Predicting Injury Severity of Drivers

- Arjun Achuthan(A20115366)

Introduction

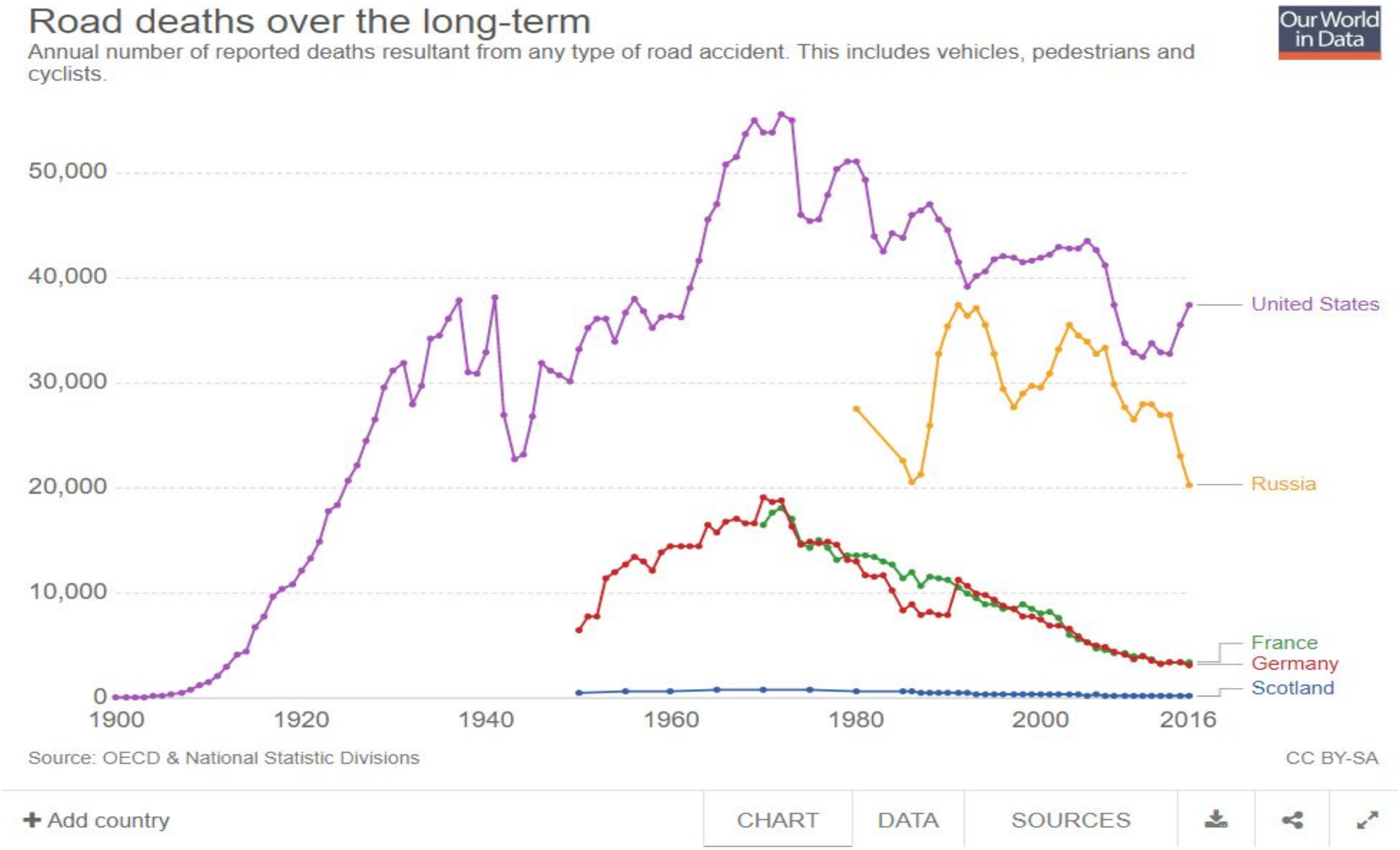
Every year we have huge number of road accidents happening in the US, with an estimate of 35,000 fatalities in 2016 alone.

Need for a thorough understanding of the circumstances and factors that lead to an accident to improve the situation.

Using car crash data provided by NHTSA and leverage our data mining skills to analyze and suggest a suitable data mining model that can best predict the injury severity.

Based on the analysis we will suggest recommendations to avoid injuries during accidents.

Car Crash deaths, trends:



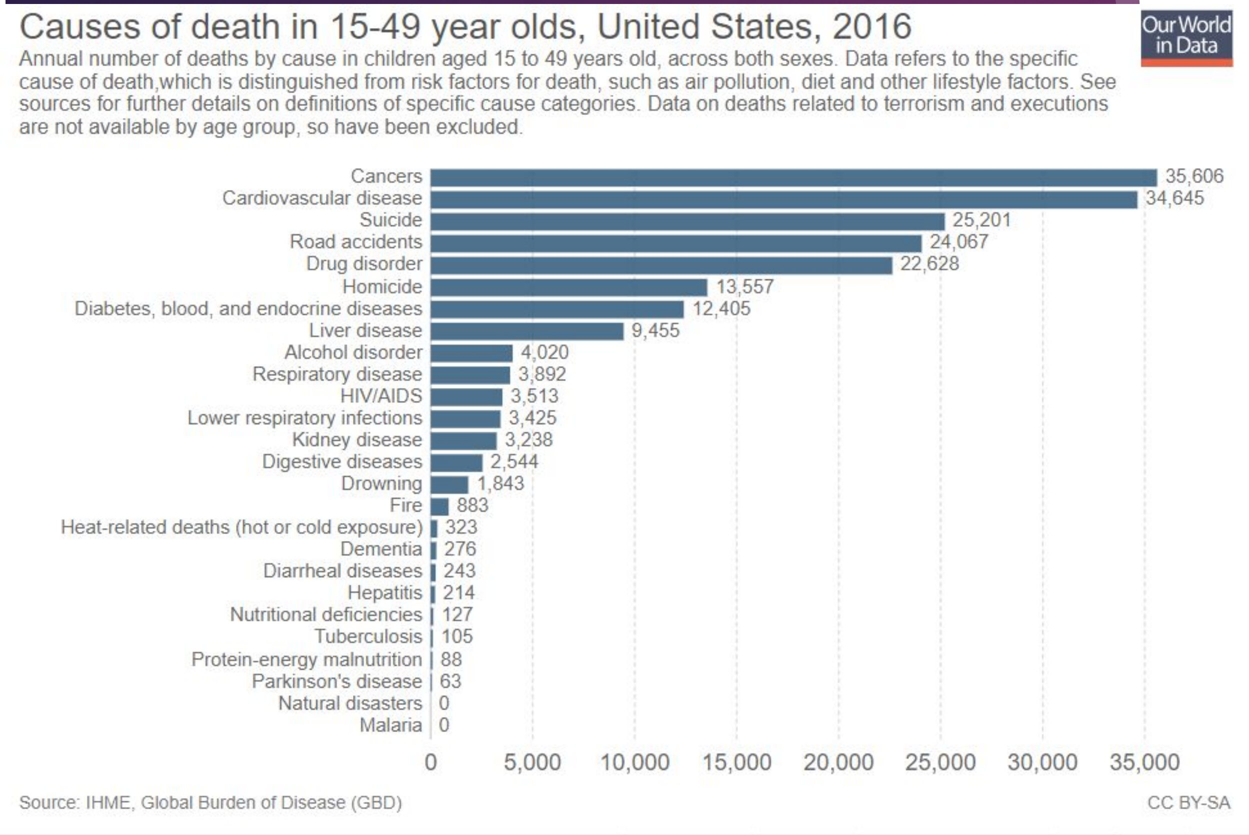
Road deaths over the long-term

Annual number of reported deaths resultant from any type of road accident. This includes vehicles, pedestrians and cyclists.

Source: OECD & National Statistic Divisions

CC BY-SA

Cause of deaths:

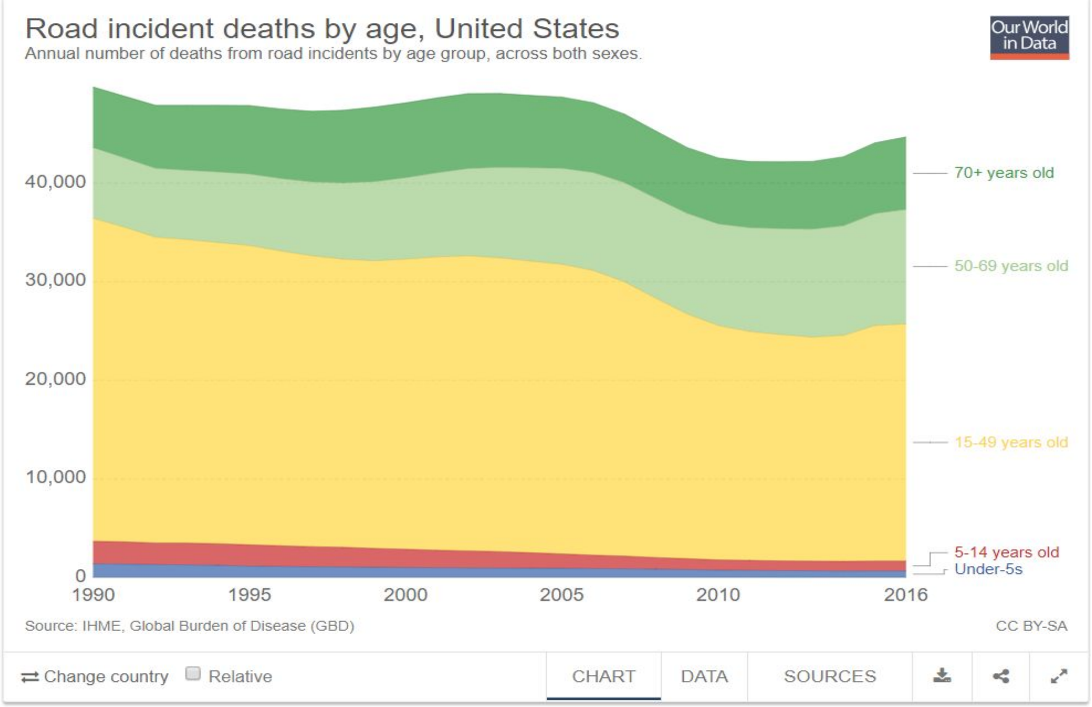


Causes of death in 15-49 year olds, United States, 2016

Annual number of deaths by cause in children aged 15 to 49 years old, across both sexes. Data refers to the specific cause of death,which is distinguished from risk factors for death, such as air pollution, diet and other lifestyle factors. See sources for further details on definitions of specific cause categories. Data on deaths related to terrorism and executions are not available by age group, so have been excluded.

Source: IHME, Global Burden of Disease (GBD)

CC BY-SA

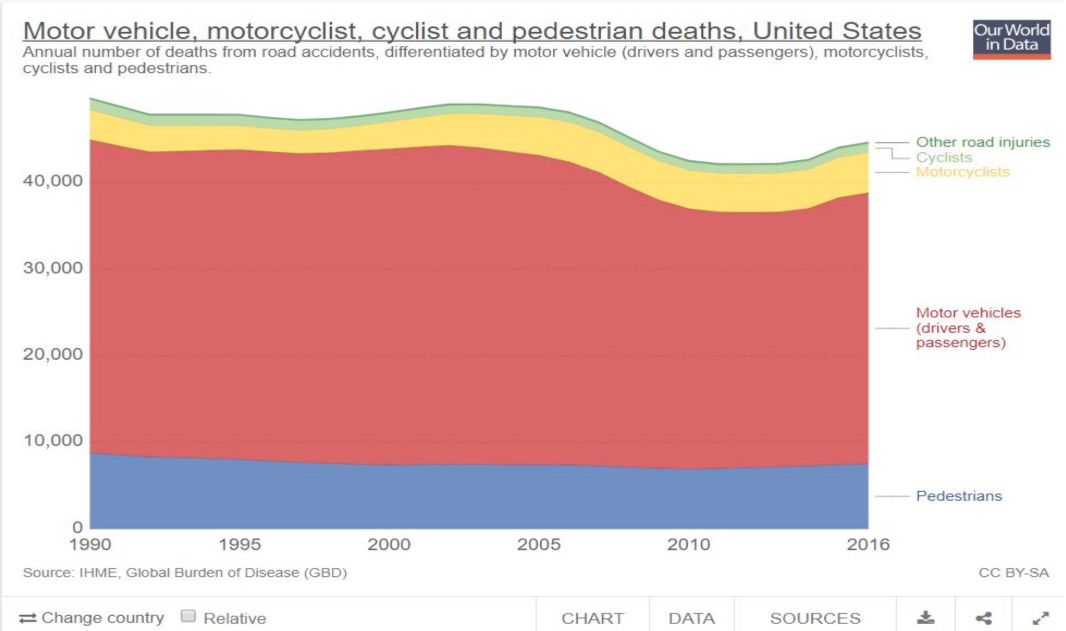


Road incident deaths by age, United States

Annual number of deaths from road incidents by age group, across both sexes.

Source: IHME, Global Burden of Disease (GBD)

CC BY-SA



Motor vehicle, motorcyclist, cyclist and pedestrian deaths, United States

Annual number of deaths from road accidents, differentiated by motor vehicle (drivers and passengers), motorcyclists, cyclists and pedestrians.

Source: IHME, Global Burden of Disease (GBD)

CC BY-SA

Project Information:

Project Implementation Methodology:

- CRISP- DM

Data Mining Tool used:

- KNIME

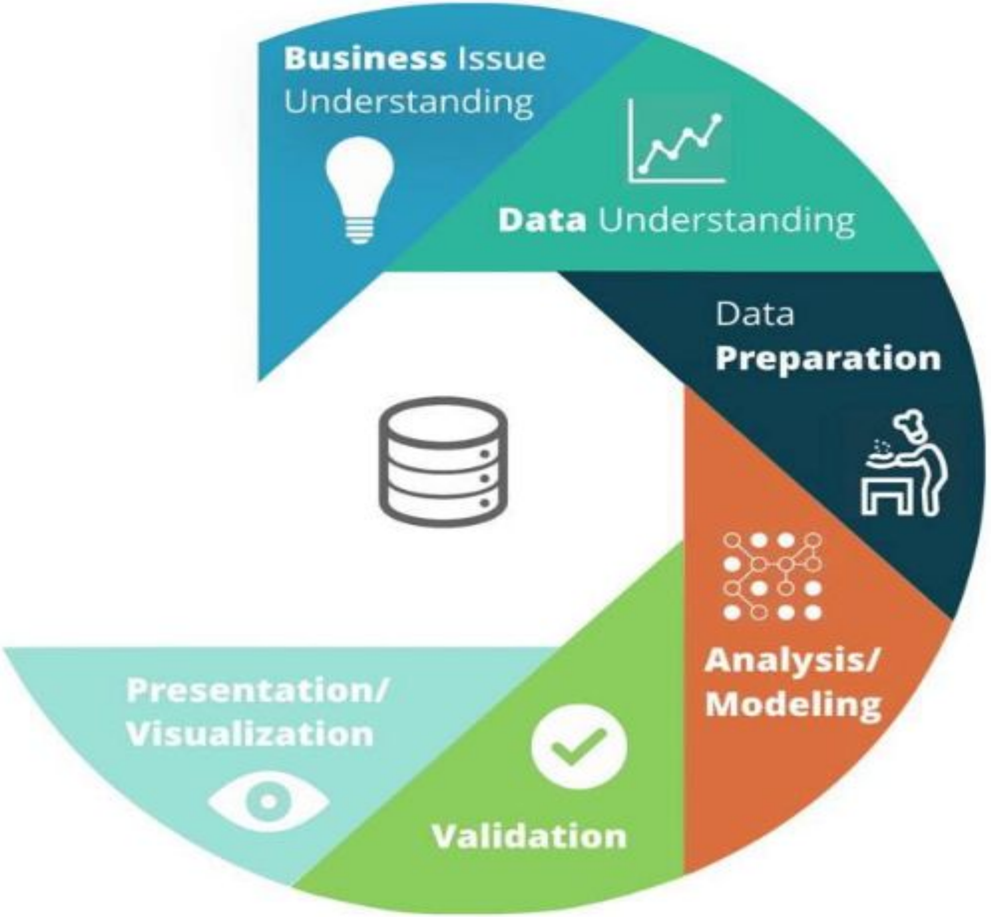
Data Mining Model:

- Random Forest
- Logistic Regression
- Neural Network

CRISP - DM

CRoss-Industry Standard Process for Data Mining

- Data Mining methodology
- Process Model
- For anyone
- Provides a complete blueprint
- Life cycle: 6 phases



1. Business Understanding

About NHTSA and their Mission:

NHTSA collets car crash data to support its mission to reduce motor vehicle crashes, injuries and deaths on our National Highways and roads.

They use data from many sources which includes National Automotive Sampling System (NASS) and General Estimates System (GES).

The National Automotive Sampling System (NASS) - General Estimates System (GES) data are obtained from a nationally representative probability sample selected from all police-reported crashes.

- **Project Objectives:**

  - To identify the key factors that contribute towards the injury severity of driver.
  - Building a suitable analytics model with the available data to predict likelihood of injury and its severity.
  - To provide suggestions and recommendation based on the insights from our analysis.

2. Data Understanding

➢ We were provided with 4 SAS dataset tables namely:

- **Accident**

  The Accident data file includes crash data with each record corresponding to an incident of accident involving one or more vehicles.

- **Distract**

  The Distract data file identifies if each of the drivers involves in crash was distracted prior to crash.

- **Vehicle**

  The Vehicle data file includes in-transport motor vehicle data as well as driver and precast data.

- **Person**

  The Person data file includes all the motorist and non-motorists involves in an accident

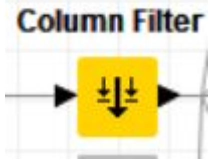
➢ A thorough analysis of attributes from each of the tables from the **(NASS)(GES) Analytical user's manual**

3. Data Preparation

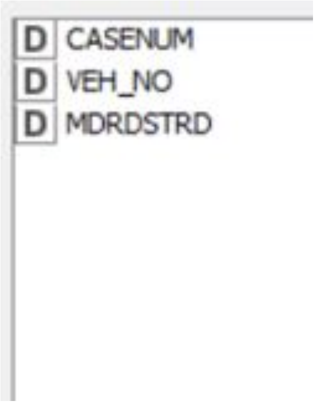
Variable selection:

Initially, variable selection was done based on domain knowledge and common sense. Later we used techniques like forward feature selection and backward feature elimination in Knime.

The variables were then filtered using the node:

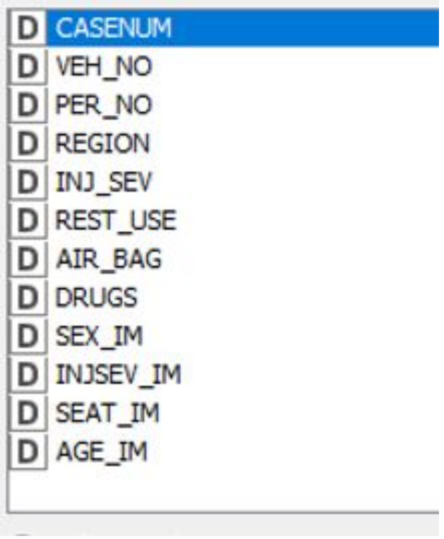


Distract Table



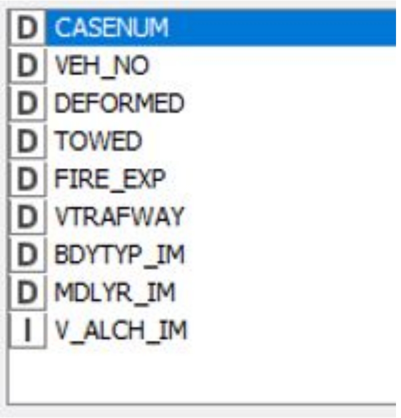
D CASENUM
D VEH_NO
D MDRDSTRD

Person table



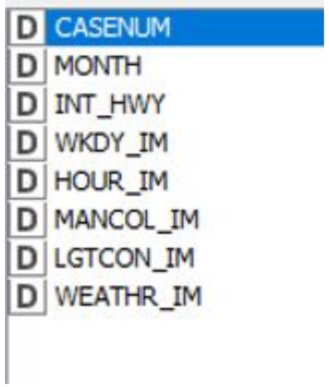
D CASENUM
D VEH_NO
D PER_NO
D REGION
D INJ_SEV
D REST_USE
D AIR_BAG
D DRUGS
D SEX_IM
D INJSEV_IM
D SEAT_IM
D AGE_IM

Vehicle Table



D CASENUM
D VEH_NO
D DEFORMED
D TOWED
D FIRE_EXP
D VTRAFWAY
D BDYTYP_IM
D MDLYR_IM
I V_ALCH_IM

Accident Table

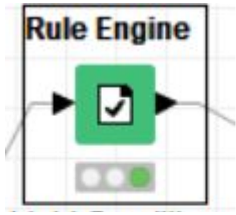


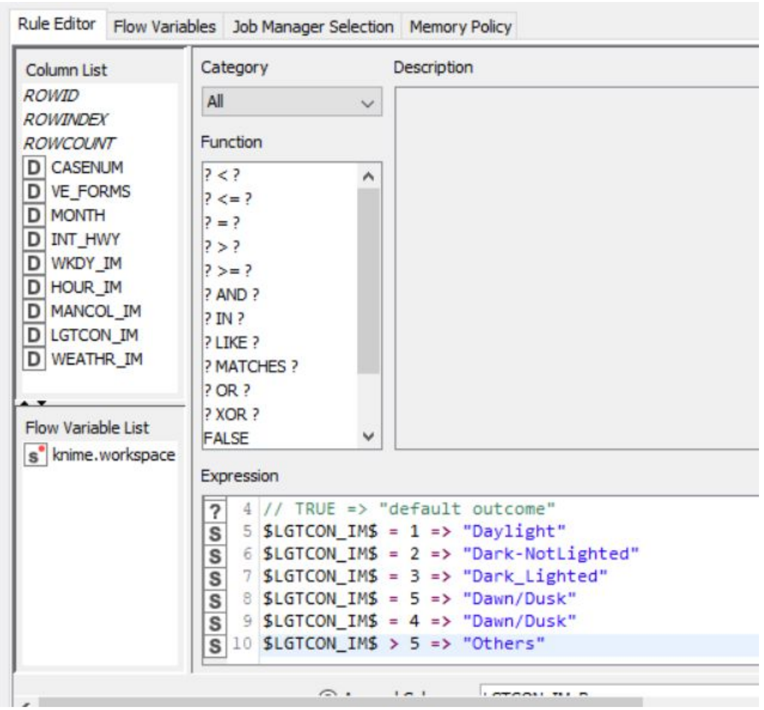
D CASENUM
D MONTH
D INT_HWY
D WKDY_IM
D HOUR_IM
D MANCOL_IM
D LGTCON_IM
D WEATHR_IM

Binning:

We have binned the selected predictor variables and tried to restrict the number of bins to less than 6.

Node used:

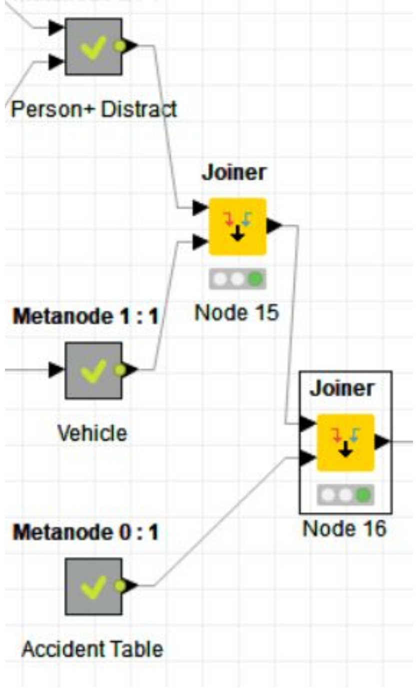


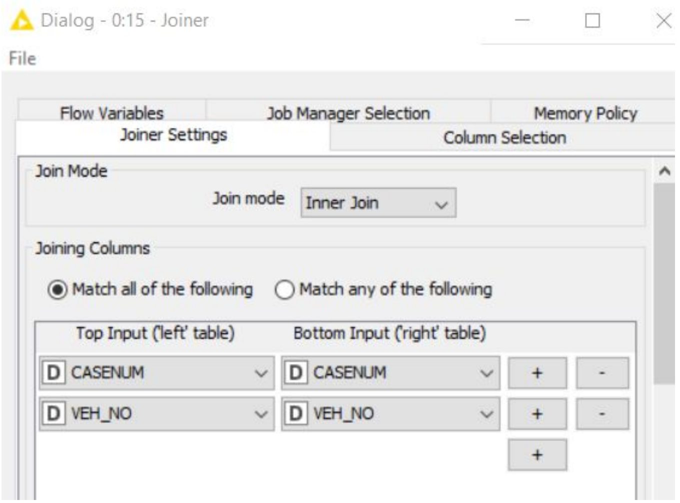


```
// TRUE => "default outcome"
$LGTCON_IM$ = 1 => "Daylight"
$LGTCON_IM$ = 2 => "Dark-NotLighted"
$LGTCON_IM$ = 3 => "Dark_Lighted"
$LGTCON_IM$ = 5 => "Dawn/Dusk"
$LGTCON_IM$ = 4 => "Dawn/Dusk"
$LGTCON_IM$ > 5 => "Others"
```

Joining tables:

Person, Distract, Vehicle and Accident tables were joined using Keys: CASENUM and VEH_NUM



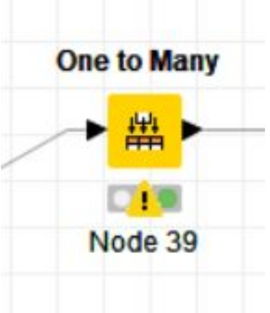


Join mode: Inner Join

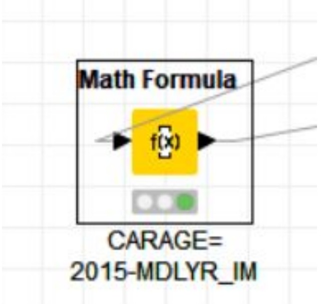
● Match all of the following ○ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)
D CASENUM	D CASENUM
D VEH_NO	D VEH_NO

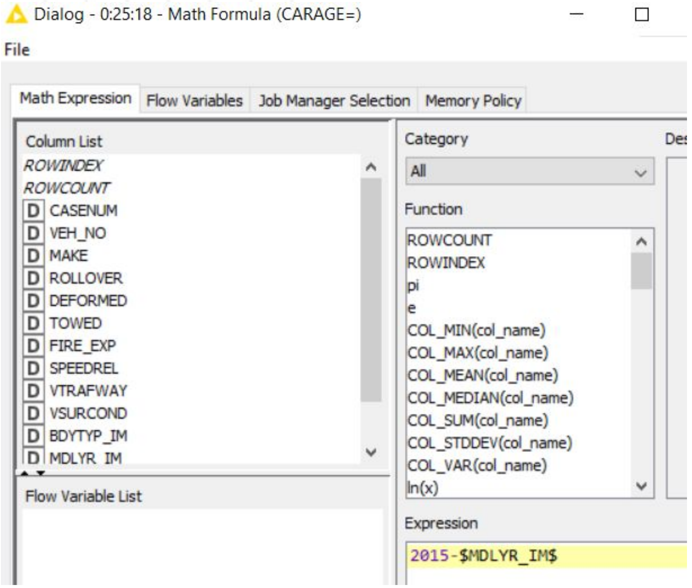
New Columns:



One-To-many node was used to Transform nominal data to numerical by creating dummy columns.
We have used this for the Neural Network.



Age of the vehicle was calculated from the MDLYR_IM column and a new is column was created called CARAGE with the age of the car.



Column List
ROWINDEX
ROWCOUNT
D CASENUM
D VEH_NO
D MAKE
D ROLLOVER
D DEFORMED
D TOWED
D FIRE_EXP
D SPEEDREL
D VTRAFWAY
D VSURCOND
D BDYTYP_IM
D MDLYR_IM

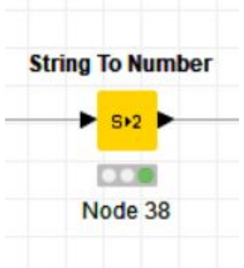
Flow Variable List

Category
All

Function
ROWCOUNT
ROWINDEX
pi
e
COL_MIN(col_name)
COL_MAX(col_name)
COL_MEAN(col_name)
COL_MEDIAN(col_name)
COL_SUM(col_name)
COL_STDDEV(col_name)
COL_VAR(col_name)
ln(x)

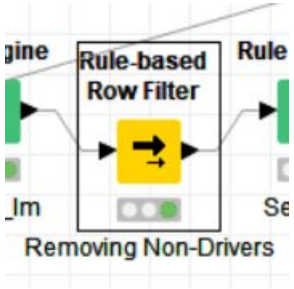
Expression
```
2015-$MDLYR_IM$
```

Data Conversions:

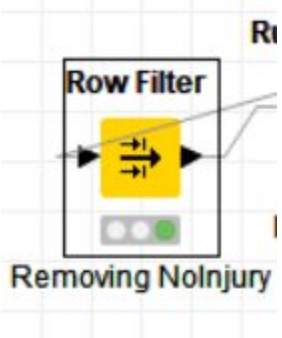


Converts string data type to Numeric data type.

Row Filters:



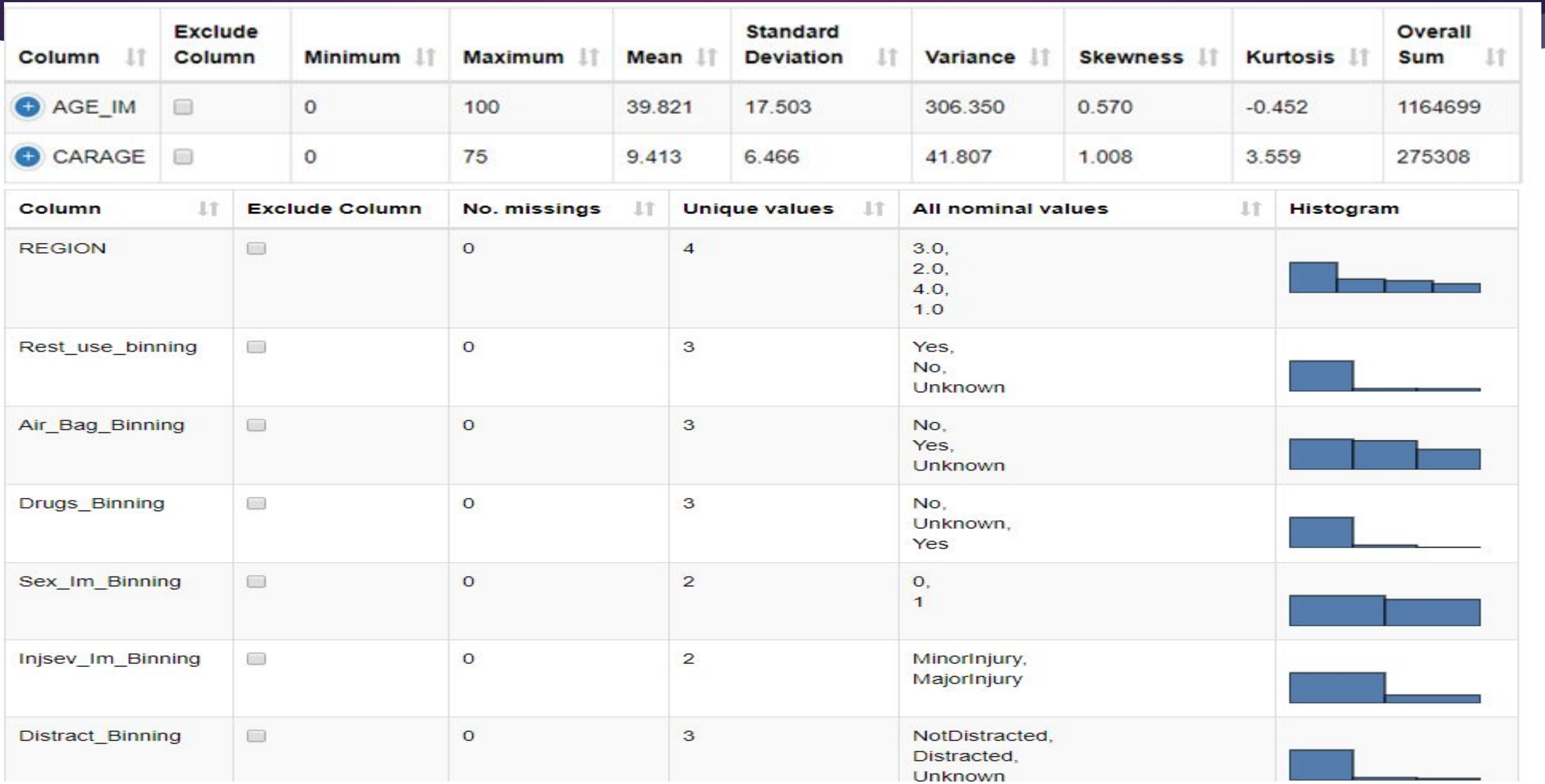
Rule based Row Filter used to remove records Non-Driver records



Row filter used to remove Drivers with no reported injury.

Variable Name	Variable Meaning	Attribute Type	Source Table
REGION	Accidents in the four regions	Nominal	Person
AGE_IM	Age of the drive	Numeric	Person
Rest_use_binning	Usage of safety restrictions	Nominal	Person
Air_Bag_Binning	Usage of airbags	Nominal	Person
Drugs_Binning	Whether the drive takes drugs before the accident	Nominal	Person
Sex_Im_Binning	Sex of the driver	Nominal	Person
Distract_Binning	Whether the driver is distracted before the accident	Nominal	Distract
DEFORMED_N	Whether the vehicle is deformed or how deformed in crash	Nominal	Vehicle
TOWED_N	Whether the vehicle is towed after crash	Nominal	Vehicle
VTRAFWAY_N	Trafficway information	Nominal	Vehicle
BDYTYP_IM_N	Body types of the vehicles	Nominal	Vehicle
CARAGE	Car age	Numeric	Vehicle (derived)
FIRE_EXP_N	Whether fire exists	Nominal	Vehicle
V_ALCH_IM_N	Whether the driver drink alcohol before the crash	Nominal	Vehicle
LGTCON_IM_B	Light conditions	Nominal	Accident
MANCOL_IM_B	Manners of collisions	Nominal	Accident
WKDY_IM_B	Which day in a week	Nominal	Accident
Weather_IM_B	Weather information	Nominal	Accident
Month_Binning	Season information	Nominal	Accident
INT_HWY_B	Whether the crash happens in interstate highways	Nominal	Accident
Hour_Binning	What time period the crash happens in a day	Nominal	Accident
Injsev_Im_Binning	The injury severity of the driver	Nominal	Person (target variable)

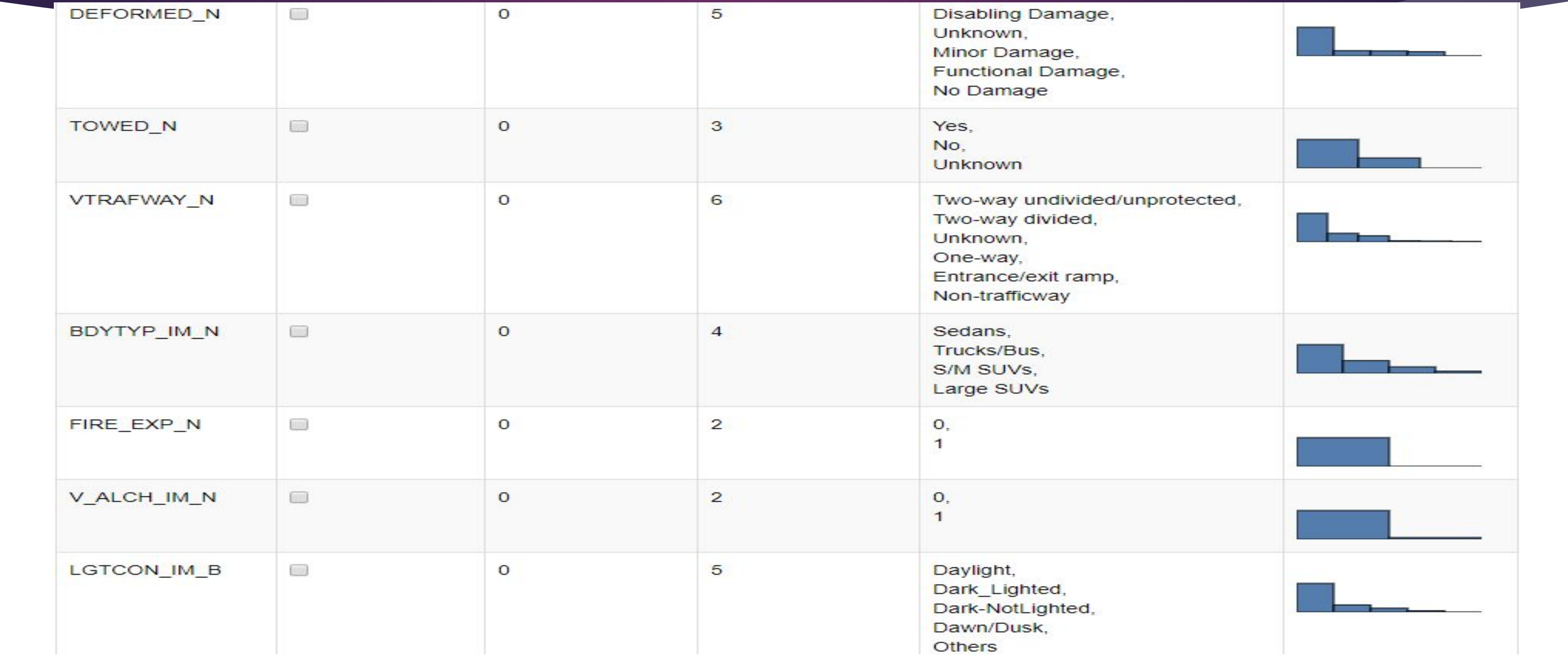
Descriptive statistics:



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum
⊕ AGE_IM	□	0	100	39.821	17.503	306.350	0.570	-0.452	1164699
⊕ CARAGE	□	0	75	9.413	6.466	41.807	1.008	3.559	275308

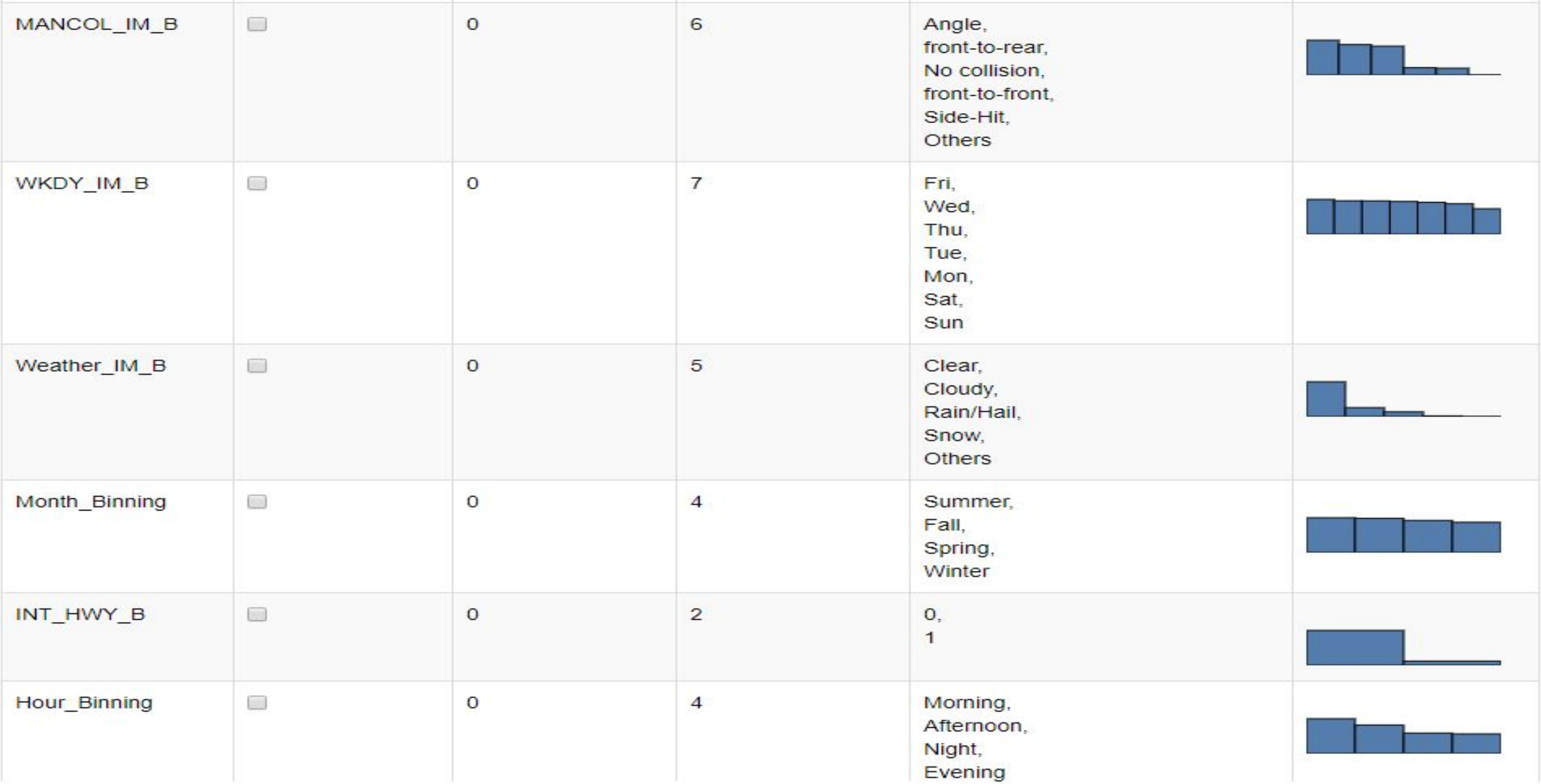
Column	Exclude Column	No. missings	Unique values	All nominal values	Histogram
REGION	□	0	4	3.0, 2.0, 4.0, 1.0	
Rest_use_binning	□	0	3	Yes, No, Unknown	
Air_Bag_Binning	□	0	3	No, Yes, Unknown	
Drugs_Binning	□	0	3	No, Unknown, Yes	
Sex_Im_Binning	□	0	2	0, 1	
Injsev_Im_Binning	□	0	2	MinorInjury, MajorInjury	
Distract_Binning	□	0	3	NotDistracted, Distracted, Unknown	

Descriptive statistics:



DEFORMED_N	☐	0	5	Disabling Damage, Unknown, Minor Damage, Functional Damage, No Damage	
TOWED_N	☐	0	3	Yes, No, Unknown	
VTRAFWAY_N	☐	0	6	Two-way undivided/unprotected, Two-way divided, Unknown, One-way, Entrance/exit ramp, Non-trafficway	
BDYTYP_IM_N	☐	0	4	Sedans, Trucks/Bus, S/M SUVs, Large SUVs	
FIRE_EXP_N	☐	0	2	0, 1	
V_ALCH_IM_N	☐	0	2	0, 1	
LGTCON_IM_B	☐	0	5	Daylight, Dark_Lighted, Dark-NotLighted, Dawn/Dusk, Others	

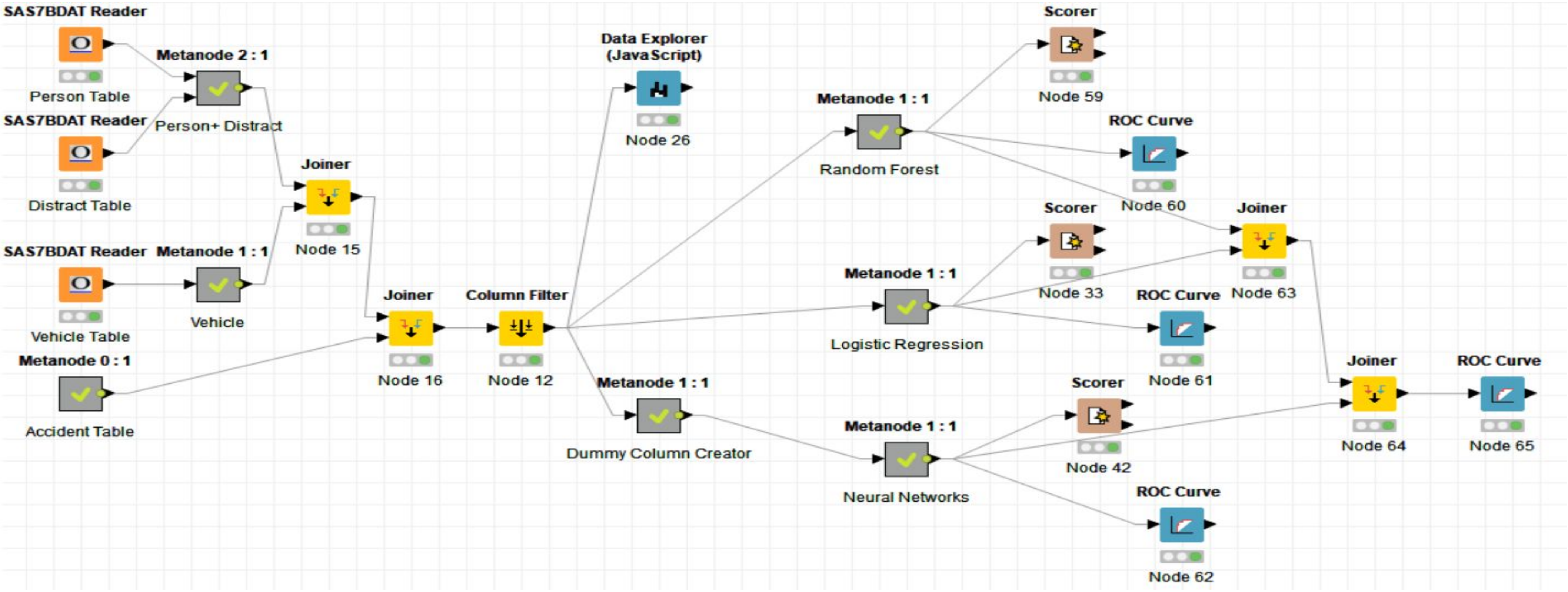
Descriptive statistics:



MANCOL_IM_B	☐	0	6	Angle, front-to-rear, No collision, front-to-front, Side-Hit, Others	
WKDY_IM_B	☐	0	7	Fri, Wed, Thu, Tue, Mon, Sat, Sun	
Weather_IM_B	☐	0	5	Clear, Cloudy, Rain/Hail, Snow, Others	
Month_Binning	☐	0	4	Summer, Fall, Spring, Winter	
INT_HWY_B	☐	0	2	0, 1	
Hour_Binning	☐	0	4	Morning, Afternoon, Night, Evening	

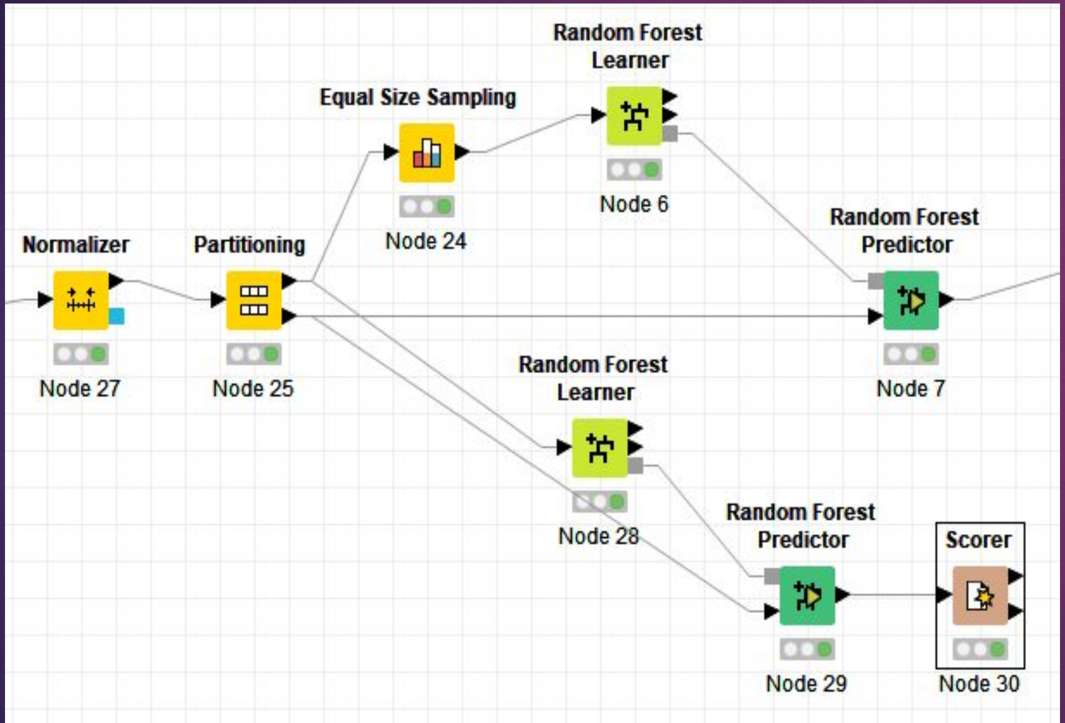
4. Modeling

Model Workflow:



Random Forest

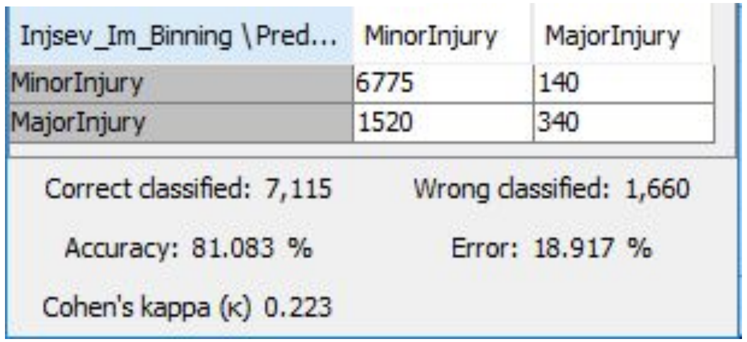
Model



- Confusion Matrix

Balancing Works!

Without Balancing:



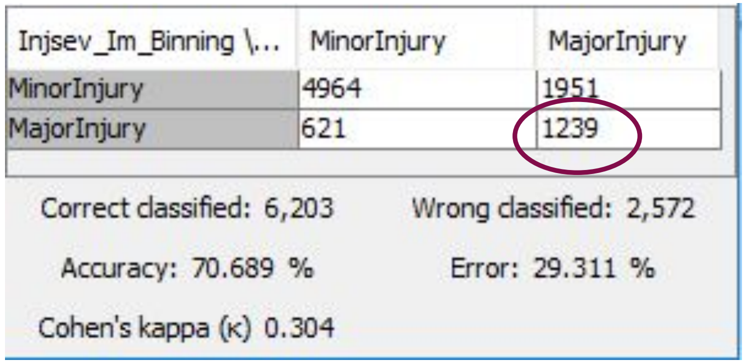
Injsev_Im_Binning \ Pred...	MinorInjury	MajorInjury
MinorInjury	6775	140
MajorInjury	1520	340

Correct classified: 7,115 Wrong classified: 1,660

Accuracy: 81.083 % Error: 18.917 %

Cohen's kappa (κ) 0.223

With Balancing (equal size sampling):

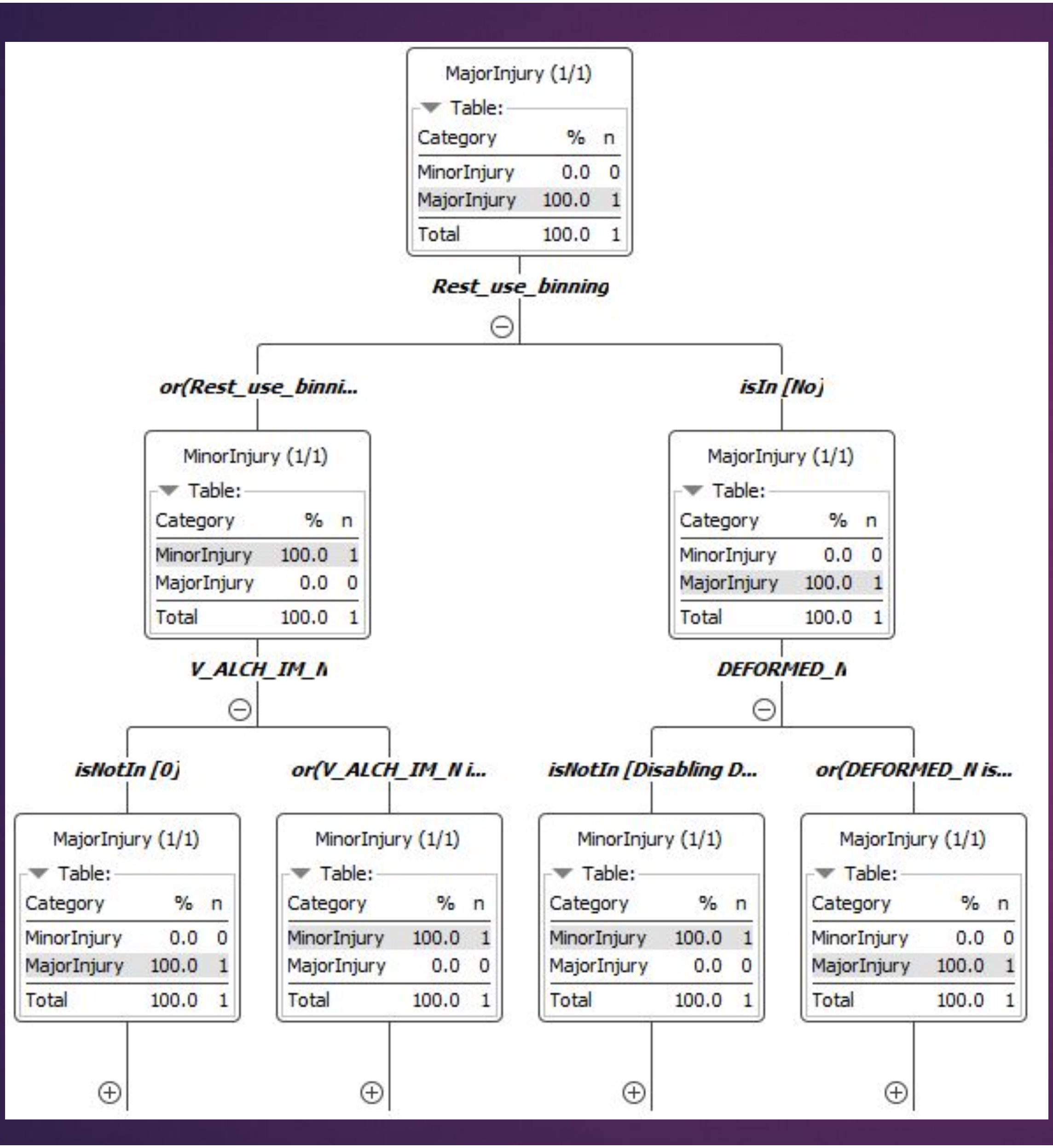


Injsev_Im_Binning \...	MinorInjury	MajorInjury
MinorInjury	4964	1951
MajorInjury	621	1239

Correct classified: 6,203 Wrong classified: 2,572

Accuracy: 70.689 % Error: 29.311 %

Cohen's kappa (κ) 0.304



MajorInjury (1/1)

Category	%	n
MinorInjury	0.0	0
MajorInjury	100.0	1
Total	100.0	1

MinorInjury (1/1)

Category	%	n
MinorInjury	100.0	1
MajorInjury	0.0	0
Total	100.0	1

MajorInjury (1/1)

Category	%	n
MinorInjury	0.0	0
MajorInjury	100.0	1
Total	100.0	1

MajorInjury (1/1)

Category	%	n
MinorInjury	0.0	0
MajorInjury	100.0	1
Total	100.0	1

MinorInjury (1/1)

Category	%	n
MinorInjury	100.0	1
MajorInjury	0.0	0
Total	100.0	1

MinorInjury (1/1)

Category	%	n
MinorInjury	100.0	1
MajorInjury	0.0	0
Total	100.0	1

MajorInjury (1/1)

Category	%	n
MinorInjury	0.0	0
MajorInjury	100.0	1
Total	100.0	1

Tree View:

IMPORTANT VARIABLES:

Rest_use_binning

V_ALCH_IM_N

DEFORMED_N

AGE_IM

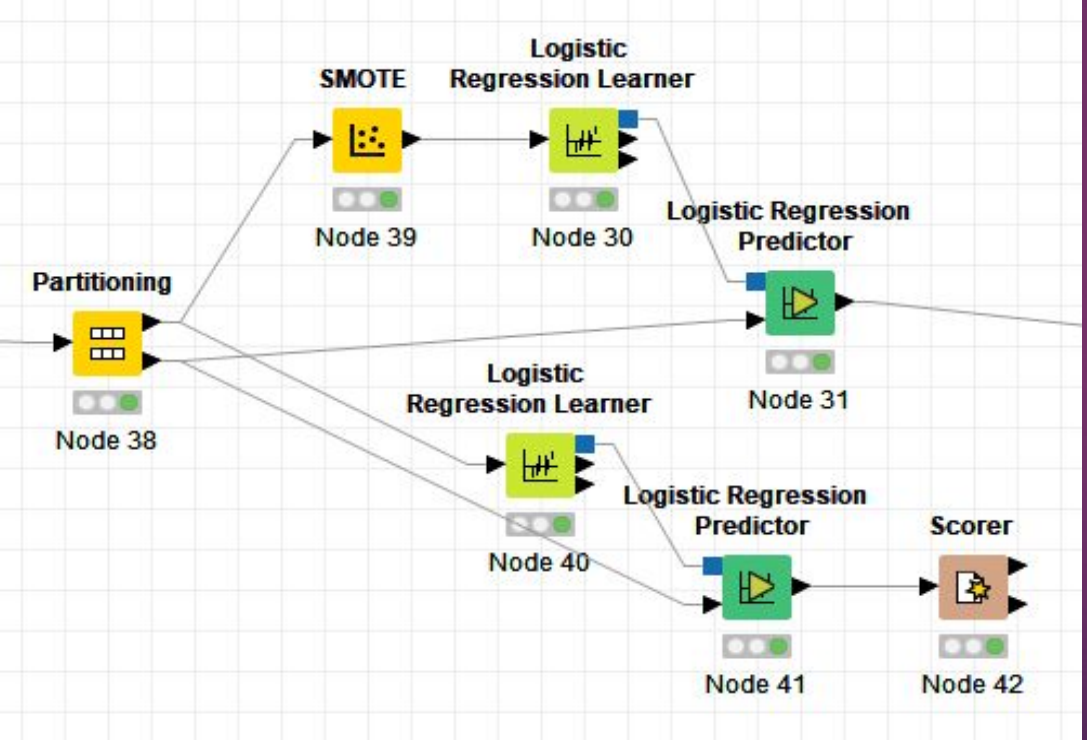
CARAGE

MANCOL_IM_B

BDYTYP_IM_N

Logistic Regression

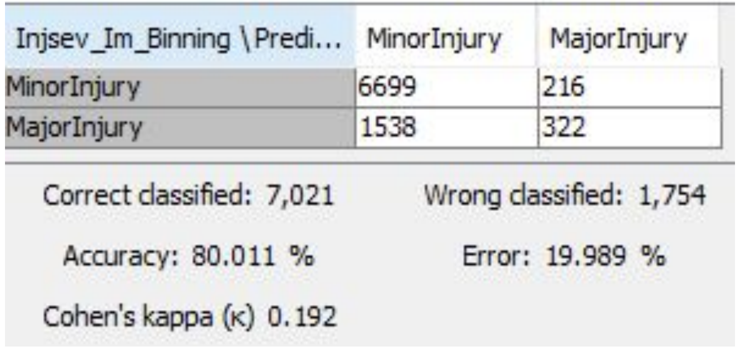
Model


SMOTE
Logistic Regression Learner
Node 39
Node 30
Logistic Regression Predictor
Partitioning
Node 38
Node 31
Logistic Regression Learner
Logistic Regression Predictor
Scorer
Node 40
Node 41
Node 42

- Confusion Matrix

Balancing Works!

Without Balancing:



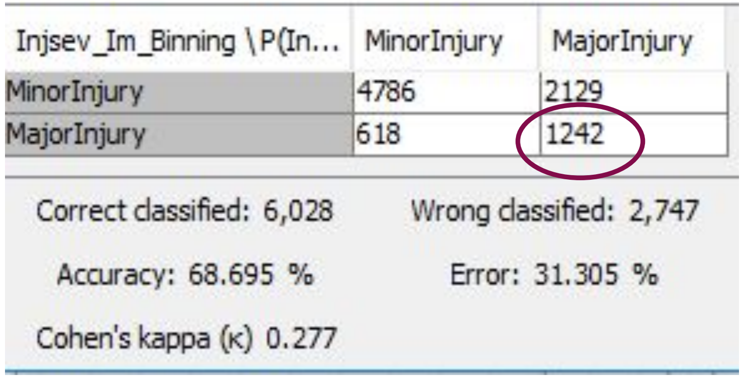
Injsev_Im_Binning \ Predi...	MinorInjury	MajorInjury
MinorInjury	6699	216
MajorInjury	1538	322

Correct classified: 7,021 Wrong classified: 1,754

Accuracy: 80.011 % Error: 19.989 %

Cohen's kappa (κ) 0.192

With Balancing (SMOTE):

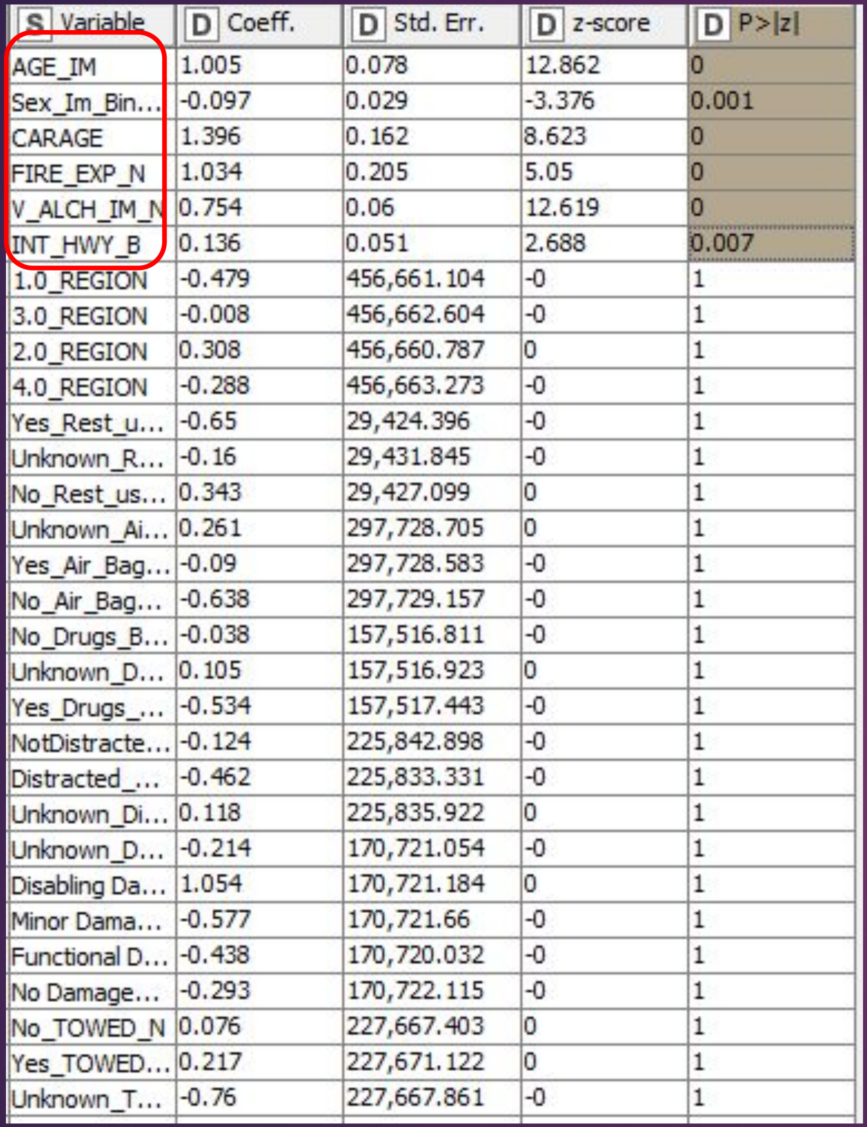


Injsev_Im_Binning \ P(In...	MinorInjury	MajorInjury
MinorInjury	4786	2129
MajorInjury	618	1242

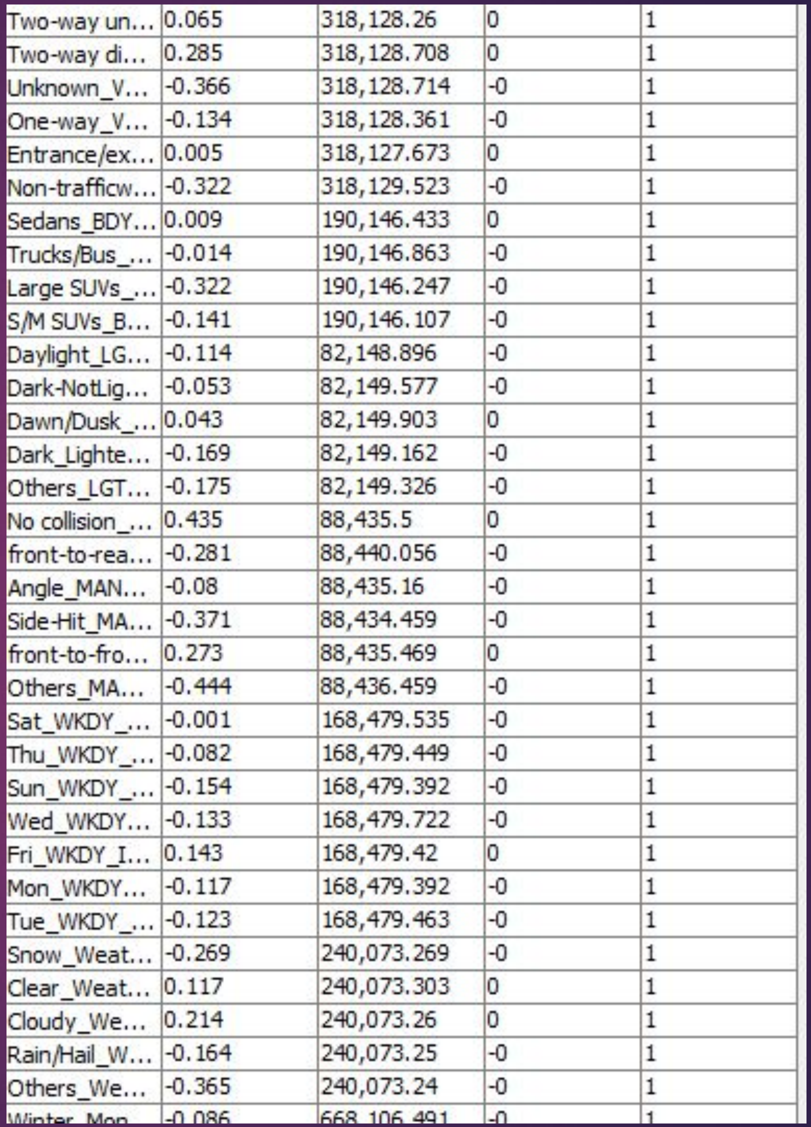
Correct classified: 6,028 Wrong classified: 2,747

Accuracy: 68.695 % Error: 31.305 %

Cohen's kappa (κ) 0.277



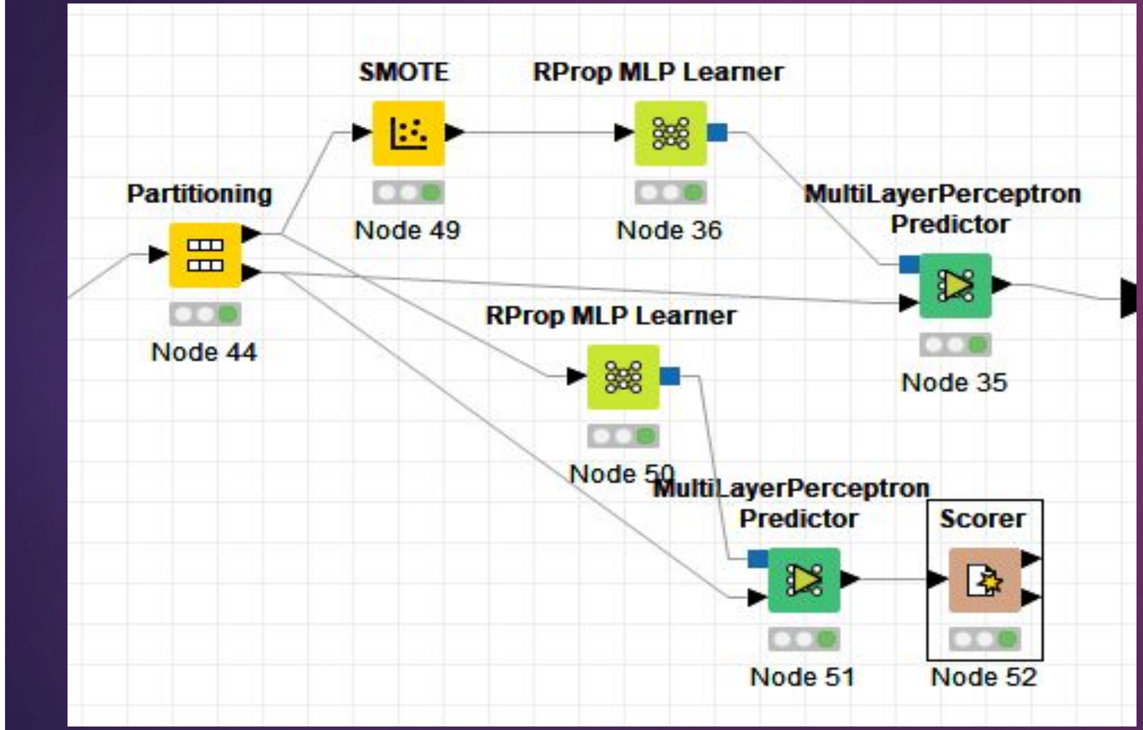
S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
AGE_IM	1.005	0.078	12.862	0
Sex_Im_Bin...	-0.097	0.029	-3.376	0.001
CARAGE	1.396	0.162	8.623	0
FIRE_EXP_N	1.034	0.205	5.05	0
V_ALCH_IM_N	0.754	0.06	12.619	0
INT_HWY_B	0.136	0.051	2.688	0.007
1.0_REGION	-0.479	456,661.104	-0	1
3.0_REGION	-0.008	456,662.604	-0	1
2.0_REGION	0.308	456,660.787	0	1
4.0_REGION	-0.288	456,663.273	-0	1
Yes_Rest_u...	-0.65	29,424.396	-0	1
Unknown_R...	-0.16	29,431.845	-0	1
No_Rest_us...	0.343	29,427.099	0	1
Unknown_Ai...	0.261	297,728.705	0	1
Yes_Air_Bag...	-0.09	297,728.583	-0	1
No_Air_Bag...	-0.638	297,729.157	-0	1
No_Drugs_B...	-0.038	157,516.811	-0	1
Unknown_D...	0.105	157,516.923	0	1
Yes_Drugs_...	-0.534	157,517.443	-0	1
NotDistracte...	-0.124	225,842.898	-0	1
Distracted_...	-0.462	225,833.331	-0	1
Unknown_Di...	0.118	225,835.922	0	1
Unknown_D...	-0.214	170,721.054	-0	1
Disabling Da...	1.054	170,721.184	0	1
Minor Dama...	-0.577	170,721.66	-0	1
Functional D...	-0.438	170,720.032	-0	1
No Damage...	-0.293	170,722.115	-0	1
No_TOWED_N	0.076	227,667.403	0	1
Yes_TOWED...	0.217	227,671.122	0	1
Unknown_T...	-0.76	227,667.861	-0	1



S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
Two-way un...	0.065	318,128.26	0	1
Two-way di...	0.285	318,128.708	0	1
Unknown_V...	-0.366	318,128.714	-0	1
One-way_V...	-0.134	318,128.361	-0	1
Entrance/ex...	0.005	318,127.673	0	1
Non-trafficw...	-0.322	318,129.523	-0	1
Sedans_BDY...	0.009	190,146.433	0	1
Trucks/Bus_...	-0.014	190,146.863	-0	1
Large SUVs_...	-0.322	190,146.247	-0	1
S/M SUVs_B...	-0.141	190,146.107	-0	1
Daylight_LG...	-0.114	82,148.896	-0	1
Dark-NotLig...	-0.053	82,149.577	-0	1
Dawn/Dusk_...	0.043	82,149.903	0	1
Dark_Lighte...	-0.169	82,149.162	-0	1
Others_LGT...	-0.175	82,149.326	-0	1
No collision_...	0.435	88,435.5	0	1
front-to-rea...	-0.281	88,440.056	-0	1
Angle_MAN...	-0.08	88,435.16	-0	1
Side-Hit_MA...	-0.371	88,434.459	-0	1
front-to-fro...	0.273	88,435.469	0	1
Others_MA...	-0.444	88,436.459	-0	1
Sat_WKDY_...	-0.001	168,479.535	-0	1
Thu_WKDY_...	-0.082	168,479.449	-0	1
Sun_WKDY_...	-0.154	168,479.392	-0	1
Wed_WKDY...	-0.133	168,479.722	-0	1
Fri_WKDY_I...	0.143	168,479.42	0	1
Mon_WKDY...	-0.117	168,479.392	-0	1
Tue_WKDY_...	-0.123	168,479.463	-0	1
Snow_Weat...	-0.269	240,073.269	-0	1
Clear_Weat...	0.117	240,073.303	0	1
Cloudy_We...	0.214	240,073.26	0	1
Rain/Hail_W...	-0.164	240,073.25	-0	1
Others_We...	-0.365	240,073.24	-0	1
Winter_Mon...	-0.086	668,106.491	-0	1

Neural Network:

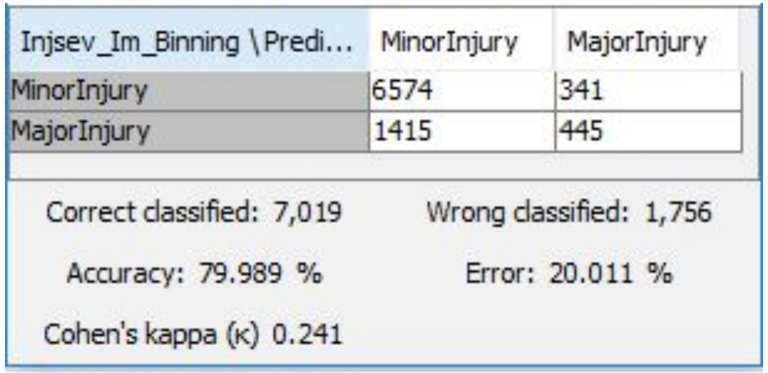
Model:



- Confusion Matrix

Balancing Works!

Without Balancing:



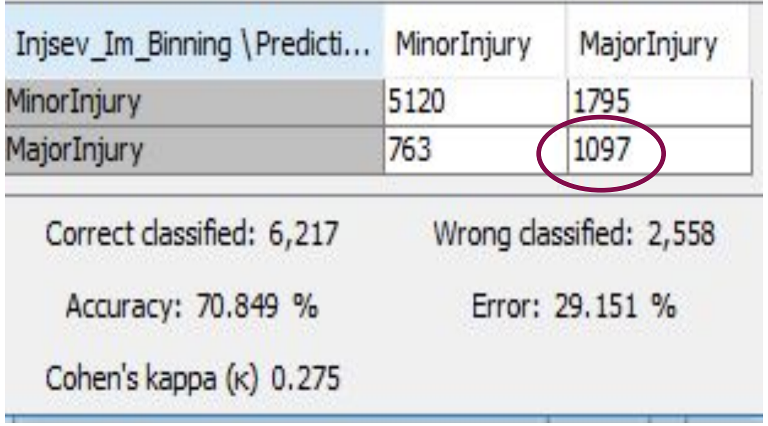
Injsev_Im_Binning \ Predi...	MinorInjury	MajorInjury
MinorInjury	6574	341
MajorInjury	1415	445

Correct classified: 7,019 Wrong classified: 1,756

Accuracy: 79.989 % Error: 20.011 %

Cohen's kappa (κ) 0.241

With Balancing (SMOTE):



Injsev_Im_Binning \ Predicti...	MinorInjury	MajorInjury
MinorInjury	5120	1795
MajorInjury	763	1097

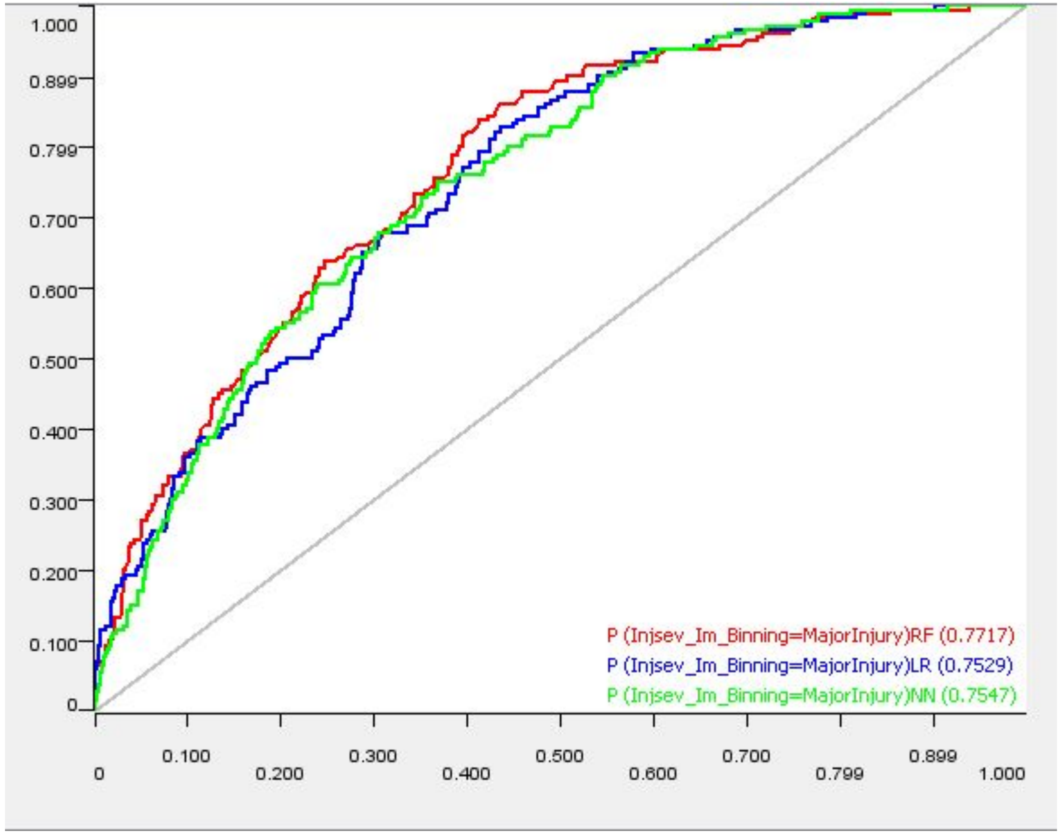
Correct classified: 6,217 Wrong classified: 2,558

Accuracy: 70.849 % Error: 29.151 %

Cohen's kappa (κ) 0.275

5. Model Evaluation/validation

- ROC Curve:



We could see from the ROC curves that all the three models performed well with our dataset. Random forest is slightly better compared to the other two models.

Accuracy Statistics

Sl No	Model (balanced)	Accuracy	Majority Class Correctly classified(Specificity)	Minority class correctly classified (Sensitivity)
1	Random Forest	70.7%	71.8%	66.6%
2	Logistic	68.7%	69.2%	66.8%
3	Neural Network	70.8%	74%	59%

We care severe injuries more. Sensitivity is critical!
Considering the total accuracy and sensitivity, random forest is the best method.

6. Deployment/suggestions

This study reveals that certain factors are highly related with severe car crash injuries

- airbags and safety restrictions (e.g. safety belts) usages
- car deformed or not in accidents
- driver drinking / taking drugs or not
- manner of collision
- fire occurrence
- light conditions
- other factors

We suggest drivers:

- Always use safety restrictions (e.g. safety belts) properly
- Select cars with air bags and make sure they are not malfunctioned
- Never drink alcohol or take certain drugs (e.g. making people drowsy) before driving
- Be cautious about vehicles too old
- Male, old-aged drivers please drive more carefully☺

6. Deployment/suggestions

We suggest car makers:

- Invest and build more robust cars
- Include air bags and other safety installments
- Develop intelligent driving systems to identify risks and avoid accidents

We suggest authorities:

- Enforce strict surveillance on drunk driving and dangerous driving (e.g. safety belt unbuckled up)
- Invest in education on safe driving and accident emergency treatment.

Questions Recommendations?

Thank you