6...	0.1785...
	Reg	Reg	Regres...	Contrib...		0.1502...	9979	0.2125...	0.9938...	2779.44	0.1392...	0.3731...	19958	9979	3001	0.1502...	0.9773...	616.66...	0.1027...
	Neural	Neural	Neural ...	Contrib...		0.19127	9979	0.1921...	0.9912...	2615.7...	0.1310...	0.3620...	19958	9979	3001	0.19127	0.9805...	610.78...	0.1017...

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.

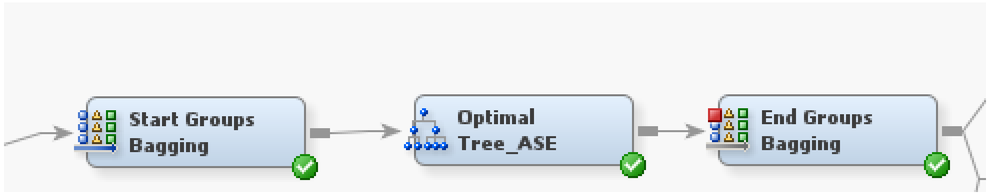




6. Performance Tuning of The Best- Selected Model:

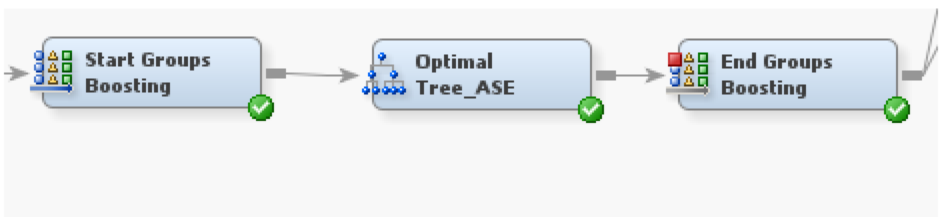
Bagging:

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.



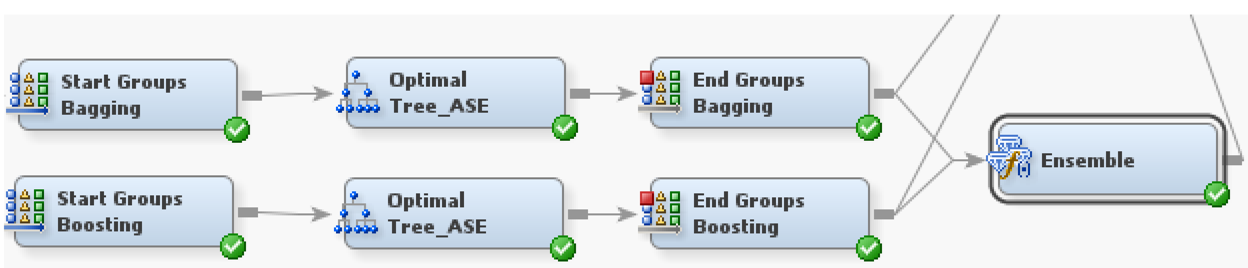
Boosting:

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.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion : Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misclassification Rate	Train: Number of Wrong Classifications	Valid: Average Squared Error
Y	EndGrp	EndGrp	End Gr...	Contrib...		0.02999	0.1418...	19958	0.9738...	9979	0.3766...	2831.7...	9979	0.2198...	2194	0.07536
	Tree3	Tree3	Optima...	Contrib...		0.02999	0.13864	19958	0.9864...	9979	0.3723...	2766.9...	.	0.2167...	.	0.0871...
	Ensmbl	Ensmbl	Ensem...	Contrib...		0.1822...	0.1530...	19958	0.8490...	9979	0.3912...	3054.3...	9979	0.2026...	2022	0.13438
	EndGrp2	EndGrp2	End Gr...	Contrib...		0.9796...	0.1939...	19958	0.7242...	9979	0.4404...	3871.6...	9979	0.4990...	4980	0.2364...

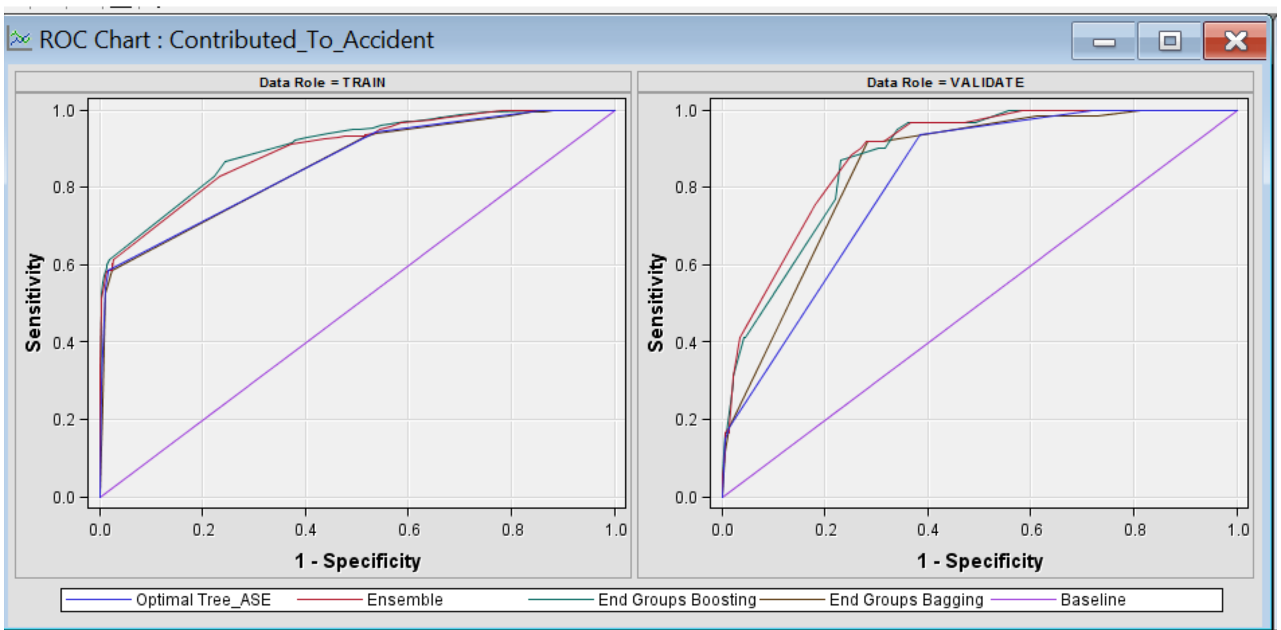
Bagging End Group:

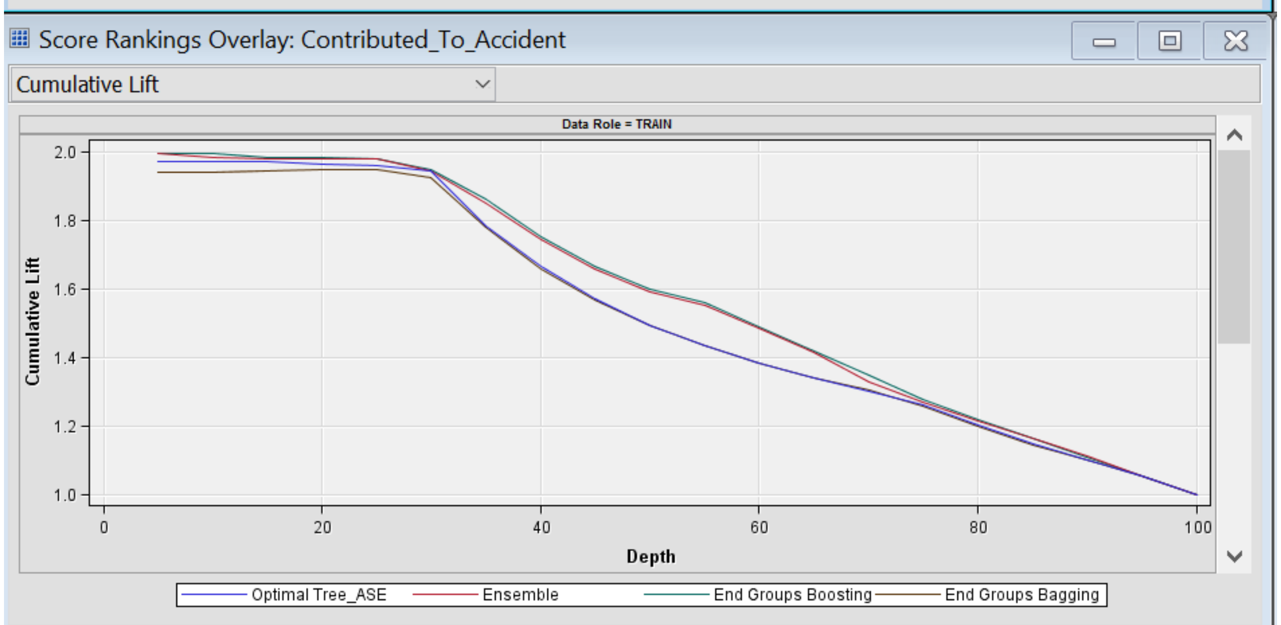
<u>Validation Misclassification rate = 0.02999</u>

<u>Validation Average Squared Error = 0.07536</u>

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

Event Classification Table

Data Role=TRAIN Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
2075	4861	119	2924

Data Role=VALIDATE Target=Contributed_To_Accident Target Label=' '

False Negative	True Negative	False Positive	True Positive
50	2900	40	11

Confusion Matrix Terms:

<table>
<tr><td></td><td></td><td colspan="2">Predicted</td></tr>
<tr><td></td><td></td><td>"+ve"</td><td>"-ve"</td></tr>
<tr><td rowspan="2">Actual</td><td>"+ve"</td><td>a</td><td>b</td></tr>
<tr><td>"-ve"</td><td>c</td><td>d</td></tr>
</table>

<table>
<tr><td></td><td></td><td colspan="2">Predicted</td></tr>
<tr><td></td><td></td><td>Accident = yes</td><td>Accident = no</td></tr>
<tr><td rowspan="2">Actual</td><td>Accident = yes</td><td>11</td><td>50</td></tr>
<tr><td>Accident = no</td><td>40</td><td>2900</td></tr>
</table>

Accuracy: Proportion of correct predictions

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

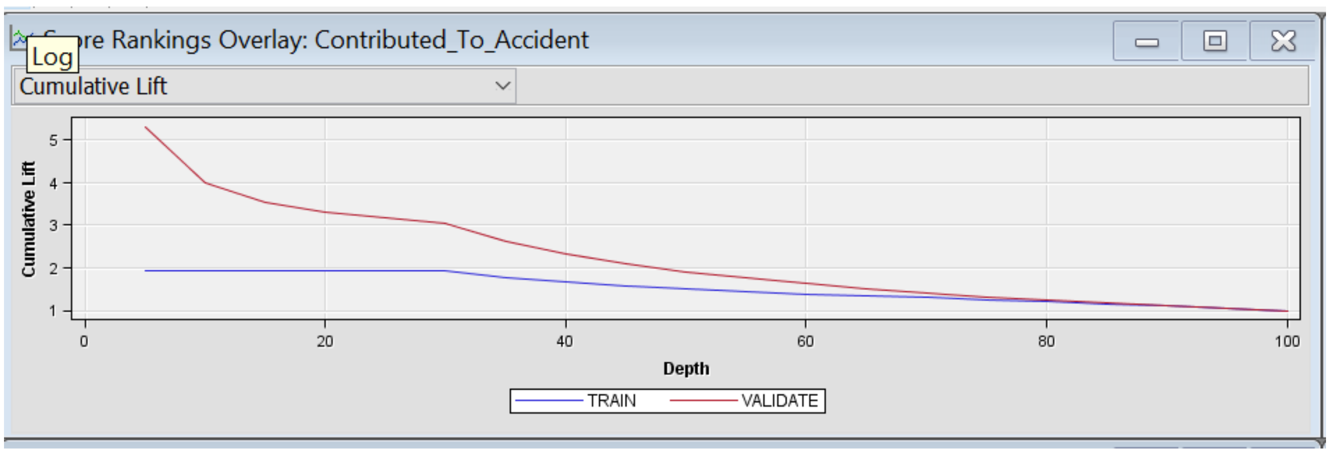
= (11+2900) / (11+50+40+2900)

= 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