Predicting injury severity using car crash data

MSIS5633 - Dr. Dursun Delen

Team 2
Hieu Nghiem
Rishi Poudyal
Manjusree Paimagham
Ahmed Sodeinde

Meet our team

Manjusree Paimagham

Rishi Poudyal

Hieu Nghiem

Ahmed Sodeinde

Business Understanding

54,969 accidents records
97,625 vehicles records
138,913 people records

Objective: **predict the injury severity of drivers in car crash whether it is low or high**

Dataset

2017 police-reported crashes data from CRSS (Crash Report Sampling System)

Injury Severity

Low: possible and minor injury

High: serious injury and fatal

Data Understanding

Accidents, vehicles and people data

Target variable: INJURY SEVERITY

Domain knowledge
Data description
Data distribution

24 predictors

Predictors:

REGION
URBANICITY
MONTH
DAY OF WEEK
HARMFUL EVENT
INTERSTATE HW
HOUR
MANNER OF COLLISION
LIGHT CONDITION
WEATHER
NUMBER OF OCCUPANTS
BODY TYPE
ROLLOVER
DEFORMED
SPEEDING RELATED
TRAFFICWAY DESC.
SURFACE CONDITION
VEHICLE AGE
INITIAL CONTACT POINT
NUMBER OF INJURIES
ALCOHOL INVOLVED
AIRBAG
SEX
AGE

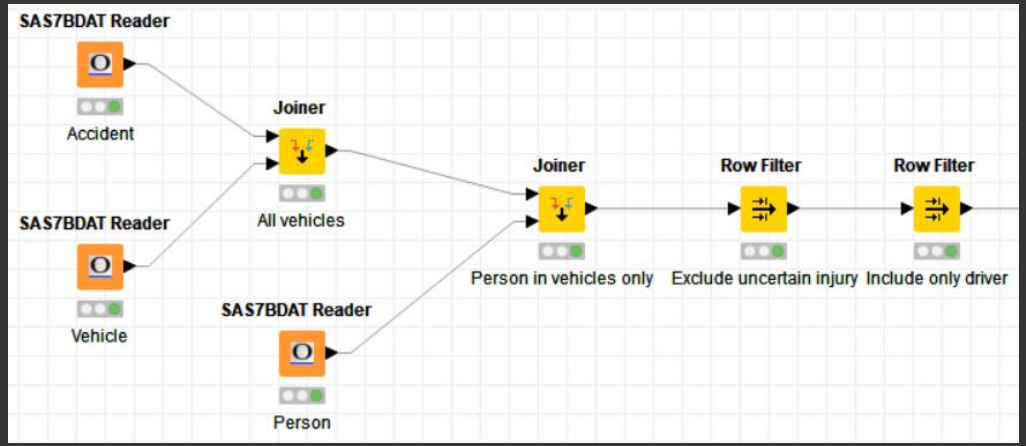
Data Preparation

Including all vehicles involved in crashes

Including only people in vehicles

Excluding uncertain / unknown injuries

Including only drivers



26,809 records

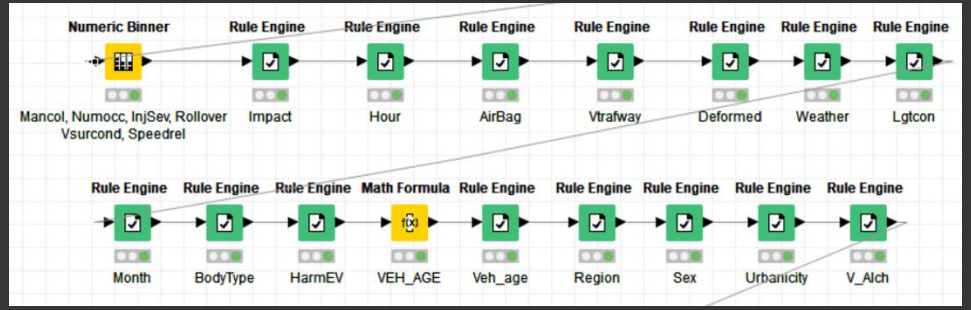
Data Preparation

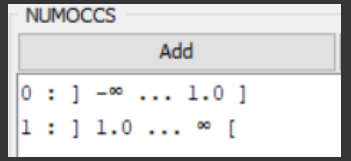
Binning the categorical variables, using the imputed variables.

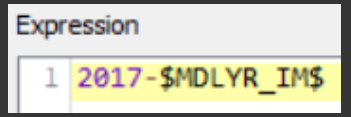
Calculating the age of the vehicles

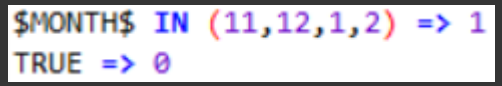
Categorizing variable levels as number: 0, 1, 2…

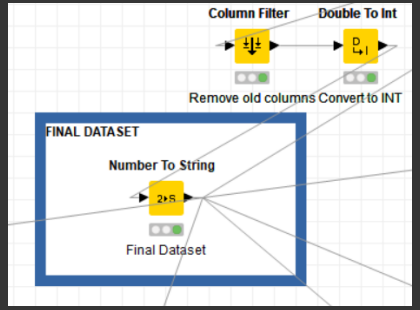
Convert number to string to make them nominal variables







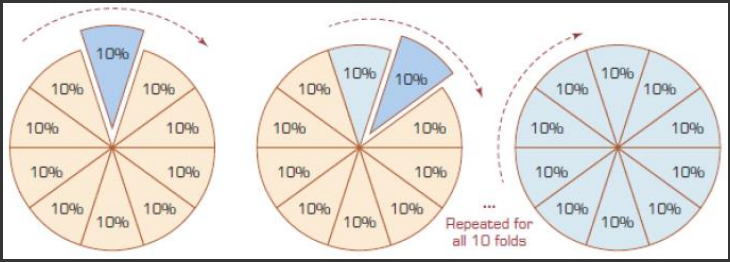




Modeling

Training and validation data:

10-Fold Cross Validation



Models chosen:

Decision Tree

Random Forest

Naïve Bayes

Logistic Regression

Artificial Neural Network

Support Vector Machine

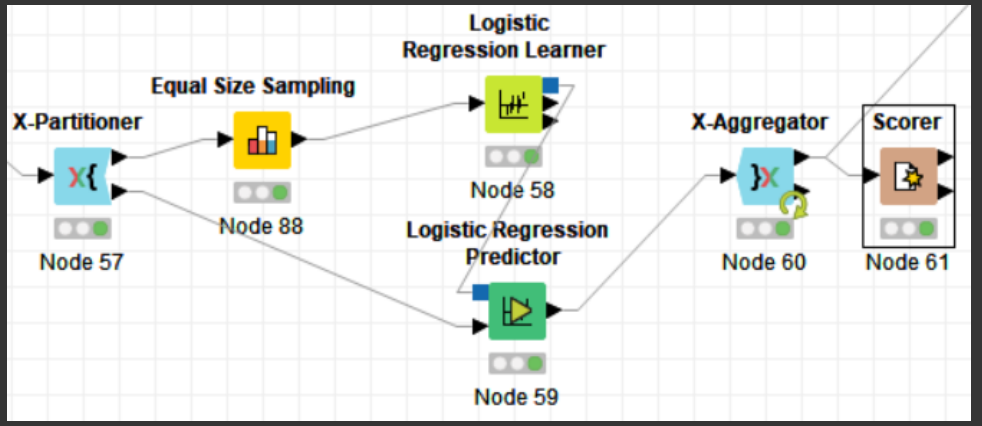
Decision Tree

Random Forest

Naïve Bayes

Logistic Regression

KNIME nodes:
X-Paritioner
Equal Size Sampling
Model Learner
Model Predictor
X-Aggregator

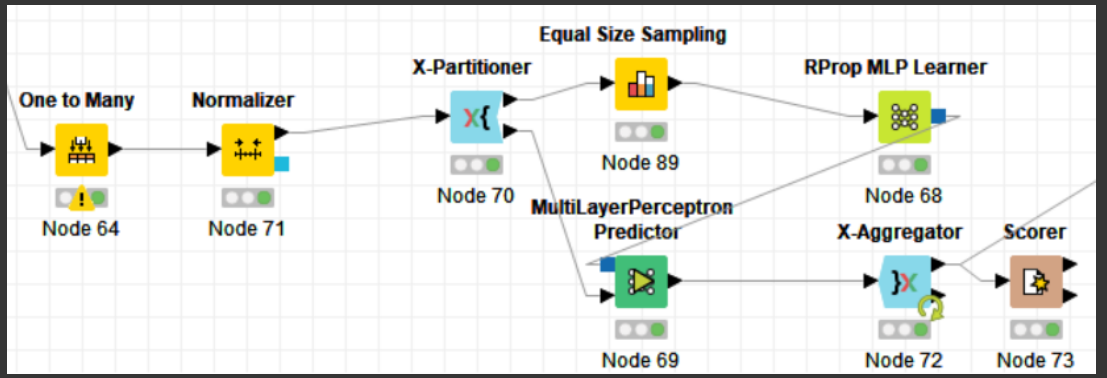


Artificial Neural Network

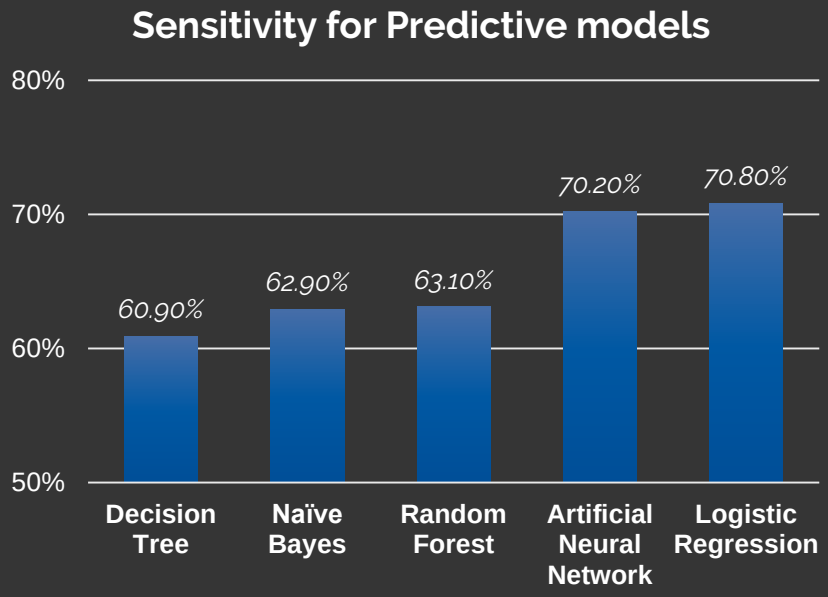
Must convert to numerical and normalize the data

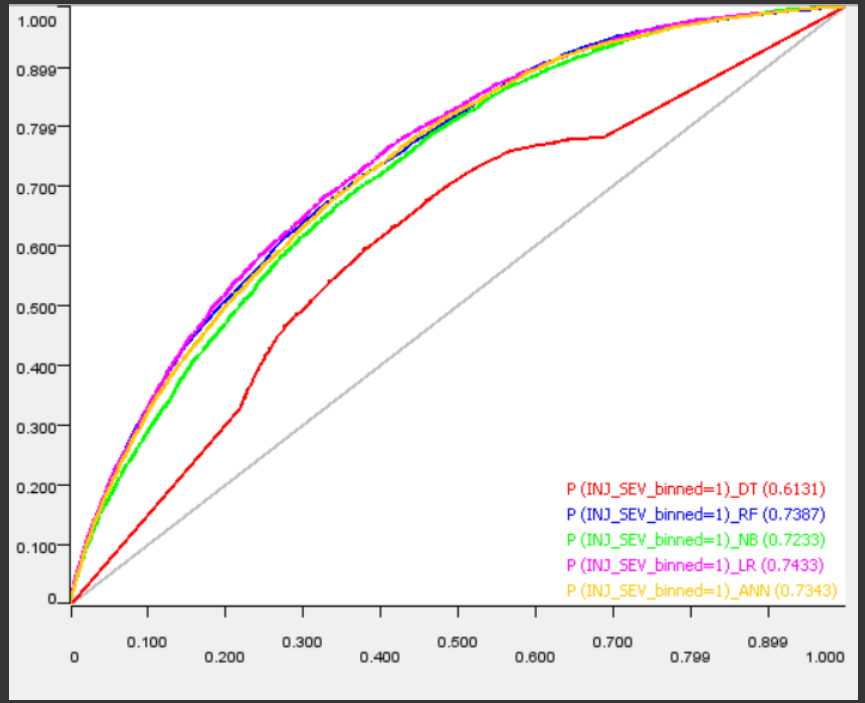
Support Vector Machine

6 hours run – no result!



Evaluation


Sensitivity for Predictive models
80%
70%
60%
50%
60.90%
62.90%
63.10%
70.20%
70.80%
Decision Tree
Naïve Bayes
Random Forest
Artificial Neural Network
Logistic Regression

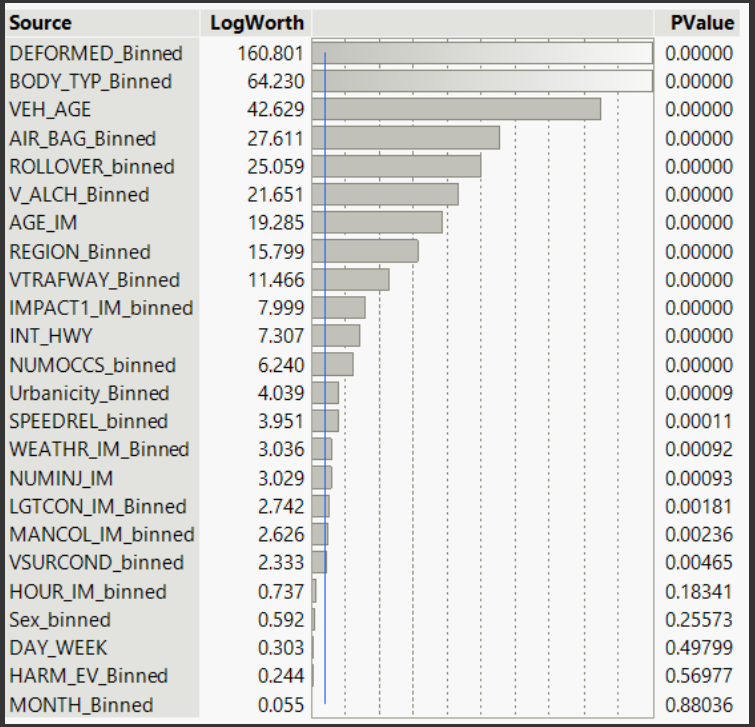

1.000
0.899
0.799
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0
0
0.100
0.200
0.300
0.400
0.500
0.600
0.700
0.799
0.899
1.000
P (INJ_SEV_binned=1)_DT (0.6131)
P (INJ_SEV_binned=1)_RF (0.7387)
P (INJ_SEV_binned=1)_NB (0.7233)
P (INJ_SEV_binned=1)_LR (0.7433)
P (INJ_SEV_binned=1)_ANN (0.7343)

Evaluation

Logistic Regression is chosen based on ROC curves and model accuracy

Most important variables are DEFORMED, BODY_TYPE and VEH_AGE

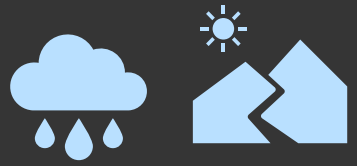
WEATHER and TIME OF DAY seem not affect injury severity



Source	LogWorth	PValue
DEFORMED_Binned	160.801	0.00000
BODY_TYP_Binned	64.230	0.00000
VEH_AGE	42.629	0.00000
AIR_BAG_Binned	27.611	0.00000
ROLLOVER_binned	25.059	0.00000
V_ALCH_Binned	21.651	0.00000
AGE_IM	19.285	0.00000
REGION_Binned	15.799	0.00000
VTRAFWAY_Binned	11.466	0.00000
IMPACT1_IM_binned	7.999	0.00000
INT_HWY	7.307	0.00000
NUMOCCS_binned	6.240	0.00000
Urbanicity_Binned	4.039	0.00009
SPEEDREL_binned	3.951	0.00011
WEATHR_IM_Binned	3.036	0.00092
NUMINJ_IM	3.029	0.00093
LGTCON_IM_Binned	2.742	0.00181
MANCOL_IM_binned	2.626	0.00236
VSURCOND_binned	2.333	0.00465
HOUR_IM_binned	0.737	0.18341
Sex_binned	0.592	0.25573
DAY_WEEK	0.303	0.49799
HARM_EV_Binned	0.244	0.56977
MONTH_Binned	0.055	0.88036

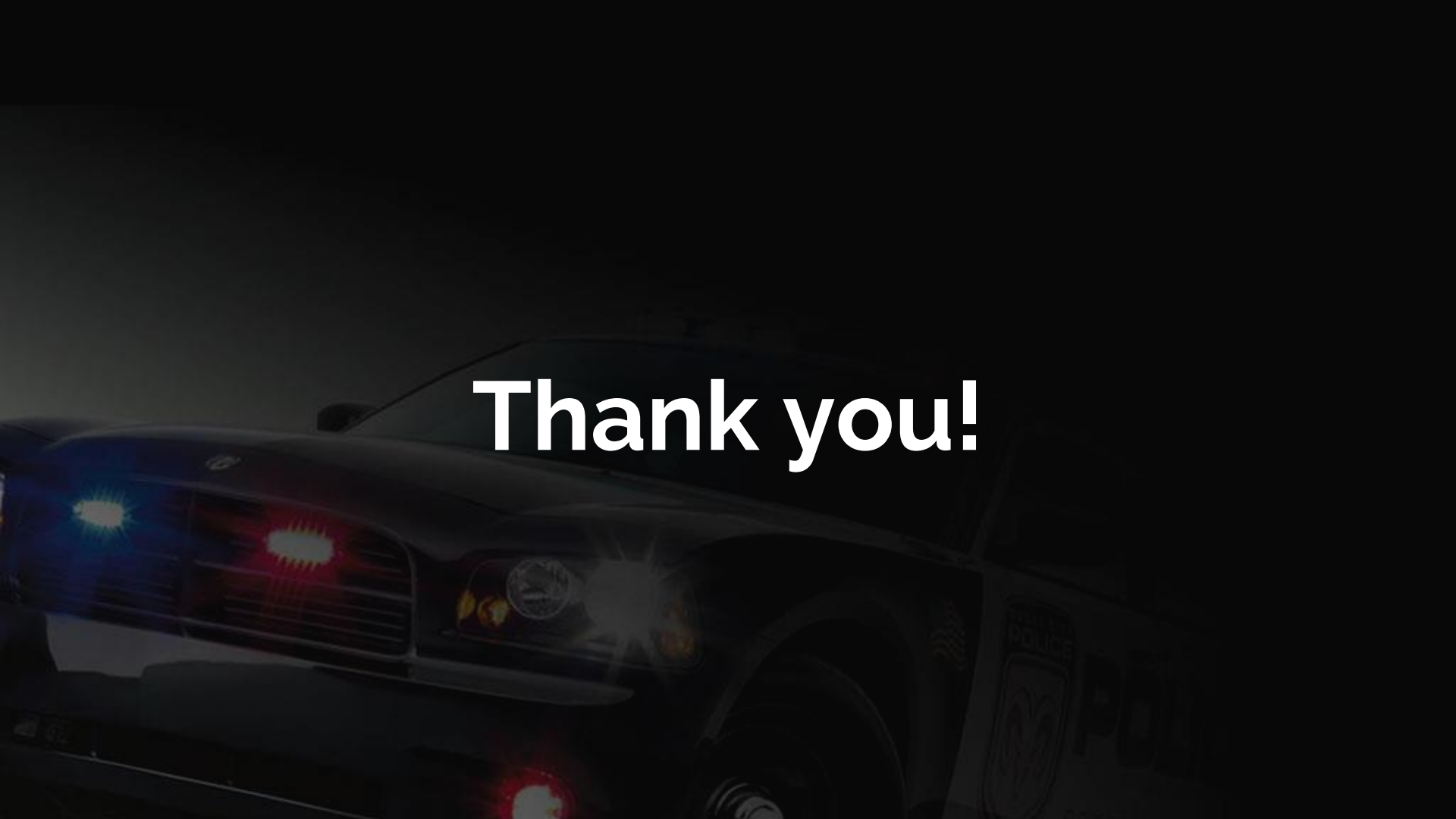
Deployment

Environment factors like weather, surface condition seem not affect injury severity



Vehicle factors like car age, body type, air bag seem to cause the most effects the risk of high severity

Propagation or law enforcement with cars have risk factors may cause high injury severity


Thank you!