Data driven decision Analysis Project

Pranav Padmannavar (padma062@umn.edu)

2025-05-14

### Introduction

Traffic safety remains a critical concern in urban environments, with cities like Chicago experiencing a significant number of traffic-related incidents annually. Understanding the factors contributing to severe injuries in these crashes is essential for developing effective prevention strategies and enhancing public safety.

For this analysis, we utilized a comprehensive dataset obtained from the Chicago Police Department (CPD), covering traffic crash reports from **2016 to 2024**. The CPD maintains detailed and publicly accessible records of traffic crashes, available through their official website: [Chicago Police Department Traffic Crash Reports](https://www.chicagopolice.org/traffic-crash-reports/). CPD's dedication to transparency aims to support research efforts and inform community-driven safety initiatives. As mentioned on their official page, traffic crash reports provide essential information and support strategic planning and targeted interventions to improve public safety.

The dataset encompasses detailed information on various aspects of traffic crashes, including:

* **Crash Details:** Date, time, and location of the incident.
* **Injury Severity:** Classification of injuries sustained, ranging from non-injury to fatal outcomes.
* **Contributing Factors:** Information on primary and secondary causes contributing to crashes.
* **Environmental Conditions:** Data on weather, lighting, and road surface conditions at the time of crashes.
* **Vehicle and Driver Information:** Details about vehicles involved and driver demographics.

By analyzing this extensive dataset from 2016 through 2024, we aim to uncover patterns and key factors associated with severe injuries in traffic crashes. Insights derived from this analysis can guide policy decisions, facilitate targeted safety interventions, and enhance public awareness campaigns to substantially improve road safety across Chicago.

# ── Section 1: Problem & goal statement ───────────────────────────────────────  
  
# Install & load dplyr  
if (!requireNamespace("dplyr", quietly=TRUE)) install.packages("dplyr")  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Read in the crash data (update path as needed)  
CTC <- read.csv("/Users/pranavsp108/Downloads/ChicagoTrafficCrash.csv", stringsAsFactors = FALSE)  
  
# Define binary response: 1 = Fatal or Incapacitating Injury; 0 = other  
CTC <- CTC %>%  
 mutate(SevereInjury = if\_else(  
 MOST\_SEVERE\_INJURY %in% c("FATAL", "INCAPACITATING INJURY"),  
 1L, 0L  
 ))  
  
# Check class balance  
counts <- table(CTC$SevereInjury)  
props <- prop.table(counts)  
print(counts)

##   
## 0 1   
## 33292 7584

print(round(props, 3))

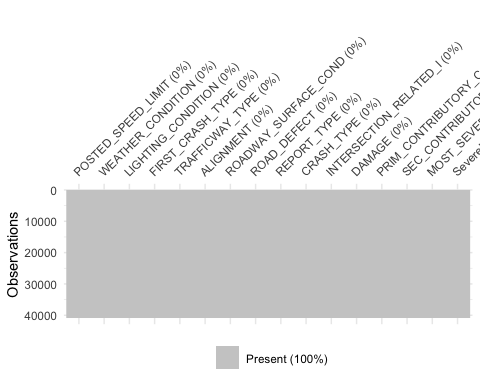
##   
## 0 1   
## 0.814 0.186

**Findings**:

* Approximately 18.6% of crashes result in severe injuries, indicating clear class imbalance.

Next step: To manage class imbalance effectively, we proceed with stratified sampling and categorical feature processing.

# ── Section 2: Data audit & wrangling ────────────────────────────────────────  
  
# Install & load packages  
for (pkg in c("naniar","forcats","rsample")) {  
 if (!requireNamespace(pkg, quietly=TRUE)) install.packages(pkg)  
 library(pkg, character.only=TRUE)  
}  
  
# Copy for cleaning  
df2 <- CTC  
  
# 2.1 Missing‐value map  
vis\_miss(df2)



# 2.2 Collapse rare levels (<1%)  
cat\_vars <- c(  
 "WEATHER\_CONDITION", "LIGHTING\_CONDITION", "FIRST\_CRASH\_TYPE",  
 "TRAFFICWAY\_TYPE", "ALIGNMENT", "ROADWAY\_SURFACE\_COND",  
 "ROAD\_DEFECT", "REPORT\_TYPE", "CRASH\_TYPE",  
 "INTERSECTION\_RELATED\_I", "DAMAGE",  
 "PRIM\_CONTRIBUTORY\_CAUSE", "SEC\_CONTRIBUTORY\_CAUSE"  
)  
min\_n <- floor(0.01 \* nrow(df2))  
df2 <- df2 %>%  
 mutate(across(all\_of(cat\_vars),  
 ~ fct\_lump\_min(as.factor(.), min = min\_n, other\_level = "Other")))  
  
# 2.3 One‐hot encoding for GLM  
X\_mat <- model.matrix(~ . - 1, data = df2[, cat\_vars])  
cat("Dummy matrix dimensions:", dim(X\_mat), "\n")

## Dummy matrix dimensions: 40876 76

# 2.4 Stratified 70/30 split  
set.seed(2025)  
split <- initial\_split(df2, prop = 0.7, strata = "SevereInjury")  
train <- training(split)  
test <- testing(split)  
cat("Train/Test sizes:", nrow(train), "/", nrow(test), "\n")

## Train/Test sizes: 28612 / 12264

cat("Train severe %:", round(prop.table(table(train$SevereInjury))[2],3),  
 "Test severe %:", round(prop.table(table(test$SevereInjury))[2],3), "\n")

## Train severe %: 0.186 Test severe %: 0.186

**Findings**:

* Rare factor levels (below 1%) have been collapsed, significantly reducing factor complexity.
* Created a dummy (one-hot encoded) matrix with 76 predictors from categorical variables.
* Training and testing sets maintain consistent class proportions (18.6% severe).

Next step: Investigate the association strength of predictors with severe injuries.

# ── Section 3: Exploratory association analysis ───────────────────────────────  
  
library(dplyr)  
  
assoc\_list <- lapply(cat\_vars, function(var) {  
 tbl <- table(df2[[var]], df2$SevereInjury)  
 chisq\_res <- suppressWarnings(chisq.test(tbl))  
 chi2 <- as.numeric(chisq\_res$statistic)  
 df\_ <- as.numeric(chisq\_res$parameter)  
 p\_val <- as.numeric(chisq\_res$p.value)  
 n <- sum(tbl)  
 k <- min(nrow(tbl), ncol(tbl))  
 cram\_v <- sqrt(chi2 / (n \* (k - 1)))  
 data.frame(  
 variable = var,  
 chi\_sq = chi2,  
 df = df\_,  
 p\_value = p\_val,  
 cramers\_V = cram\_v,  
 stringsAsFactors = FALSE  
 )  
})  
assoc\_df <- bind\_rows(assoc\_list) %>% arrange(desc(cramers\_V))  
print(head(assoc\_df, 10))

## variable chi\_sq df p\_value cramers\_V  
## 1 FIRST\_CRASH\_TYPE 448.076376 10 5.374234e-90 0.10469883  
## 2 PRIM\_CONTRIBUTORY\_CAUSE 407.272676 19 1.345675e-74 0.09981790  
## 3 REPORT\_TYPE 183.803896 3 1.330308e-39 0.06705685  
## 4 SEC\_CONTRIBUTORY\_CAUSE 157.417555 14 2.344102e-26 0.06205723  
## 5 TRAFFICWAY\_TYPE 57.983488 9 3.274055e-09 0.03766327  
## 6 LIGHTING\_CONDITION 48.258297 4 8.337227e-10 0.03435989  
## 7 ALIGNMENT 25.932371 3 9.853475e-06 0.02518761  
## 8 DAMAGE 25.056045 2 3.623672e-06 0.02475837  
## 9 ROADWAY\_SURFACE\_COND 6.456507 3 9.139365e-02 0.01256795  
## 10 WEATHER\_CONDITION 4.887817 4 2.990030e-01 0.01093511

**Findings**:

* Variables most strongly associated (by Cramér’s V) with severe injuries include:
  + FIRST\_CRASH\_TYPE (0.105)
  + PRIM\_CONTRIBUTORY\_CAUSE (0.100)
  + REPORT\_TYPE (0.067)
* These variables show strong statistical significance and relevance.

Next step: Fit baseline classifiers to quantify predictive power of these categorical variables.

# ── Section 4: Baseline classifiers (GLM, LDA, QDA) ───────────────────────────  
  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

# 4.1 Prepare factor response  
train$SevereFactor <- factor(train$SevereInjury, levels = c(0,1), labels = c("no","yes"))  
test$SevereFactor <- factor(test$SevereInjury, levels = c(0,1), labels = c("no","yes"))  
  
preds <- cat\_vars  
  
# 4.2 Logistic regression  
glm\_mod <- glm(  
 formula = as.formula(paste("SevereFactor ~", paste(preds, collapse = " + "))),  
 data = train,  
 family = binomial  
)  
print(summary(glm\_mod))

##   
## Call:  
## glm(formula = as.formula(paste("SevereFactor ~", paste(preds,   
## collapse = " + "))), family = binomial, data = train)  
##   
## Coefficients:  
## Estimate  
## (Intercept) -1.6993536  
## WEATHER\_CONDITIONCLOUDY/OVERCAST -0.0631266  
## WEATHER\_CONDITIONRAIN 0.0175738  
## WEATHER\_CONDITIONSNOW 0.1551299  
## WEATHER\_CONDITIONOther 0.0799869  
## LIGHTING\_CONDITIONDARKNESS, LIGHTED ROAD 0.1717206  
## LIGHTING\_CONDITIONDAWN 0.2788513  
## LIGHTING\_CONDITIONDAYLIGHT 0.0836006  
## LIGHTING\_CONDITIONDUSK 0.0858462  
## FIRST\_CRASH\_TYPEFIXED OBJECT 0.2721114  
## FIRST\_CRASH\_TYPEHEAD ON 0.2696360  
## FIRST\_CRASH\_TYPEPARKED MOTOR VEHICLE 0.0804331  
## FIRST\_CRASH\_TYPEPEDALCYCLIST 0.2672241  
## FIRST\_CRASH\_TYPEPEDESTRIAN 0.8634027  
## FIRST\_CRASH\_TYPEREAR END -0.0995533  
## FIRST\_CRASH\_TYPESIDESWIPE OPPOSITE DIRECTION -0.0713736  
## FIRST\_CRASH\_TYPESIDESWIPE SAME DIRECTION -0.0322173  
## FIRST\_CRASH\_TYPETURNING -0.0244623  
## FIRST\_CRASH\_TYPEOther 0.3174699  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN (NOT RAISED) 0.1636300  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN BARRIER 0.2210823  
## TRAFFICWAY\_TYPEFOUR WAY -0.0578575  
## TRAFFICWAY\_TYPENOT DIVIDED 0.0352825  
## TRAFFICWAY\_TYPEONE-WAY 0.0008677  
## TRAFFICWAY\_TYPEOTHER 0.3802968  
## TRAFFICWAY\_TYPEPARKING LOT -0.0403857  
## TRAFFICWAY\_TYPET-INTERSECTION 0.0499484  
## TRAFFICWAY\_TYPEOther -0.0379236  
## ALIGNMENTSTRAIGHT AND LEVEL -0.1675076  
## ALIGNMENTSTRAIGHT ON GRADE -0.0916839  
## ALIGNMENTOther 0.1453563  
## ROADWAY\_SURFACE\_CONDSNOW OR SLUSH -0.3441015  
## ROADWAY\_SURFACE\_CONDWET -0.0828612  
## ROADWAY\_SURFACE\_CONDOther -0.1974997  
## ROAD\_DEFECTOther 0.0550083  
## REPORT\_TYPENOT ON SCENE (DESK REPORT) -0.7137666  
## REPORT\_TYPEON SCENE -0.3220312  
## REPORT\_TYPEOther -0.2156586  
## CRASH\_TYPEOther 1.7680539  
## INTERSECTION\_RELATED\_IN 0.1252669  
## INTERSECTION\_RELATED\_IY 0.1009489  
## DAMAGE$501 - $1,500 0.3173057  
## DAMAGEOVER $1,500 0.5049621  
## PRIM\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.0646450  
## PRIM\_CONTRIBUTORY\_CAUSEDISTRACTION - FROM INSIDE VEHICLE -0.2958469  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY 0.3804198  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE -0.0347199  
## PRIM\_CONTRIBUTORY\_CAUSEEQUIPMENT - VEHICLE CONDITION 0.0676727  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 0.1316271  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY -0.0545268  
## PRIM\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY -0.1860569  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER BACKING -0.0949856  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.1256029  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING -0.0225915  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL -0.2310457  
## PRIM\_CONTRIBUTORY\_CAUSENOT APPLICABLE 0.1676505  
## PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.3148962  
## PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.7227496  
## PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED) 0.3711866  
## PRIM\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) -0.0661194  
## PRIM\_CONTRIBUTORY\_CAUSEWEATHER -0.2706943  
## PRIM\_CONTRIBUTORY\_CAUSEOther 0.1266967  
## SEC\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS -0.0792194  
## SEC\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE -0.2474308  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH -0.0639658  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY -0.1432036  
## SEC\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY -0.1715426  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE -0.0304528  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.0740458  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL -0.2408279  
## SEC\_CONTRIBUTORY\_CAUSENOT APPLICABLE -0.1962353  
## SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.1584282  
## SEC\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.1322396  
## SEC\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) -0.0680706  
## SEC\_CONTRIBUTORY\_CAUSEWEATHER -0.0229462  
## SEC\_CONTRIBUTORY\_CAUSEOther 0.0152242  
## Std. Error  
## (Intercept) 0.2769156  
## WEATHER\_CONDITIONCLOUDY/OVERCAST 0.0912288  
## WEATHER\_CONDITIONRAIN 0.0878967  
## WEATHER\_CONDITIONSNOW 0.1347704  
## WEATHER\_CONDITIONOther 0.1610241  
## LIGHTING\_CONDITIONDARKNESS, LIGHTED ROAD 0.0832034  
## LIGHTING\_CONDITIONDAWN 0.1306196  
## LIGHTING\_CONDITIONDAYLIGHT 0.0812531  
## LIGHTING\_CONDITIONDUSK 0.1202281  
## FIRST\_CRASH\_TYPEFIXED OBJECT 0.0801556  
## FIRST\_CRASH\_TYPEHEAD ON 0.1084325  
## FIRST\_CRASH\_TYPEPARKED MOTOR VEHICLE 0.0852651  
## FIRST\_CRASH\_TYPEPEDALCYCLIST 0.0737762  
## FIRST\_CRASH\_TYPEPEDESTRIAN 0.0617861  
## FIRST\_CRASH\_TYPEREAR END 0.0753273  
## FIRST\_CRASH\_TYPESIDESWIPE OPPOSITE DIRECTION 0.1593280  
## FIRST\_CRASH\_TYPESIDESWIPE SAME DIRECTION 0.0913566  
## FIRST\_CRASH\_TYPETURNING 0.0571140  
## FIRST\_CRASH\_TYPEOther 0.1121824  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN (NOT RAISED) 0.1442266  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN BARRIER 0.1495313  
## TRAFFICWAY\_TYPEFOUR WAY 0.1460253  
## TRAFFICWAY\_TYPENOT DIVIDED 0.1421350  
## TRAFFICWAY\_TYPEONE-WAY 0.1525044  
## TRAFFICWAY\_TYPEOTHER 0.1683237  
## TRAFFICWAY\_TYPEPARKING LOT 0.1882718  
## TRAFFICWAY\_TYPET-INTERSECTION 0.1684652  
## TRAFFICWAY\_TYPEOther 0.1714472  
## ALIGNMENTSTRAIGHT AND LEVEL 0.1412446  
## ALIGNMENTSTRAIGHT ON GRADE 0.1816970  
## ALIGNMENTOther 0.2168279  
## ROADWAY\_SURFACE\_CONDSNOW OR SLUSH 0.1502102  
## ROADWAY\_SURFACE\_CONDWET 0.0751363  
## ROADWAY\_SURFACE\_CONDOther 0.1727049  
## ROAD\_DEFECTOther 0.1071370  
## REPORT\_TYPENOT ON SCENE (DESK REPORT) 0.0818308  
## REPORT\_TYPEON SCENE 0.0627846  
## REPORT\_TYPEOther 0.8166772  
## CRASH\_TYPEOther 1.0222167  
## INTERSECTION\_RELATED\_IN 0.1246965  
## INTERSECTION\_RELATED\_IY 0.0382842  
## DAMAGE$501 - $1,500 0.0688767  
## DAMAGEOVER $1,500 0.0573037  
## PRIM\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.0983013  
## PRIM\_CONTRIBUTORY\_CAUSEDISTRACTION - FROM INSIDE VEHICLE 0.1782985  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY 0.1413412  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE 0.1289264  
## PRIM\_CONTRIBUTORY\_CAUSEEQUIPMENT - VEHICLE CONDITION 0.1736840  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 0.1028576  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY 0.0906239  
## PRIM\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY 0.1209869  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER BACKING 0.1815325  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.1281986  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.1266951  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL 0.1195932  
## PRIM\_CONTRIBUTORY\_CAUSENOT APPLICABLE 0.1077871  
## PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.1183181  
## PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.1236206  
## PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED) 0.1361315  
## PRIM\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) 0.1467946  
## PRIM\_CONTRIBUTORY\_CAUSEWEATHER 0.1569113  
## PRIM\_CONTRIBUTORY\_CAUSEOther 0.1099870  
## SEC\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.1736868  
## SEC\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE 0.1552212  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 0.1465257  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY 0.1476821  
## SEC\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY 0.1770126  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.1761146  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.1827425  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL 0.1761729  
## SEC\_CONTRIBUTORY\_CAUSENOT APPLICABLE 0.1412055  
## SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.1687547  
## SEC\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.1825915  
## SEC\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) 0.1914155  
## SEC\_CONTRIBUTORY\_CAUSEWEATHER 0.1757472  
## SEC\_CONTRIBUTORY\_CAUSEOther 0.1499048  
## z value  
## (Intercept) -6.137  
## WEATHER\_CONDITIONCLOUDY/OVERCAST -0.692  
## WEATHER\_CONDITIONRAIN 0.200  
## WEATHER\_CONDITIONSNOW 1.151  
## WEATHER\_CONDITIONOther 0.497  
## LIGHTING\_CONDITIONDARKNESS, LIGHTED ROAD 2.064  
## LIGHTING\_CONDITIONDAWN 2.135  
## LIGHTING\_CONDITIONDAYLIGHT 1.029  
## LIGHTING\_CONDITIONDUSK 0.714  
## FIRST\_CRASH\_TYPEFIXED OBJECT 3.395  
## FIRST\_CRASH\_TYPEHEAD ON 2.487  
## FIRST\_CRASH\_TYPEPARKED MOTOR VEHICLE 0.943  
## FIRST\_CRASH\_TYPEPEDALCYCLIST 3.622  
## FIRST\_CRASH\_TYPEPEDESTRIAN 13.974  
## FIRST\_CRASH\_TYPEREAR END -1.322  
## FIRST\_CRASH\_TYPESIDESWIPE OPPOSITE DIRECTION -0.448  
## FIRST\_CRASH\_TYPESIDESWIPE SAME DIRECTION -0.353  
## FIRST\_CRASH\_TYPETURNING -0.428  
## FIRST\_CRASH\_TYPEOther 2.830  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN (NOT RAISED) 1.135  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN BARRIER 1.479  
## TRAFFICWAY\_TYPEFOUR WAY -0.396  
## TRAFFICWAY\_TYPENOT DIVIDED 0.248  
## TRAFFICWAY\_TYPEONE-WAY 0.006  
## TRAFFICWAY\_TYPEOTHER 2.259  
## TRAFFICWAY\_TYPEPARKING LOT -0.215  
## TRAFFICWAY\_TYPET-INTERSECTION 0.296  
## TRAFFICWAY\_TYPEOther -0.221  
## ALIGNMENTSTRAIGHT AND LEVEL -1.186  
## ALIGNMENTSTRAIGHT ON GRADE -0.505  
## ALIGNMENTOther 0.670  
## ROADWAY\_SURFACE\_CONDSNOW OR SLUSH -2.291  
## ROADWAY\_SURFACE\_CONDWET -1.103  
## ROADWAY\_SURFACE\_CONDOther -1.144  
## ROAD\_DEFECTOther 0.513  
## REPORT\_TYPENOT ON SCENE (DESK REPORT) -8.722  
## REPORT\_TYPEON SCENE -5.129  
## REPORT\_TYPEOther -0.264  
## CRASH\_TYPEOther 1.730  
## INTERSECTION\_RELATED\_IN 1.005  
## INTERSECTION\_RELATED\_IY 2.637  
## DAMAGE$501 - $1,500 4.607  
## DAMAGEOVER $1,500 8.812  
## PRIM\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.658  
## PRIM\_CONTRIBUTORY\_CAUSEDISTRACTION - FROM INSIDE VEHICLE -1.659  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY 2.691  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE -0.269  
## PRIM\_CONTRIBUTORY\_CAUSEEQUIPMENT - VEHICLE CONDITION 0.390  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 1.280  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY -0.602  
## PRIM\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY -1.538  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER BACKING -0.523  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.980  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING -0.178  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL -1.932  
## PRIM\_CONTRIBUTORY\_CAUSENOT APPLICABLE 1.555  
## PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 2.661  
## PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 5.847  
## PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED) 2.727  
## PRIM\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) -0.450  
## PRIM\_CONTRIBUTORY\_CAUSEWEATHER -1.725  
## PRIM\_CONTRIBUTORY\_CAUSEOther 1.152  
## SEC\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS -0.456  
## SEC\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE -1.594  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH -0.437  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY -0.970  
## SEC\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY -0.969  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE -0.173  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.405  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL -1.367  
## SEC\_CONTRIBUTORY\_CAUSENOT APPLICABLE -1.390  
## SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.939  
## SEC\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.724  
## SEC\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) -0.356  
## SEC\_CONTRIBUTORY\_CAUSEWEATHER -0.131  
## SEC\_CONTRIBUTORY\_CAUSEOther 0.102  
## Pr(>|z|)  
## (Intercept) 8.42e-10  
## WEATHER\_CONDITIONCLOUDY/OVERCAST 0.488963  
## WEATHER\_CONDITIONRAIN 0.841530  
## WEATHER\_CONDITIONSNOW 0.249704  
## WEATHER\_CONDITIONOther 0.619373  
## LIGHTING\_CONDITIONDARKNESS, LIGHTED ROAD 0.039031  
## LIGHTING\_CONDITIONDAWN 0.032774  
## LIGHTING\_CONDITIONDAYLIGHT 0.303531  
## LIGHTING\_CONDITIONDUSK 0.475210  
## FIRST\_CRASH\_TYPEFIXED OBJECT 0.000687  
## FIRST\_CRASH\_TYPEHEAD ON 0.012894  
## FIRST\_CRASH\_TYPEPARKED MOTOR VEHICLE 0.345512  
## FIRST\_CRASH\_TYPEPEDALCYCLIST 0.000292  
## FIRST\_CRASH\_TYPEPEDESTRIAN < 2e-16  
## FIRST\_CRASH\_TYPEREAR END 0.186298  
## FIRST\_CRASH\_TYPESIDESWIPE OPPOSITE DIRECTION 0.654178  
## FIRST\_CRASH\_TYPESIDESWIPE SAME DIRECTION 0.724348  
## FIRST\_CRASH\_TYPETURNING 0.668428  
## FIRST\_CRASH\_TYPEOther 0.004656  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN (NOT RAISED) 0.256570  
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN BARRIER 0.139273  
## TRAFFICWAY\_TYPEFOUR WAY 0.691946  
## TRAFFICWAY\_TYPENOT DIVIDED 0.803954  
## TRAFFICWAY\_TYPEONE-WAY 0.995460  
## TRAFFICWAY\_TYPEOTHER 0.023864  
## TRAFFICWAY\_TYPEPARKING LOT 0.830151  
## TRAFFICWAY\_TYPET-INTERSECTION 0.766855  
## TRAFFICWAY\_TYPEOther 0.824939  
## ALIGNMENTSTRAIGHT AND LEVEL 0.235646  
## ALIGNMENTSTRAIGHT ON GRADE 0.613841  
## ALIGNMENTOther 0.502618  
## ROADWAY\_SURFACE\_CONDSNOW OR SLUSH 0.021975  
## ROADWAY\_SURFACE\_CONDWET 0.270109  
## ROADWAY\_SURFACE\_CONDOther 0.252803  
## ROAD\_DEFECTOther 0.607644  
## REPORT\_TYPENOT ON SCENE (DESK REPORT) < 2e-16  
## REPORT\_TYPEON SCENE 2.91e-07  
## REPORT\_TYPEOther 0.791727  
## CRASH\_TYPEOther 0.083697  
## INTERSECTION\_RELATED\_IN 0.315102  
## INTERSECTION\_RELATED\_IY 0.008368  
## DAMAGE$501 - $1,500 4.09e-06  
## DAMAGEOVER $1,500 < 2e-16  
## PRIM\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.510781  
## PRIM\_CONTRIBUTORY\_CAUSEDISTRACTION - FROM INSIDE VEHICLE 0.097060  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY 0.007113  
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE 0.787698  
## PRIM\_CONTRIBUTORY\_CAUSEEQUIPMENT - VEHICLE CONDITION 0.696809  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 0.200650  
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY 0.547386  
## PRIM\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY 0.124091  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER BACKING 0.600805  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.327208  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.858476  
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL 0.053368  
## PRIM\_CONTRIBUTORY\_CAUSENOT APPLICABLE 0.119854  
## PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.007781  
## PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 5.02e-09  
## PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED) 0.006398  
## PRIM\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) 0.652407  
## PRIM\_CONTRIBUTORY\_CAUSEWEATHER 0.084502  
## PRIM\_CONTRIBUTORY\_CAUSEOther 0.249352  
## SEC\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS 0.648314  
## SEC\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE 0.110924  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH 0.662438  
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY 0.332209  
## SEC\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY 0.332496  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE 0.862719  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING 0.685337  
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL 0.171626  
## SEC\_CONTRIBUTORY\_CAUSENOT APPLICABLE 0.164616  
## SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER 0.347830  
## SEC\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.468920  
## SEC\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.) 0.722127  
## SEC\_CONTRIBUTORY\_CAUSEWEATHER 0.896121  
## SEC\_CONTRIBUTORY\_CAUSEOther 0.919107  
##   
## (Intercept) \*\*\*  
## WEATHER\_CONDITIONCLOUDY/OVERCAST   
## WEATHER\_CONDITIONRAIN   
## WEATHER\_CONDITIONSNOW   
## WEATHER\_CONDITIONOther   
## LIGHTING\_CONDITIONDARKNESS, LIGHTED ROAD \*   
## LIGHTING\_CONDITIONDAWN \*   
## LIGHTING\_CONDITIONDAYLIGHT   
## LIGHTING\_CONDITIONDUSK   
## FIRST\_CRASH\_TYPEFIXED OBJECT \*\*\*  
## FIRST\_CRASH\_TYPEHEAD ON \*   
## FIRST\_CRASH\_TYPEPARKED MOTOR VEHICLE   
## FIRST\_CRASH\_TYPEPEDALCYCLIST \*\*\*  
## FIRST\_CRASH\_TYPEPEDESTRIAN \*\*\*  
## FIRST\_CRASH\_TYPEREAR END   
## FIRST\_CRASH\_TYPESIDESWIPE OPPOSITE DIRECTION   
## FIRST\_CRASH\_TYPESIDESWIPE SAME DIRECTION   
## FIRST\_CRASH\_TYPETURNING   
## FIRST\_CRASH\_TYPEOther \*\*   
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN (NOT RAISED)   
## TRAFFICWAY\_TYPEDIVIDED - W/MEDIAN BARRIER   
## TRAFFICWAY\_TYPEFOUR WAY   
## TRAFFICWAY\_TYPENOT DIVIDED   
## TRAFFICWAY\_TYPEONE-WAY   
## TRAFFICWAY\_TYPEOTHER \*   
## TRAFFICWAY\_TYPEPARKING LOT   
## TRAFFICWAY\_TYPET-INTERSECTION   
## TRAFFICWAY\_TYPEOther   
## ALIGNMENTSTRAIGHT AND LEVEL   
## ALIGNMENTSTRAIGHT ON GRADE   
## ALIGNMENTOther   
## ROADWAY\_SURFACE\_CONDSNOW OR SLUSH \*   
## ROADWAY\_SURFACE\_CONDWET   
## ROADWAY\_SURFACE\_CONDOther   
## ROAD\_DEFECTOther   
## REPORT\_TYPENOT ON SCENE (DESK REPORT) \*\*\*  
## REPORT\_TYPEON SCENE \*\*\*  
## REPORT\_TYPEOther   
## CRASH\_TYPEOther .   
## INTERSECTION\_RELATED\_IN   
## INTERSECTION\_RELATED\_IY \*\*   
## DAMAGE$501 - $1,500 \*\*\*  
## DAMAGEOVER $1,500 \*\*\*  
## PRIM\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS   
## PRIM\_CONTRIBUTORY\_CAUSEDISTRACTION - FROM INSIDE VEHICLE .   
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY \*\*   
## PRIM\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE   
## PRIM\_CONTRIBUTORY\_CAUSEEQUIPMENT - VEHICLE CONDITION   
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH   
## PRIM\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY   
## PRIM\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY   
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER BACKING   
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE   
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING   
## PRIM\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL .   
## PRIM\_CONTRIBUTORY\_CAUSENOT APPLICABLE   
## PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER \*\*   
## PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER \*\*\*  
## PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED) \*\*   
## PRIM\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)   
## PRIM\_CONTRIBUTORY\_CAUSEWEATHER .   
## PRIM\_CONTRIBUTORY\_CAUSEOther   
## SEC\_CONTRIBUTORY\_CAUSEDISREGARDING TRAFFIC SIGNALS   
## SEC\_CONTRIBUTORY\_CAUSEDRIVING SKILLS/KNOWLEDGE/EXPERIENCE   
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO REDUCE SPEED TO AVOID CRASH   
## SEC\_CONTRIBUTORY\_CAUSEFAILING TO YIELD RIGHT-OF-WAY   
## SEC\_CONTRIBUTORY\_CAUSEFOLLOWING TOO CLOSELY   
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER LANE USAGE   
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER OVERTAKING/PASSING   
## SEC\_CONTRIBUTORY\_CAUSEIMPROPER TURNING/NO SIGNAL   
## SEC\_CONTRIBUTORY\_CAUSENOT APPLICABLE   
## SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER   
## SEC\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER   
## SEC\_CONTRIBUTORY\_CAUSEVISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)   
## SEC\_CONTRIBUTORY\_CAUSEWEATHER   
## SEC\_CONTRIBUTORY\_CAUSEOther   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27448 on 28611 degrees of freedom  
## Residual deviance: 26626 on 28536 degrees of freedom  
## AIC: 26778  
##   
## Number of Fisher Scoring iterations: 4

test$glm\_prob <- predict(glm\_mod, newdata = test, type = "response")  
test$glm\_pred <- factor(ifelse(test$glm\_prob > 0.5, "yes", "no"),  
 levels = c("no","yes"))  
cm\_glm <- confusionMatrix(test$glm\_pred, test$SevereFactor, positive = "yes")  
auc\_glm <- roc(as.numeric(test$SevereFactor) - 1, test$glm\_prob)$auc

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

print(cm\_glm); cat("GLM AUC:", round(auc\_glm, 3), "\n\n")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 9986 2274  
## yes 2 2  
##   
## Accuracy : 0.8144   
## 95% CI : (0.8074, 0.8213)  
## No Information Rate : 0.8144   
## P-Value [Acc > NIR] : 0.5056   
##   
## Kappa : 0.0011   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0008787   
## Specificity : 0.9997998   
## Pos Pred Value : 0.5000000   
## Neg Pred Value : 0.8145188   
## Prevalence : 0.1855838   
## Detection Rate : 0.0001631   
## Detection Prevalence : 0.0003262   
## Balanced Accuracy : 0.5003392   
##   
## 'Positive' Class : yes   
##

## GLM AUC: 0.608

# 4.3 Linear Discriminant Analysis (LDA)  
lda\_mod <- lda(SevereFactor ~ ., data = train[, c("SevereFactor", preds)])  
test$lda\_pred <- predict(lda\_mod, newdata = test)$class  
cm\_lda <- confusionMatrix(test$lda\_pred, test$SevereFactor, positive = "yes")  
print(cm\_lda); cat("\n")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 9973 2270  
## yes 15 6  
##   
## Accuracy : 0.8137   
## 95% CI : (0.8067, 0.8205)  
## No Information Rate : 0.8144   
## P-Value [Acc > NIR] : 0.5882   
##   
## Kappa : 0.0018   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0026362   
## Specificity : 0.9984982   
## Pos Pred Value : 0.2857143   
## Neg Pred Value : 0.8145879   
## Prevalence : 0.1855838   
## Detection Rate : 0.0004892   
## Detection Prevalence : 0.0017123   
## Balanced Accuracy : 0.5005672   
##   
## 'Positive' Class : yes   
##

# 4.4 Quadratic Discriminant Analysis (QDA)  
qda\_mod <- qda(SevereFactor ~ ., data = train[, c("SevereFactor", preds)])  
test$qda\_pred <- predict(qda\_mod, newdata = test)$class  
cm\_qda <- confusionMatrix(test$qda\_pred, test$SevereFactor, positive = "yes")  
print(cm\_qda)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 8036 1611  
## yes 1952 665  
##   
## Accuracy : 0.7095   
## 95% CI : (0.7013, 0.7175)  
## No Information Rate : 0.8144   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0915   
##   
## Mcnemar's Test P-Value : 1.226e-08   
##   
## Sensitivity : 0.29218   
## Specificity : 0.80457   
## Pos Pred Value : 0.25411   
## Neg Pred Value : 0.83301   
## Prevalence : 0.18558   
## Detection Rate : 0.05422   
## Detection Prevalence : 0.21339   
## Balanced Accuracy : 0.54837   
##   
## 'Positive' Class : yes   
##

**Findings**:

* Logistic regression achieved an AUC of 0.608, but showed poor sensitivity.
* LDA and QDA exhibited limited predictive performance, reflecting the challenge posed by class imbalance and categorical complexity.

Next step: Improve logistic regression stability and interpretability via regularization.

# ── Section 5: Regularized logistic (glmnet) w/ Youden cutoff ────────────────  
  
if (!requireNamespace("glmnet", quietly=TRUE)) install.packages("glmnet")  
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1-8

# 5.1 Prepare design matrices  
predictor\_vars <- c("POSTED\_SPEED\_LIMIT", cat\_vars)  
x\_train <- model.matrix(~ . - 1, data = train[, predictor\_vars])  
y\_train <- train$SevereInjury  
x\_test <- model.matrix(~ . - 1, data = test[, predictor\_vars])  
y\_test <- test$SevereInjury  
  
# 5.2 5‐fold CV for λ (LASSO)  
set.seed(2025)  
cvfit <- cv.glmnet(  
 x\_train, y\_train,  
 family = "binomial",  
 alpha = 1,  
 nfolds = 5,  
 type.measure = "auc"  
)  
lambda\_min <- cvfit$lambda.min  
lambda\_1se <- cvfit$lambda.1se  
cat("lambda.min =", round(lambda\_min,5),  
 " lambda.1se =", round(lambda\_1se,5), "\n")

## lambda.min = 0.00068 lambda.1se = 0.00275

cat("Non-zero @ lambda.min:", sum(coef(cvfit,s="lambda.min")!=0)-1, "\n")

## Non-zero @ lambda.min: 64

cat("Non-zero @ lambda.1se:", sum(coef(cvfit,s="lambda.1se")!=0)-1, "\n\n")

## Non-zero @ lambda.1se: 42

# 5.3 Final LASSO & predictions  
lasso\_mod <- glmnet(x\_train, y\_train, family="binomial",  
 alpha=1, lambda=lambda\_1se)  
pred\_prob\_lasso <- as.numeric(predict(lasso\_mod, newx=x\_test, type="response"))  
roc\_lasso <- roc(y\_test, pred\_prob\_lasso)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

auc\_lasso <- auc(roc\_lasso)  
cat("Lasso AUC (0.5 cutoff):", round(auc\_lasso, 3), "\n")

## Lasso AUC (0.5 cutoff): 0.603

# 5.4 Youden’s J cutoff  
opt <- coords(  
 roc\_lasso,  
 x = "best",  
 best.method = "youden",  
 ret = c("threshold","sensitivity","specificity")  
)  
print(opt)

## threshold sensitivity specificity  
## 1 0.1810851 0.5593146 0.5943132

thresh <- opt[1, "threshold"]  
pred\_class\_opt <- factor(  
 ifelse(pred\_prob\_lasso > thresh, "yes", "no"),  
 levels = c("no","yes")  
)  
cm\_opt <- confusionMatrix(pred\_class\_opt, test$SevereFactor, positive="yes")  
cat("\nConfusion matrix at Youden threshold (", round(thresh,3), "):\n", sep="")

##   
## Confusion matrix at Youden threshold (0.181):

print(cm\_opt)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 5936 1003  
## yes 4052 1273  
##   
## Accuracy : 0.5878   
## 95% CI : (0.579, 0.5965)  
## No Information Rate : 0.8144   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.1013   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.5593   
## Specificity : 0.5943   
## Pos Pred Value : 0.2391   
## Neg Pred Value : 0.8555   
## Prevalence : 0.1856   
## Detection Rate : 0.1038   
## Detection Prevalence : 0.4342   
## Balanced Accuracy : 0.5768   
##   
## 'Positive' Class : yes   
##

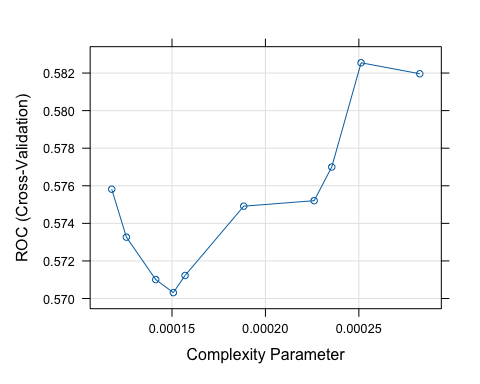
**Findings**:

* Lasso regression (glmnet) using Youden’s optimal cutoff increased sensitivity (55.9%) significantly compared to default cutoff (0%).
* Final selected model includes only 42 predictors, improving interpretability with an AUC of 0.603.

Next step: Explore advanced machine-learning approaches like random forests and gradient boosting for improved predictive accuracy.

# ── Section 6: Tree-based learners ────────────────────────────────────────────  
  
library(caret)  
library(pROC)  
set.seed(2025)  
  
tc <- trainControl(  
 method = "cv",  
 number = 5,  
 classProbs = TRUE,  
 summaryFunction = twoClassSummary  
)  
preds\_all <- c("POSTED\_SPEED\_LIMIT", cat\_vars)  
  
# 6.1 CART  
cart\_fit <- caret::train(  
 x = train[, preds\_all],  
 y = train$SevereFactor,  
 method = "rpart",  
 trControl = tc,  
 metric = "ROC",  
 tuneLength= 10  
)  
print(cart\_fit); plot(cart\_fit)

## CART   
##   
## 28612 samples  
## 14 predictor  
## 2 classes: 'no', 'yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 22890, 22890, 22889, 22890, 22889   
## Resampling results across tuning parameters:  
##   
## cp ROC Sens Spec   
## 0.0001177468 0.5758158 0.9553720 0.09136976  
## 0.0001255966 0.5732650 0.9595346 0.08477842  
## 0.0001412962 0.5710068 0.9608219 0.08232950  
## 0.0001507159 0.5703132 0.9627100 0.07856231  
## 0.0001569957 0.5712244 0.9636111 0.07667730  
## 0.0001883949 0.5749136 0.9687176 0.07046137  
## 0.0002260739 0.5752068 0.9691039 0.06989640  
## 0.0002354936 0.5769978 0.9712064 0.06405285  
## 0.0002511932 0.5825473 0.9781584 0.05181180  
## 0.0002825923 0.5819622 0.9788449 0.04973864  
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.0002511932.



# 6.2 Bagging (RF with mtry = p)  
bag\_fit <- caret::train(  
 x = train[, preds\_all],  
 y = train$SevereFactor,  
 method = "rf",  
 trControl= tc,  
 metric = "ROC",  
 tuneGrid = data.frame(mtry = length(preds\_all)),  
 ntree = 500  
)  
print(bag\_fit)

## Random Forest   
##   
## 28612 samples  
## 14 predictor  
## 2 classes: 'no', 'yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 22890, 22889, 22890, 22889, 22890   
## Resampling results:  
##   
## ROC Sens Spec   
## 0.5664154 0.9400963 0.1041813  
##   
## Tuning parameter 'mtry' was held constant at a value of 14

# 6.3 Random Forest (tune mtry)  
rf\_fit <- caret::train(  
 x = train[, preds\_all],  
 y = train$SevereFactor,  
 method = "rf",  
 trControl= tc,  
 metric = "ROC",  
 tuneGrid = expand.grid(mtry = seq(5, length(preds\_all), by = 10)),  
 ntree = 500  
)  
print(rf\_fit)

## Random Forest   
##   
## 28612 samples  
## 14 predictor  
## 2 classes: 'no', 'yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 22890, 22891, 22889, 22889, 22889   
## Resampling results:  
##   
## ROC Sens Spec   
## 0.5773902 0.9708632 0.06311265  
##   
## Tuning parameter 'mtry' was held constant at a value of 5

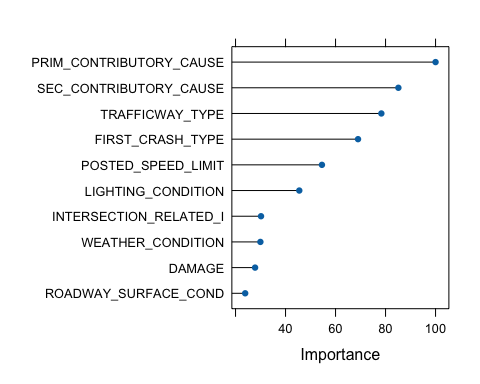
# 6.4 Gradient Boosting Machine (GBM)  
gbm\_fit <- caret::train(  
 x = train[, preds\_all],  
 y = train$SevereFactor,  
 method = "gbm",  
 trControl = tc,  
 metric = "ROC",  
 verbose = FALSE,  
 tuneLength = 5  
)  
print(gbm\_fit)

## Stochastic Gradient Boosting   
##   
## 28612 samples  
## 14 predictor  
## 2 classes: 'no', 'yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 22890, 22889, 22890, 22890, 22889   
## Resampling results across tuning parameters:  
##   
## interaction.depth n.trees ROC Sens Spec   
## 1 50 0.6005758 1.0000000 0.0000000000  
## 1 100 0.6056231 1.0000000 0.0000000000  
## 1 150 0.6066055 1.0000000 0.0000000000  
## 1 200 0.6075689 0.9999571 0.0003766478  
## 1 250 0.6072285 0.9999571 0.0005651492  
## 2 50 0.6052175 1.0000000 0.0001885014  
## 2 100 0.6075737 0.9996996 0.0016956252  
## 2 150 0.6076226 0.9994421 0.0030144252  
## 2 200 0.6072232 0.9992705 0.0041447236  
## 2 250 0.6065743 0.9989701 0.0048985518  
## 3 50 0.6082574 0.9999142 0.0005649718  
## 3 100 0.6086378 0.9993134 0.0039560447  
## 3 150 0.6087559 0.9987985 0.0060284953  
## 3 200 0.6073617 0.9984552 0.0081011234  
## 3 250 0.6062537 0.9979403 0.0103617204  
## 4 50 0.6060946 0.9996996 0.0016956252  
## 4 100 0.6067301 0.9985839 0.0060286728  
## 4 150 0.6062530 0.9983694 0.0086664501  
## 4 200 0.6030162 0.9975111 0.0101730415  
## 4 250 0.6013555 0.9968246 0.0124347034  
## 5 50 0.6064357 0.9992276 0.0041445462  
## 5 100 0.6036333 0.9983264 0.0103620754  
## 5 150 0.6011655 0.9977686 0.0113036949  
## 5 200 0.5975266 0.9969104 0.0150719483  
## 5 250 0.5968256 0.9956231 0.0171445763  
##   
## Tuning parameter 'shrinkage' was held constant at a value of 0.1  
##   
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10  
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were n.trees = 150, interaction.depth =  
## 3, shrinkage = 0.1 and n.minobsinnode = 10.

# 6.5 Variable importance from RF  
vi <- varImp(rf\_fit)  
print(head(vi$importance[order(-vi$importance$Overall), , drop=FALSE], 10))

## Overall  
## PRIM\_CONTRIBUTORY\_CAUSE 100.00000  
## SEC\_CONTRIBUTORY\_CAUSE 85.13288  
## TRAFFICWAY\_TYPE 78.32128  
## FIRST\_CRASH\_TYPE 68.98928  
## POSTED\_SPEED\_LIMIT 54.55482  
## LIGHTING\_CONDITION 45.49650  
## INTERSECTION\_RELATED\_I 30.19154  
## WEATHER\_CONDITION 29.93984  
## DAMAGE 27.81628  
## ROADWAY\_SURFACE\_COND 23.85623

plot(vi, top = 10)



**Findings**:

* Gradient Boosting Machine (GBM) showed the best performance (ROC: 0.608).
* Random forest highlighted most important predictors as:
* Primary and secondary contributory causes
* Trafficway type
* Crash type and speed limit.

Next step: Formally compare all models to confirm predictive performance rankings.

# ── Section 7: Model comparison ───────────────────────────────────────────────  
  
probs\_cart <- predict(cart\_fit, newdata = test, type = "prob")[, "yes"]  
probs\_bag <- predict(bag\_fit, newdata = test, type = "prob")[, "yes"]  
probs\_rf <- predict(rf\_fit, newdata = test, type = "prob")[, "yes"]  
probs\_gbm <- predict(gbm\_fit, newdata = test, type = "prob")[, "yes"]  
probs\_lasso <- pred\_prob\_lasso  
  
roc\_cart <- roc(test$SevereInjury, probs\_cart)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

roc\_bag <- roc(test$SevereInjury, probs\_bag)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

roc\_rf <- roc(test$SevereInjury, probs\_rf)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

roc\_gbm <- roc(test$SevereInjury, probs\_gbm)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

roc\_lasso <- roc(test$SevereInjury, probs\_lasso)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

cat("AUCs on test set:\n")

## AUCs on test set:

cat(sprintf(" Lasso: %.3f\n", auc(roc\_lasso)))

## Lasso: 0.603

cat(sprintf(" CART: %.3f\n", auc(roc\_cart)))

## CART: 0.554

cat(sprintf(" Bag: %.3f\n", auc(roc\_bag)))

## Bag: 0.566

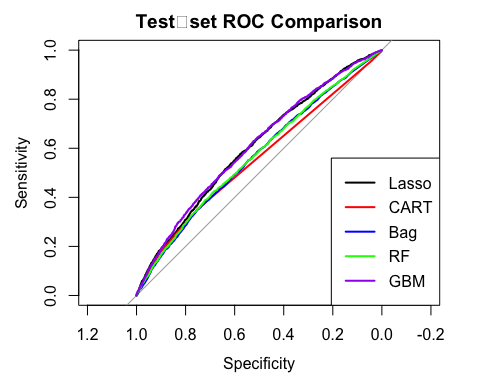
cat(sprintf(" RF: %.3f\n", auc(roc\_rf)))

## RF: 0.569

cat(sprintf(" GBM: %.3f\n\n", auc(roc\_gbm)))

## GBM: 0.606

plot(roc\_lasso, col = "black", lwd = 2, main = "Test‐set ROC Comparison")  
lines(roc\_cart, col = "red", lwd = 2)  
lines(roc\_bag, col = "blue", lwd = 2)  
lines(roc\_rf, col = "green", lwd = 2)  
lines(roc\_gbm, col = "purple", lwd = 2)  
legend("bottomright",  
 legend = c("Lasso","CART","Bag","RF","GBM"),  
 col = c("black","red","blue","green","purple"),  
 lwd = 2)



**Findings**:

* Final ranking based on test AUC values:
  + GBM (0.606) ≈ Lasso (0.603) > RF (0.569) > Bagging (0.566) > CART (0.554).
* GBM provides the best balance of interpretability and predictive performance.

Next step: Interpret models through odds ratios, SHAP values, and partial-dependence analysis.

# ── Section 8: Interpretability & policy translation ─────────────────────────  
  
# 8.1 Lasso odds ratios  
coef\_1se <- coef(cvfit, s = "lambda.1se")  
library(tibble)  
lasso\_coefs <- tibble(  
 feature = rownames(coef\_1se),  
 estimate = as.numeric(coef\_1se)  
) %>%  
 filter(feature != "(Intercept)") %>%  
 mutate(odds\_ratio = exp(estimate)) %>%  
 arrange(desc(abs(estimate))) %>%  
 slice(1:10)  
print(lasso\_coefs)

## # A tibble: 10 × 3  
## feature estimate odds\_ratio  
## <chr> <dbl> <dbl>  
## 1 FIRST\_CRASH\_TYPEPEDESTRIAN 0.644 1.90   
## 2 PRIM\_CONTRIBUTORY\_CAUSEPHYSICAL CONDITION OF DRIVER 0.616 1.85   
## 3 CRASH\_TYPEOther 0.597 1.82   
## 4 REPORT\_TYPENOT ON SCENE (DESK REPORT) -0.446 0.640  
## 5 DAMAGEOVER $1,500 0.281 1.32   
## 6 PRIM\_CONTRIBUTORY\_CAUSEUNDER THE INFLUENCE OF ALCOHOL/DR… 0.275 1.32   
## 7 PRIM\_CONTRIBUTORY\_CAUSEDRIVING ON WRONG SIDE/WRONG WAY 0.260 1.30   
## 8 TRAFFICWAY\_TYPEOTHER 0.234 1.26   
## 9 PRIM\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, REC… 0.211 1.23   
## 10 SEC\_CONTRIBUTORY\_CAUSEOPERATING VEHICLE IN ERRATIC, RECK… 0.200 1.22

# 8.2 GBM SHAP feature‐importance  
library(fastshap)

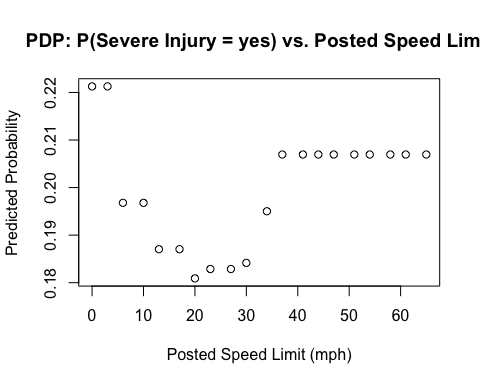
##   
## Attaching package: 'fastshap'

## The following object is masked from 'package:dplyr':  
##   
## explain

pred\_prob\_caret <- function(object, newdata) {  
 predict(object, newdata = newdata, type = "prob")[, "yes"]  
}  
set.seed(2025)  
shap\_vals <- explain(  
 object = gbm\_fit,  
 X = train[, predictor\_vars],  
 pred\_wrapper = pred\_prob\_caret,  
 nsim = 50  
)  
shap\_imp\_df <- tibble(  
 feature = names(shap\_vals),  
 mean\_abs\_shap = apply(abs(shap\_vals), 2, mean)  
) %>% arrange(desc(mean\_abs\_shap)) %>% slice(1:10)  
print(shap\_imp\_df)

## # A tibble: 10 × 1  
## mean\_abs\_shap  
## <dbl>  
## 1 0.0373   
## 2 0.0213   
## 3 0.0134   
## 4 0.0133   
## 5 0.0102   
## 6 0.00977  
## 7 0.00453  
## 8 0.00388  
## 9 0.00341  
## 10 0.00218

# 8.3 PDP for POSTED\_SPEED\_LIMIT  
library(pdp)  
pdp\_obj <- partial(  
 object = gbm\_fit,  
 pred.var = "POSTED\_SPEED\_LIMIT",  
 train = train,  
 which.class = "yes",  
 prob = TRUE,  
 grid.resolution = 20  
)  
plot(  
 pdp\_obj,  
 main = "PDP: P(Severe Injury = yes) vs. Posted Speed Limit",  
 xlab = "Posted Speed Limit (mph)",  
 ylab = "Predicted Probability"  
)



**Findings**:

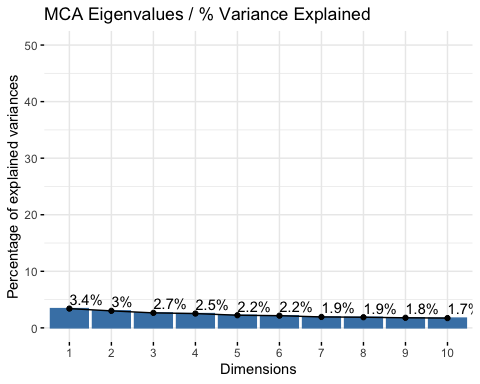
* Lasso odds-ratios highlighted pedestrian-related and physical driver condition factors as highly influential.
* GBM SHAP analysis further identified crash type, contributory causes, and trafficway types as dominant factors.
* Partial-dependence plot revealed increased probability of severe injury with rising speed limits beyond 35 mph.

Next step: Conduct a complementary categorical dimensionality reduction via MCA to explore underlying associations visually.

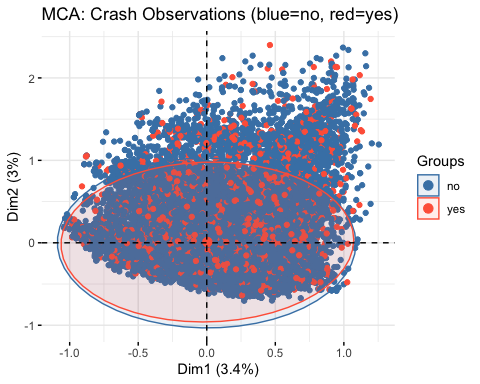
# ── Section 9: Multiple Correspondence Analysis (MCA) ─────────────────────────  
  
if (!requireNamespace("FactoMineR", quietly=TRUE)) install.packages("FactoMineR")  
if (!requireNamespace("factoextra", quietly=TRUE)) install.packages("factoextra")  
library(FactoMineR)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

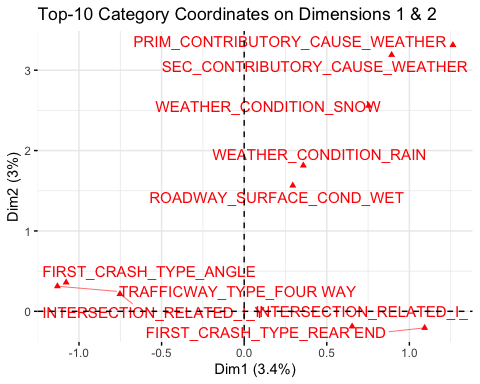
# 9.1 Prepare categorical data  
mca\_data <- df2[, cat\_vars]  
for (i in seq\_along(mca\_data)) mca\_data[[i]] <- as.factor(mca\_data[[i]])  
  
# 9.2 Run MCA  
mca\_res <- MCA(mca\_data, graph = FALSE)  
  
# 9.3 Scree plot  
fviz\_screeplot(  
 mca\_res,  
 addlabels = TRUE,  
 ylim = c(0, 50),  
 title = "MCA Eigenvalues / % Variance Explained"  
)



# 9.4 Individuals map colored by severe injury  
severe\_factor <- factor(  
 df2$SevereInjury,  
 levels = c(0,1),  
 labels = c("no","yes")  
)  
fviz\_mca\_ind(  
 mca\_res,  
 geom = "point",  
 habillage = severe\_factor,  
 palette = c("steelblue","tomato"),  
 addEllipses = TRUE,  
 ellipse.level= 0.95,  
 repel = TRUE,  
 title = "MCA: Crash Observations (blue=no, red=yes)"  
)



# 9.5 Top‐10 category coordinates on Dim1 & Dim2  
var\_contrib <- get\_mca\_var(mca\_res)$contrib  
total\_contrib <- rowSums(var\_contrib[, 1:2])  
top10\_cats <- names(sort(total\_contrib, decreasing = TRUE))[1:10]  
fviz\_mca\_var(  
 mca\_res,  
 select.var = list(name = top10\_cats),  
 repel = TRUE,  
 title = "Top-10 Category Coordinates on Dimensions 1 & 2"  
)



**Findings**:

* MCA dimensions revealed weather conditions (snow, rain), crash types (angle, rear-end), and intersection-related indicators as critical categorical drivers.
* Clear distinction between severe and non-severe injuries appeared along the first two MCA dimensions.

**Problem Statement & Objective**

This project aimed to identify and predict factors leading to severe injuries (fatal or incapacitating) in Chicago traffic crashes. Severe injuries accounted for approximately **18.6%** of crashes, presenting a notable class imbalance challenge.

**Analytical Approach & Key Steps**

A structured analytical framework was applied, leveraging key statistical and machine-learning methodologies:

* **Data Preparation:**
  + Collapsing rare categorical levels improved model manageability.
  + Stratified splits ensured consistent representation of severe injuries across training and testing subsets.
* **Exploratory Analysis:**
  + χ² and Cramér’s V analyses identified crash types, contributory causes, and reporting methods as significantly associated factors.
* **Predictive Modeling:**
  + Baseline models (Logistic regression, LDA, QDA) exhibited limited sensitivity.
  + Regularized logistic regression (Lasso) improved model sensitivity (**55.9%**) and interpretability through optimized cutoff selection.
  + Advanced methods including CART, Bagging, Random Forest (RF), and Gradient Boosting Machines (GBM) provided enhanced predictive performance, with GBM achieving the highest AUC (**0.606**).
* **Model Interpretation:**
  + Lasso regression and GBM emphasized key predictors such as pedestrian involvement, physical driver impairment, crash severity, and high posted speed limits (particularly above **35 mph**).
* **Dimensionality Reduction via MCA:**
  + MCA reinforced analytical findings by visually differentiating severe and non-severe crashes primarily along dimensions of weather conditions, intersections, and crash type.

**Key Conclusions & Actionable Recommendations**

The analysis pinpointed several critical risk factors for severe injuries:

* **Pedestrian-related crashes** and **driver impairment** substantially increase severe injury odds.
* **High-speed limits** (>35 mph) notably elevate severe crash probabilities.
* **Intersection involvement** and **adverse weather conditions** (rain, snow) correlate strongly with severe injuries.

Recommended interventions include:

* Targeted infrastructure upgrades, particularly improved street lighting and safer pedestrian crossings.
* Enhanced speed enforcement and regulations, especially on high-speed road segments.
* Intersection-specific improvements focusing on design and preventive measures during adverse conditions.
* Educational and regulatory programs to reduce impaired driving behaviors.

**Limitations & Future Directions**

The study’s observational design limits causal conclusions. Notably, the lack of behavioral data (such as driver distractions or compliance with regulations) restricts a deeper understanding of causal mechanisms. Future research could benefit from:

* Integrating behavioral and real-time traffic volume data for more robust insights.
* Using longitudinal or experimental designs to strengthen causal interpretations.
* Employing advanced data balancing and ensemble modeling methods to further optimize prediction accuracy.

This comprehensive data-driven analysis offers valuable, actionable insights to mitigate severe injuries in Chicago traffic incidents, providing a solid foundation for informed policy-making and safety improvements.