# Presentation Big Data

2024-03-23

### Crash injuries prediction model for the Virginia Department of Transporation

### Import libraries

```
# Load packages
library(foreign)
library(tidyverse)
library(margins)
library(ROCR)
library(caTools)
library(tree) # tree
library(rpart) # tree
library(rpart.plot) # tree plot
library(caret) #confusion matrix
library(e1071) #confusion matrix
library(ISLR)
library(MASS)
library(randomForest)
library(gbm)
library(ipred)
library(ggplot2)
library(hrbrthemes)
library(texreg)
```

### Import dataset and sample

```
# Import dataset
df <- read.dbf("C:/Users/usuario/Desktop/Masters Degree CEU/Big Data/Project/CrashData_Basic.dbf", as.i
```

### Data Tranformation

```
# Create the variable injured
df = df %>%
  mutate(injury = case_when(PERSONS_IN == 0 ~ "No",
                             PERSONS_IN >= 1 ~ "Yes"))
variable <- as.factor(df$injury)
df$injury <- variable
```

```r
# Select relevant variables
df <- subset(df, select = c(injury, K_PEOPLE, PERSONS_IN, CRASH_DT, VEH_COUNT, WEATHER_CO, LIGHT_COND,
         ALCOHOL_NO, BELTED_UNB, BIKE_NONBI, COLLISION_,
         DISTRACTED, ANIMAL, DROWSY_NOT, DRUG_NODRU, MOTOR_NONM,
         PED_NONPED, SPEED_NOTS, SENIOR_NOT, YOUNG_NOTY, OWNERSHIP))

# datetime modification
df <- df %>%
  mutate(year = substr(CRASH_DT, 1, 4),
         month = substr(CRASH_DT, 6, 7),
         day = substr(CRASH_DT, 9, 10))
df$year <- as.factor(df$year)
df$month <- as.factor(df$month)
df$day <- as.factor(df$day)

# Remove variables
df <- subset(df, select = -c(CRASH_DT, PERSONS_IN, day))

# Transform weather and light conditions
df <- df %>%
  filter(WEATHER_CO != 99)%>%
  filter(LIGHT_COND != 99)%>%
  filter(COLLISION_ != 99)

# class
summary(df)
```

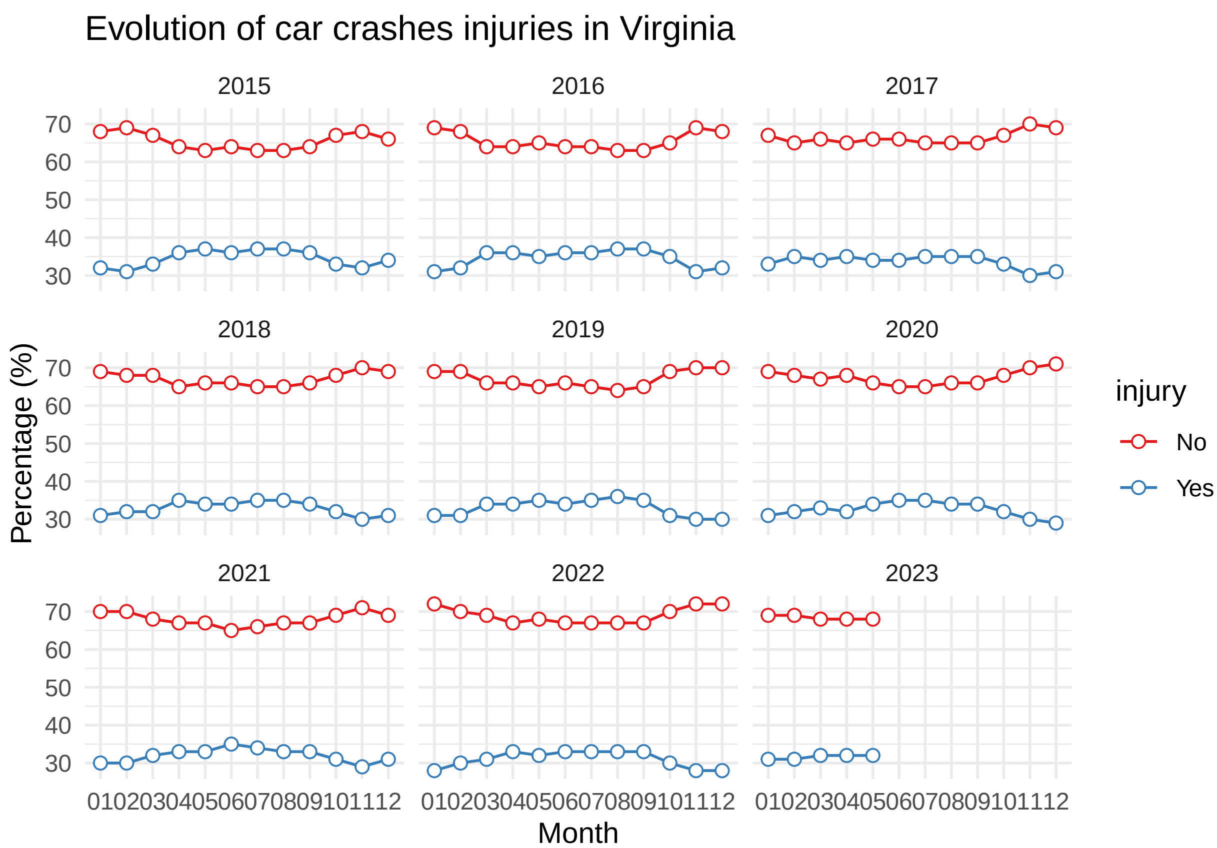
```
##  injury          K_PEOPLE          VEH_COUNT       WEATHER_CO     LIGHT_COND
##  No :695883   Min.   :0.000000   Min.   : 1.00   1      :858553   1: 28142
##  Yes:341762   1st Qu.:0.000000   1st Qu.: 1.00   5      :137029   2:681827
##               Median :0.000000   Median : 2.00   6      : 16170   3: 30134
##               Mean   :0.006935   Mean   : 1.84   4      : 14339   4:138477
##               3rd Qu.:0.000000   3rd Qu.: 2.00   3      :  5106   5:154692
##               Max.   :6.000000   Max.   :75.00   7      :  4161   6:  2480
##                                                  (Other):  2287   7:  1893
##  ALCOHOL_NO BELTED_UNB BIKE_NONBI    COLLISION_     DISTRACTED ANIMAL
##  0:978027   0:995399   0:1032508   1      :321315   0:841872   0:973587
##  1: 59618   1: 42246   1:   5137   2      :266381   1:195773   1: 64058
##                                    9      :202090
##                                    4      : 81706
##                                    10     : 48467
##                                    16     : 30590
##                                    (Other): 87096
##  DROWSY_NOT  DRUG_NODRU  MOTOR_NONM  PED_NONPED  SPEED_NOTS SENIOR_NOT
##  0:1008539   0:1028018   0:1021146   0:1024636   0:831504   0:866121
##  1:  29106   1:   9627   1:  16499   1:  13009   1:206141   1:171524
##
##
##
##
##
##  YOUNG_NOTY OWNERSHIP       year             month
##  0:843819   1:674176   2018   :131848   10     : 96649
```

```
##  1:193826   2: 32591   2016    :128525   11      : 94998
##             3:322013   2019    :128172   05      : 94257
##             4:  2274   2017    :127374   12      : 88742
##             5:  1623   2015    :125799   01      : 86583
##             6:  4968   2022    :122434   03      : 85106
##                       (Other):273493   (Other):491310
```

### Outcome variable analysis

## Data visualization

```r
agg1 <- df %>%
  group_by(year, month, injury) %>%
  summarise(n = n()) %>%
  mutate(prop = n/sum(n),
  pct = round((prop*100), 0))

ggplot(agg1, aes(x=month, y=pct, group = injury, colour = injury)) +
  geom_line()  +
  geom_point( size=2, shape=21, fill="white") +
  scale_color_brewer(palette = 'Set1') +
  theme_minimal() +
  labs(x = "Month",
       y = "Percentage (%)",
       title= "Evolution of car crashes injuries in Virginia")+
  facet_wrap(~ year)
```




## Injuries by year

```
table(df$injury, df$year)
```

```
##
##        2015  2016  2017  2018  2019  2020  2021  2022  2023
##   No  82558 84090 84629 88615 86004 71215 80537 84509 33726
##   Yes 43241 44435 42745 43233 42168 34385 37961 37925 15669
```

### Models Implementation

Split the data between train and test

```
# Split data
set.seed(321)
spl = sample.split(df$injury, SplitRatio = 0.7)
train = subset(df, spl==TRUE)
test = subset(df, spl==FALSE)
table(train$injury) #Check balance
```

```
##
```

```
##     No    Yes
## 487118 239233
```

```r
table(test$injury) #Check balance
```

```
##
##     No    Yes
## 208765 102529
```

Down sampling in the training data

```r
# Index of values with yes and no
Yes <- which(train$injury == "Yes")
No <- which(train$injury == "No")

# Sample the indices
downsample <- sample(No, length(Yes))
train <- train[c(downsample, Yes),]
table(train$injury)
```

```
##
##     No    Yes
## 239233 239233
```

## Logistic regression model

```r
# Fit the logistic regression model
set.seed(530)
mod1 <- glm(injury ~ VEH_COUNT + ALCOHOL_NO + BELTED_UNB + BIKE_NONBI +
            OWNERSHIP + COLLISION_ + WEATHER_CO + LIGHT_COND + DISTRACTED +
            ANIMAL + DROWSY_NOT + DRUG_NODRU + MOTOR_NONM + PED_NONPED +
            SPEED_NOTS + SENIOR_NOT + YOUNG_NOTY + year,
            data = train, family = binomial(link = "logit"), na.action = na.omit)
summary(mod1)
```

```
##
## Call:
## glm(formula = injury ~ VEH_COUNT + ALCOHOL_NO + BELTED_UNB +
##     BIKE_NONBI + OWNERSHIP + COLLISION_ + WEATHER_CO + LIGHT_COND +
##     DISTRACTED + ANIMAL + DROWSY_NOT + DRUG_NODRU + MOTOR_NONM +
##     PED_NONPED + SPEED_NOTS + SENIOR_NOT + YOUNG_NOTY + year,
##     family = binomial(link = "logit"), data = train, na.action = na.omit)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.844107   0.026963 -31.306  < 2e-16 ***
## VEH_COUNT     0.275860   0.006826  40.416  < 2e-16 ***
## ALCOHOL_NO1   0.201646   0.013889  14.519  < 2e-16 ***
## BELTED_UNB1   1.514568   0.017045  88.859  < 2e-16 ***
```

```
## BIKE_NONBI1   3.528301   0.101741  34.679  < 2e-16 ***
## OWNERSHIP2   -0.324353   0.018774 -17.277  < 2e-16 ***
## OWNERSHIP3    0.290796   0.007045  41.275  < 2e-16 ***
## OWNERSHIP4   -0.374365   0.071752  -5.217 1.81e-07 ***
## OWNERSHIP5   -0.256475   0.083383  -3.076 0.002099 **
## OWNERSHIP6   -0.301441   0.047965  -6.285 3.29e-10 ***
## COLLISION_10 -1.252551   0.037437 -33.457  < 2e-16 ***
## COLLISION_11 -0.673558   0.068179  -9.879  < 2e-16 ***
## COLLISION_12  0.068566   0.192664   0.356 0.721926
## COLLISION_13 -0.025265   0.770917  -0.033 0.973856
## COLLISION_14  0.887317   1.136638   0.781 0.435008
## COLLISION_15 -0.826139   0.045782 -18.045  < 2e-16 ***
## COLLISION_16  0.264604   0.019389  13.647  < 2e-16 ***
## COLLISION_2   0.168596   0.008192  20.580  < 2e-16 ***
## COLLISION_3   0.769758   0.021520  35.770  < 2e-16 ***
## COLLISION_4  -0.711835   0.013219 -53.848  < 2e-16 ***
## COLLISION_5  -0.100452   0.025086  -4.004 6.22e-05 ***
## COLLISION_6  -0.164787   0.034898  -4.722 2.34e-06 ***
## COLLISION_7  -0.078177   0.221845  -0.352 0.724543
## COLLISION_8   0.603194   0.026013  23.188  < 2e-16 ***
## COLLISION_9   0.267822   0.012584  21.283  < 2e-16 ***
## WEATHER_CO10 -0.056950   0.153859  -0.370 0.711274
## WEATHER_CO11  0.090991   0.149776   0.608 0.543507
## WEATHER_CO3   0.065266   0.045362   1.439 0.150215
## WEATHER_CO4   0.097151   0.026016   3.734 0.000188 ***
## WEATHER_CO5  -0.124028   0.009158 -13.543  < 2e-16 ***
## WEATHER_CO6  -0.525978   0.026326 -19.980  < 2e-16 ***
## WEATHER_CO7  -0.359690   0.049864  -7.213 5.45e-13 ***
## WEATHER_CO8   0.176022   0.421806   0.417 0.676454
## WEATHER_CO9   0.085922   0.083043   1.035 0.300824
## LIGHT_COND2   0.096748   0.019814   4.883 1.05e-06 ***
## LIGHT_COND3   0.082482   0.026447   3.119 0.001816 **
## LIGHT_COND4  -0.027505   0.021265  -1.293 0.195857
## LIGHT_COND5   0.047948   0.021250   2.256 0.024045 *
## LIGHT_COND6  -0.057427   0.067112  -0.856 0.392170
## LIGHT_COND7  -1.620316   0.115771 -13.996  < 2e-16 ***
## DISTRACTED1   0.027927   0.008114   3.442 0.000578 ***
## ANIMAL1      -0.039419   0.029355  -1.343 0.179326
## DROWSY_NOT1   0.250618   0.018947  13.227  < 2e-16 ***
## DRUG_NODRU1   0.530949   0.032076  16.553  < 2e-16 ***
## MOTOR_NONM1   2.293551   0.032732  70.070  < 2e-16 ***
## PED_NONPED1   3.785278   0.188873  20.041  < 2e-16 ***
## SPEED_NOTS1   0.130658   0.007948  16.439  < 2e-16 ***
## SENIOR_NOT1   0.187787   0.008331  22.541  < 2e-16 ***
## YOUNG_NOTY1  -0.072363   0.007888  -9.173  < 2e-16 ***
## year2016      0.019570   0.012265   1.596 0.110575
## year2017     -0.048657   0.012326  -3.948 7.90e-05 ***
## year2018     -0.049371   0.012210  -4.043 5.27e-05 ***
## year2019     -0.071428   0.012354  -5.782 7.38e-09 ***
## year2020     -0.096648   0.013030  -7.417 1.19e-13 ***
## year2021     -0.134715   0.012629 -10.667  < 2e-16 ***
## year2022     -0.158278   0.012567 -12.595  < 2e-16 ***
## year2023     -0.155289   0.016570  -9.372  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 663295  on 478465  degrees of freedom
## Residual deviance: 605620  on 478409  degrees of freedom
## AIC: 605734
##
## Number of Fisher Scoring iterations: 6
```

```r
screenreg(list(mod1))
```

```
##
## ==============================
##                 Model 1
## ------------------------------
## (Intercept)          -0.84 ***
##                      (0.03)
## VEH_COUNT             0.28 ***
##                      (0.01)
## ALCOHOL_NO1           0.20 ***
##                      (0.01)
## BELTED_UNB1           1.51 ***
##                      (0.02)
## BIKE_NONBI1           3.53 ***
##                      (0.10)
## OWNERSHIP2           -0.32 ***
##                      (0.02)
## OWNERSHIP3            0.29 ***
##                      (0.01)
## OWNERSHIP4           -0.37 ***
##                      (0.07)
## OWNERSHIP5           -0.26 **
##                      (0.08)
## OWNERSHIP6           -0.30 ***
##                      (0.05)
## COLLISION_10         -1.25 ***
##                      (0.04)
## COLLISION_11         -0.67 ***
##                      (0.07)
## COLLISION_12          0.07
##                      (0.19)
## COLLISION_13         -0.03
##                      (0.77)
## COLLISION_14          0.89
##                      (1.14)
## COLLISION_15         -0.83 ***
##                      (0.05)
## COLLISION_16          0.26 ***
##                      (0.02)
## COLLISION_2           0.17 ***
##                      (0.01)
## COLLISION_3           0.77 ***
##                      (0.02)
```

```
## COLLISION_4          -0.71 ***
##                      (0.01)
## COLLISION_5          -0.10 ***
##                      (0.03)
## COLLISION_6          -0.16 ***
##                      (0.03)
## COLLISION_7          -0.08
##                      (0.22)
## COLLISION_8           0.60 ***
##                      (0.03)
## COLLISION_9           0.27 ***
##                      (0.01)
## WEATHER_CO10         -0.06
##                      (0.15)
## WEATHER_CO11          0.09
##                      (0.15)
## WEATHER_CO3           0.07
##                      (0.05)
## WEATHER_CO4           0.10 ***
##                      (0.03)
## WEATHER_CO5          -0.12 ***
##                      (0.01)
## WEATHER_CO6          -0.53 ***
##                      (0.03)
## WEATHER_CO7          -0.36 ***
##                      (0.05)
## WEATHER_CO8           0.18
##                      (0.42)
## WEATHER_CO9           0.09
##                      (0.08)
## LIGHT_COND2           0.10 ***
##                      (0.02)
## LIGHT_COND3           0.08 **
##                      (0.03)
## LIGHT_COND4          -0.03
##                      (0.02)
## LIGHT_COND5           0.05 *
##                      (0.02)
## LIGHT_COND6          -0.06
##                      (0.07)
## LIGHT_COND7          -1.62 ***
##                      (0.12)
## DISTRACTED1           0.03 ***
##                      (0.01)
## ANIMAL1              -0.04
##                      (0.03)
## DROWSY_NOT1           0.25 ***
##                      (0.02)
## DRUG_NODRU1           0.53 ***
##                      (0.03)
## MOTOR_NONM1           2.29 ***
##                      (0.03)
## PED_NONPED1           3.79 ***
##                      (0.19)
```

```
## SPEED_NOTS1           0.13 ***
##                      (0.01)
## SENIOR_NOT1           0.19 ***
##                      (0.01)
## YOUNG_NOTY1          -0.07 ***
##                      (0.01)
## year2016              0.02
##                      (0.01)
## year2017             -0.05 ***
##                      (0.01)
## year2018             -0.05 ***
##                      (0.01)
## year2019             -0.07 ***
##                      (0.01)
## year2020             -0.10 ***
##                      (0.01)
## year2021             -0.13 ***
##                      (0.01)
## year2022             -0.16 ***
##                      (0.01)
## year2023             -0.16 ***
##                      (0.02)
## ------------------------------
## AIC              605734.28
## BIC              606365.74
## Log Likelihood  -302810.14
## Deviance         605620.28
## Num. obs.        478466
## ==============================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

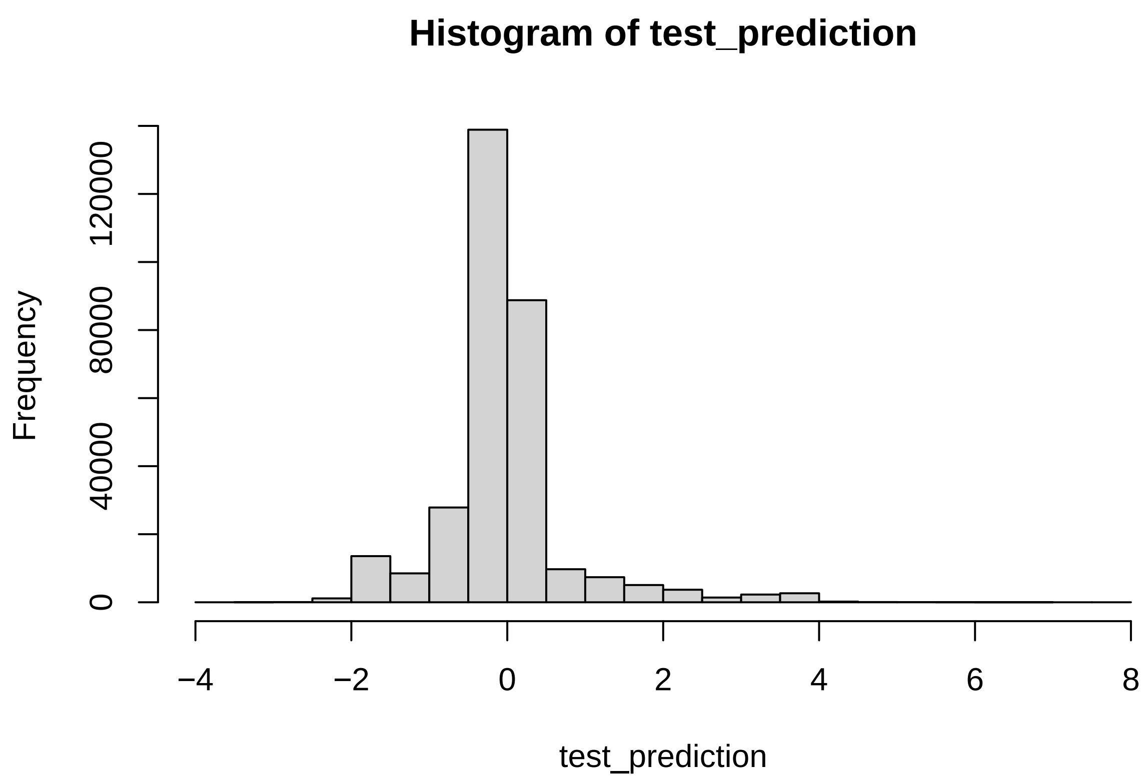
```r
# Odds ratio
(exp(mod1$coefficients[-1])-1)*100
```

```
##    VEH_COUNT  ALCOHOL_NO1  BELTED_UNB1  BIKE_NONBI1   OWNERSHIP2   OWNERSHIP3
##    31.766385    22.341473   354.745545  3306.602880   -27.700516    33.749231
##   OWNERSHIP4   OWNERSHIP5   OWNERSHIP6 COLLISION_10 COLLISION_11 COLLISION_12
##   -31.227413   -22.622571   -26.024859   -71.422518   -49.010902     7.097183
## COLLISION_13 COLLISION_14 COLLISION_15 COLLISION_16  COLLISION_2  COLLISION_3
##    -2.494843   142.860382   -56.226386    30.291549    18.364243   115.924407
##  COLLISION_4  COLLISION_5  COLLISION_6  COLLISION_7  COLLISION_8  COLLISION_9
##   -50.925709    -9.557109   -15.192553    -7.519908    82.794876    30.711387
## WEATHER_CO10 WEATHER_CO11  WEATHER_CO3  WEATHER_CO4  WEATHER_CO5  WEATHER_CO6
##    -5.535889     9.525953     6.744261    10.202624   -11.664453   -40.902296
##  WEATHER_CO7  WEATHER_CO8  WEATHER_CO9  LIGHT_COND2  LIGHT_COND3  LIGHT_COND4
##   -30.210751    19.246452     8.972109    10.158232     8.597945    -2.713015
##  LIGHT_COND5  LIGHT_COND6  LIGHT_COND7  DISTRACTED1      ANIMAL1  DROWSY_NOT1
##     4.911623    -5.580949   -80.216387     2.832019    -3.865224    28.481977
##  DRUG_NODRU1  MOTOR_NONM1  PED_NONPED1  SPEED_NOTS1  SENIOR_NOT1  YOUNG_NOTY1
##    70.054586   891.007008  4304.789288    13.957787    20.657600    -6.980659
##     year2016     year2017     year2018     year2019     year2020     year2021
##     1.976238    -4.749257    -4.817165    -6.893703    -9.212477   -12.603513
##     year2022     year2023
##   -14.638772   -14.383263
```

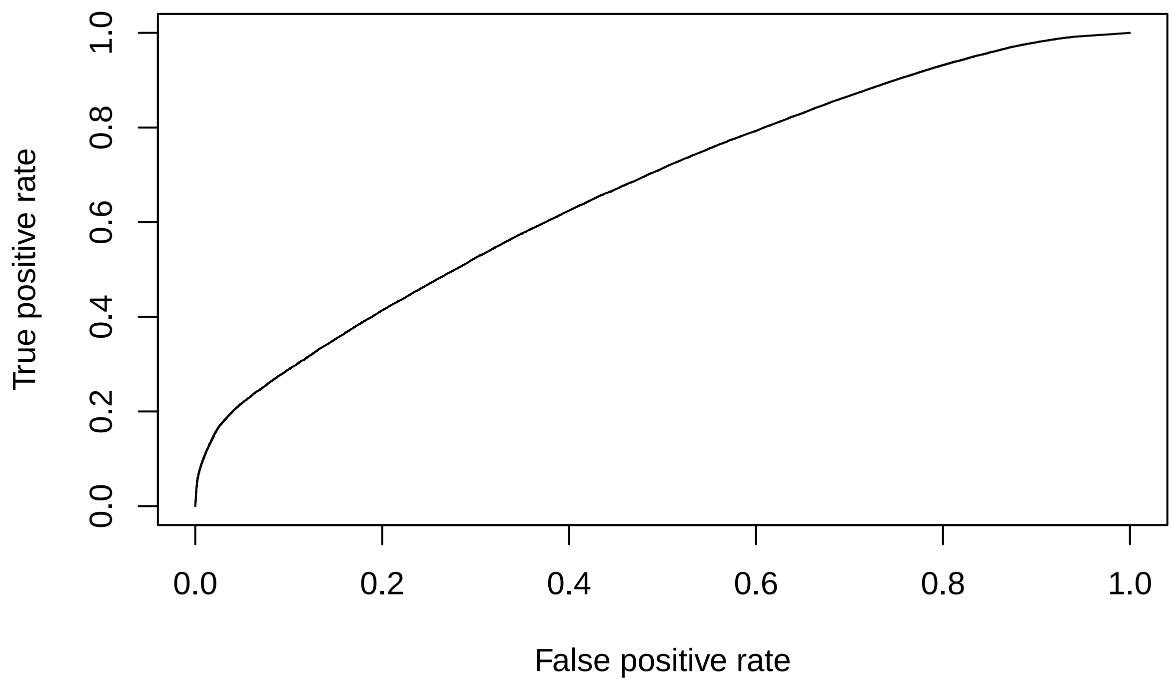
## Evaluate the logistic regression performance on the testing set

Implementing a roc curve

```
test_prediction <- predict(mod1, newdata = test)
hist(test_prediction)
```




```
pred = prediction(test_prediction, test$injury)
perf = performance(pred, "tpr", "fpr")
plot(perf)
```




Implementing a confusion Matrix and accuracy

```
table(test$injury, test_prediction>0.5)
```

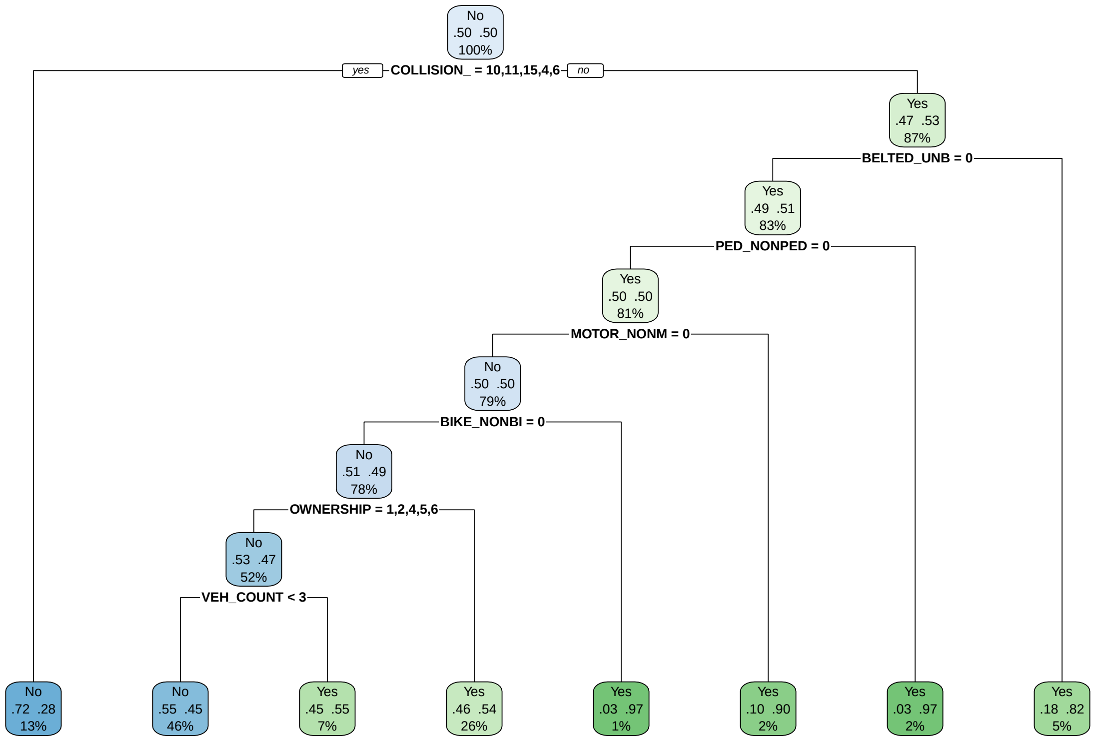
```
##
##         FALSE   TRUE
##   No   198485  10280
##   Yes   80322  22207
```

```
(198485+22207)/(198485+22207+10280+80322)
```

```
## [1] 0.7089504
```

# Decision Tree

```
# Decision tree
set.seed(12)
mod2 <- rpart(injury ~ VEH_COUNT + ALCOHOL_NO + BELTED_UNB + BIKE_NONBI +
            OWNERSHIP + COLLISION_ + WEATHER_CO + LIGHT_COND + DISTRACTED +
            ANIMAL + DROWSY_NOT + DRUG_NODRU + MOTOR_NONM + PED_NONPED +
            SPEED_NOTS + SENIOR_NOT + YOUNG_NOTY + year, data=train, method = "class",
            control = rpart.control(maxdepth = 10, minbucket = 7, minsplit = 10))
rpart.plot(mod2, extra = 104)
```




## Evaluate the decision treeperformance on the testing set

```r
# predict injury or not on train data
train_prediction = predict(mod2, data=train, type = "class")

# Confusion Matrix on train
tab2 = table(Predicted = train_prediction, Actual = train$injury)
confusionMatrix(tab2)
```

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicted      No    Yes
##       No   162870 116033
##       Yes   76363 123200
##
##                Accuracy : 0.5979
##                  95% CI : (0.5965, 0.5993)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1958
##
```

```
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6808
##             Specificity : 0.5150
##          Pos Pred Value : 0.5840
##          Neg Pred Value : 0.6173
##              Prevalence : 0.5000
##          Detection Rate : 0.3404
##    Detection Prevalence : 0.5829
##       Balanced Accuracy : 0.5979
##
##        'Positive' Class : No
##
```

```r
# predict injury or not on test data
test_prediction2 = predict(mod2, newdata=test, type = "class")

# Confusion Matrix on test
tab3 = table(Predicted = test_prediction2, Actual = test$injury)
confusionMatrix(tab3)
```

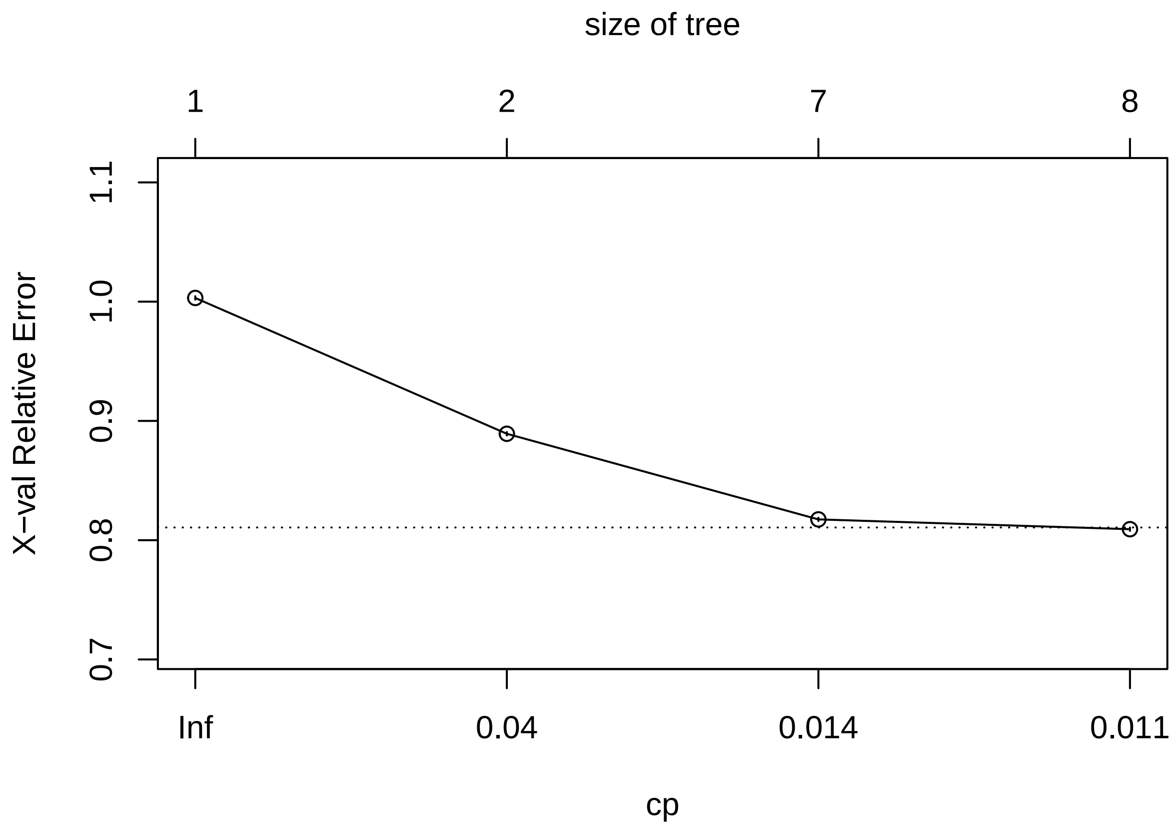
```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted     No    Yes
##       No  142781  49908
##       Yes  65984  52621
##
##                Accuracy : 0.6277
##                  95% CI : (0.626, 0.6294)
##     No Information Rate : 0.6706
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1896
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6839
##             Specificity : 0.5132
##          Pos Pred Value : 0.7410
##          Neg Pred Value : 0.4437
##              Prevalence : 0.6706
##          Detection Rate : 0.4587
##    Detection Prevalence : 0.6190
##       Balanced Accuracy : 0.5986
##
##        'Positive' Class : No
##
```

### Decision Tree pruning

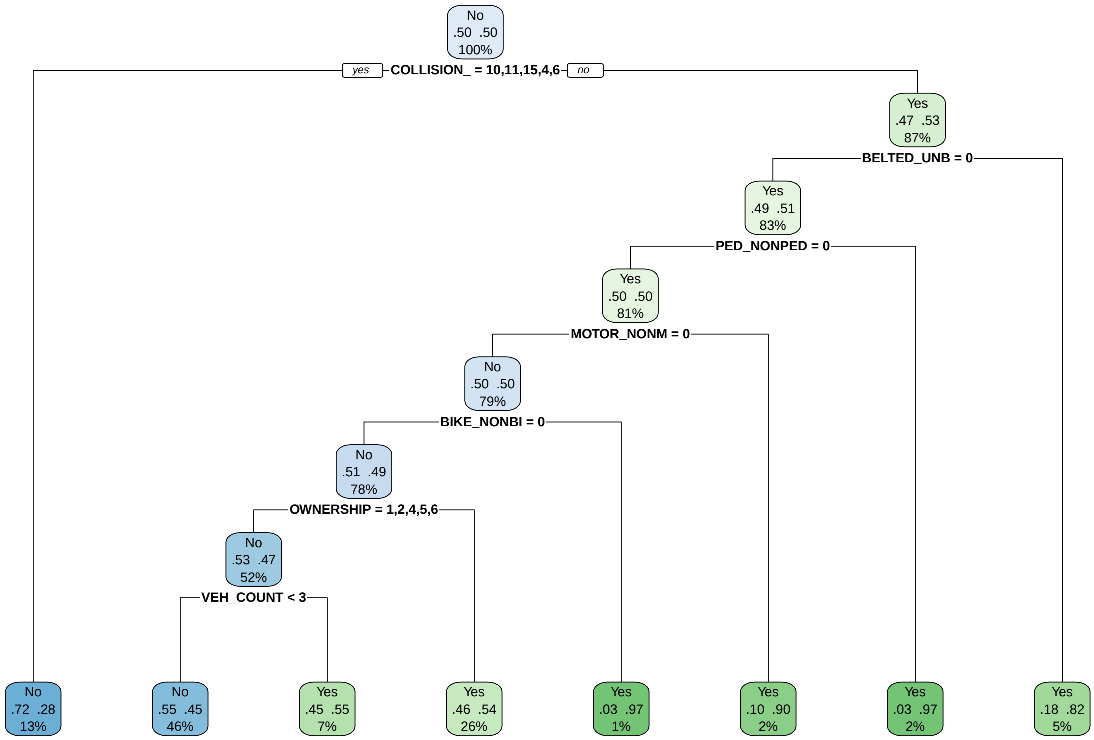
```r
# Complexity plot
printcp(mod2)
```

```
##
## Classification tree:
## rpart(formula = injury ~ VEH_COUNT + ALCOHOL_NO + BELTED_UNB +
##     BIKE_NONBI + OWNERSHIP + COLLISION_ + WEATHER_CO + LIGHT_COND +
##     DISTRACTED + ANIMAL + DROWSY_NOT + DRUG_NODRU + MOTOR_NONM +
##     PED_NONPED + SPEED_NOTS + SENIOR_NOT + YOUNG_NOTY + year,
##     data = train, method = "class", control = rpart.control(maxdepth = 10,
##         minbucket = 7, minsplit = 10))
##
## Variables actually used in tree construction:
## [1] BELTED_UNB BIKE_NONBI COLLISION_ MOTOR_NONM OWNERSHIP  PED_NONPED VEH_COUNT
##
## Root node error: 239233/478466 = 0.5
##
## n= 478466
##
##         CP nsplit rel error  xerror      xstd
## 1 0.112037      0   1.00000 1.00310 0.0014457
## 2 0.014123      1   0.88796 0.88923 0.0014368
## 3 0.013125      6   0.81735 0.81743 0.0014214
## 4 0.010000      7   0.80422 0.80922 0.0014191
```

```r
plotcp(mod2)
```




```r
# pruning
mod3 = prune(mod2, cp=0.011)
rpart.plot(mod3, extra = 104)
```




Evaluate the pruning performance on the testing set

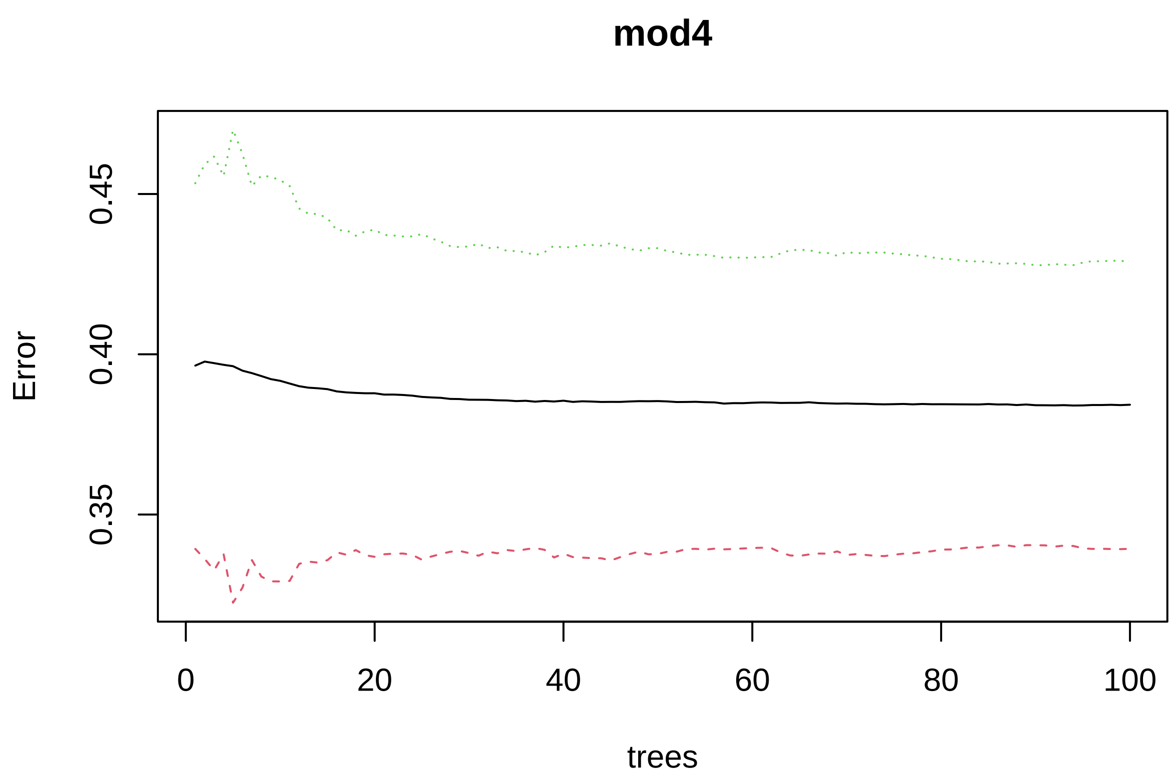
```r
# predict injury or not on test data
test_prediction3 = predict(mod3, newdata=test, type = "class")

# Confusion Matrix on train
tab4 = table(Predicted = test_prediction3, Actual = test$injury)
confusionMatrix(tab4)
```

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicted      No    Yes
##       No   142781  49908
##       Yes   65984  52621
##
##                Accuracy : 0.6277
##                  95% CI : (0.626, 0.6294)
##     No Information Rate : 0.6706
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1896
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6839
```

```
##            Specificity : 0.5132
##         Pos Pred Value : 0.7410
##         Neg Pred Value : 0.4437
##             Prevalence : 0.6706
##         Detection Rate : 0.4587
##   Detection Prevalence : 0.6190
##      Balanced Accuracy : 0.5986
##
##       'Positive' Class : No
##
```

## Random Forest

```
set.seed(431)
mod4=randomForest(injury ~ VEH_COUNT + ALCOHOL_NO + BELTED_UNB + BIKE_NONBI +
            OWNERSHIP + COLLISION_ + WEATHER_CO + LIGHT_COND + DISTRACTED +
            ANIMAL + DROWSY_NOT + DRUG_NODRU + MOTOR_NONM + PED_NONPED +
            SPEED_NOTS + SENIOR_NOT + YOUNG_NOTY + year, data=train,
            mtry=4,importance=TRUE, ntree=100)
plot(mod4)
```




Evaluate the random forest performance on the testing set

```r
# predict injury or not on test data
test_prediction4 = predict(mod4, newdata=test, type = "class")

# Confusion Matrix on train
tab5 = table(Predicted = test_prediction4, Actual = test$injury)
confusionMatrix(tab5)
```

```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted     No    Yes
##       No  137613  43787
##       Yes  71152  58742
##
##                Accuracy : 0.6308
##                  95% CI : (0.6291, 0.6325)
##     No Information Rate : 0.6706
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2173
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6592
##             Specificity : 0.5729
##          Pos Pred Value : 0.7586
##          Neg Pred Value : 0.4522
##              Prevalence : 0.6706
##          Detection Rate : 0.4421
##    Detection Prevalence : 0.5827
##       Balanced Accuracy : 0.6161
##
##        'Positive' Class : No
##
```

Variable importance

```r
importance(mod4)
```

```
##                     No        Yes MeanDecreaseAccuracy MeanDecreaseGini
## VEH_COUNT    17.966281  4.9562254            45.826191        1792.0318
## ALCOHOL_NO   16.314951  3.4465899            29.468676         577.0574
## BELTED_UNB  160.808228 78.3045330           136.397310        4798.0268
## BIKE_NONBI   61.579781 42.1744334            67.992573        1492.5634
## OWNERSHIP    12.456580 46.5600588            88.249637        2114.3650
## COLLISION_   33.719635  1.2367817            80.045257        8829.8832
## WEATHER_CO   15.368890  3.0740577            23.929980        1405.7668
## LIGHT_COND    7.989843  6.8750529            33.721803        1913.1855
## DISTRACTED   10.161461 -2.0080353            13.698238         588.9727
## ANIMAL        8.304783  0.1736327             9.331178        1656.6972
## DROWSY_NOT    2.797025  8.7454134            18.579646         348.4508
## DRUG_NODRU   28.732712  5.4242269            31.917133         341.6790
```

```
## MOTOR_NONM 148.167224 63.0216973           102.648888        3492.4116
## PED_NONPED  57.244969  1.0400208            23.046533        2790.4319
## SPEED_NOTS  12.971106  3.4018552            27.759250         694.4888
## SENIOR_NOT  20.481074 11.1564062            46.003629         656.4176
## YOUNG_NOTY   9.291894  4.6300420            21.469476         556.6132
## year         9.664153 10.9111582            14.355525        2881.7686
```

```r
varImpPlot(mod4)
```