Risk Factors for Injury Severity in Road Accidents

2025-06-27

This kaggle dataset provides the research collected by the Addis Adaba Sub-city police department of road traffic accidents from 2017 through 2020. https://www.kaggle.com/datasets/kanuriviveknag/road-accidents-severity-dataset

My goal is to identify significant predictors of injury occurrence in road traffic accidents using logistic regression.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(dplyr)
library(ggplot2)
Data = read_csv("RTA Dataset 2.csv"); count(Data)
```

```
## Rows: 12316 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (29): Day_of_week, Age_band_of_driver, Sex_of_driver, Educational_level...
## dbl   (2): Number_of_vehicles_involved, Number_of_casualties
## time  (1): Time
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 12316
```

Given over 12,300 entries, I want to first organize my data set and select only the subset of variables that will be relevant for my task.

```
Data = read_csv("RTA Dataset 2.csv") %>%
  select(Accident_severity, Sex_of_driver, Area_accident_occured, Light_conditions,
         Educational_level, Road_surface_conditions) %>%
  rename(Severity = Accident_severity, Sex = Sex_of_driver, Area = Area_accident_occured,
         Light = Light_conditions, Education = Educational_level, Road = Road_surface_conditions) %>%
  # Renamed for simplicity
  drop_na() %>%
  filter(!Sex %in% c("Unknown"), !Education %in% c("Unknown"), !Area %in% c("Unknown"),
         !Light %in% c("Unknown"), !Road %in% c("Unknown"))
```

```
## Rows: 12316 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (29): Day_of_week, Age_band_of_driver, Sex_of_driver, Educational_level...
## dbl   (2): Number_of_vehicles_involved, Number_of_casualties
## time  (1): Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# To sparse the data I dropped all unknowns within the variables.

Data = Data %>%
  mutate(Area = fct_collapse(as.factor(Area),"Urban_institutions" = c("Hospital areas",
"Office areas", "School areas", "Industrial areas"), "Residential/Market" = c("Residential areas",
"Market areas", "Church areas", "Other"), "Rural" = c("Rural village areas", "Outside rural areas",
"Rural village areasOffice areas"),"Recreational" = "Recreational areas"),
Light = ifelse(Light == "Daylight", "Daylight", "Low Visibility"), Light = as.factor(Light),
    Road = fct_collapse(as.factor(Road), "Severe" = c("Flood over 3cm. deep", "Snow"),
"Wet" = "Wet or damp", "Dry" = "Dry"),
    Sex = as.factor(Sex),
    Injury = ifelse(Severity %in% c("Slight Injury", "Serious Injury", "Fatal Injury"), 1, 0))
```

Above, I continued to clean the data by combining similar entries into broader categories. I then factored predictor variables into categorical variables and created a binary variable for injuries sustained.

```
logit_model = glm(Injury ~ Sex + Area + Light + Road, family = binomial, data = Data)
summary(logit_model)
```

```
##
## Call:
## glm(formula = Injury ~ Sex + Area + Light + Road, family = binomial,
##     data = Data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)              5.0988     0.4656  10.951  < 2e-16 ***
## SexMale                 -0.5987     0.4585  -1.306   0.1916
## AreaUrban_institutions   0.2660     0.1889   1.409   0.1590
## AreaRural                0.5371     0.7185   0.748   0.4548
## AreaRecreational        -0.6990     0.3751  -1.863   0.0624 .
## LightLow Visibility     -0.8445     0.1747  -4.835 1.33e-06 ***
## RoadSevere              13.3165   477.9144   0.028   0.9778
## RoadWet                  0.2995     0.2094   1.430   0.1527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1502.4  on 11086  degrees of freedom
## Residual deviance: 1470.4  on 11079  degrees of freedom
## AIC: 1486.4
##
## Number of Fisher Scoring iterations: 16
```

A large standard error of 477.9144 for Road Severity indicates great uncertainty in the estimation of the tested parameters. This implies that there is potentially multicollinearity within this section, and that the predictor variables are possibly correlated to each other. To counteract, I merged rarer entries into a broader "Other" group.

```
Data = Data %>%
  mutate(Area = fct_lump(Area, n = 3, other_level = "Other"), Road = fct_lump(Road, n = 2,
other_level = "Other")) %>%
  filter(Road != "Other")
# Remove overly sparse Road category

logit_model = glm(Injury ~ Sex + Area + Light + Road, family = binomial, data = Data)
summary(logit_model)
```

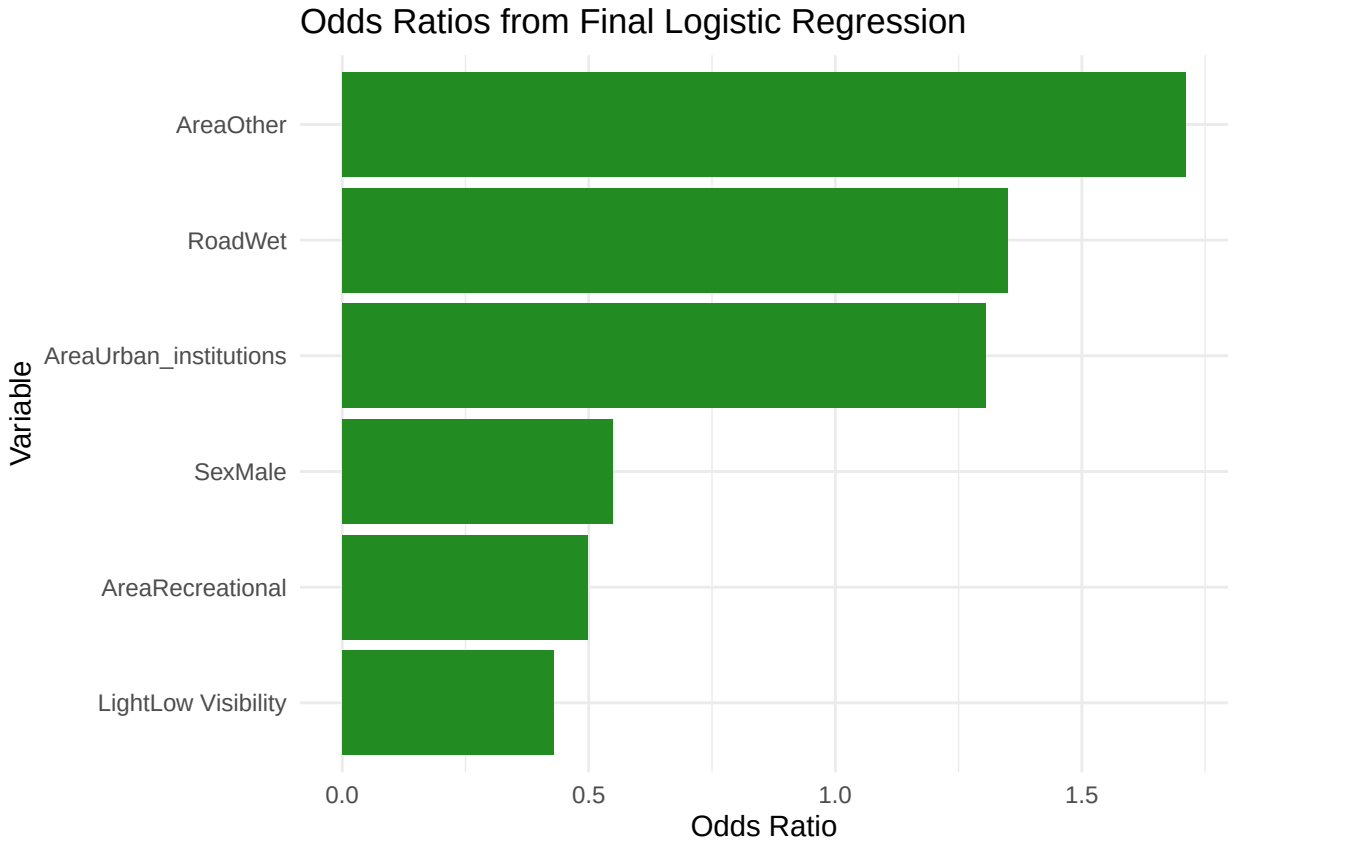
```
##
## Call:
## glm(formula = Injury ~ Sex + Area + Light + Road, family = binomial,
##     data = Data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              5.0988     0.4656  10.951  < 2e-16 ***
## SexMale                 -0.5987     0.4584  -1.306   0.1916
## AreaUrban_institutions   0.2660     0.1888   1.409   0.1590
## AreaRecreational        -0.6990     0.3751  -1.863   0.0624 .
## AreaOther                0.5371     0.7184   0.748   0.4547
## LightLow Visibility     -0.8445     0.1747  -4.835 1.33e-06 ***
## RoadWet                  0.2995     0.2094   1.430   0.1527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1500.7  on 11020  degrees of freedom
## Residual deviance: 1470.4  on 11014  degrees of freedom
## AIC: 1484.4
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
# Calculate and visualize Odds Ratios
Odds = data.frame(Variable = names(coef(logit_model)), OR = exp(coef(logit_model)))
Odds
```

```
##                                     Variable          OR
## (Intercept)                       (Intercept) 163.8239317
## SexMale                             SexMale   0.5495322
## AreaUrban_institutions AreaUrban_institutions   1.3047416
## AreaRecreational             AreaRecreational   0.4971033
## AreaOther                           AreaOther   1.7109973
## LightLow Visibility       LightLow Visibility   0.4297621
## RoadWet                             RoadWet   1.3491626
```

```
ggplot(Odds[-1,], aes(x = reorder(Variable, OR), y = OR)) +
  geom_bar(stat = "identity", fill = "forestgreen") +
  coord_flip() +
  labs(title = "Odds Ratios from Final Logistic Regression", x = "Variable", y = "Odds Ratio") +
  theme_minimal()
```



From the above model, we see that male drivers had a 45.05% lower odds of injury, but this result is not statistically significant (p=0.1916). Urban areas had a 30.47% higher chance of injury compared to residential areas, but again, this is not significant (p=.0159). Recreational areas had marginally lower odds of injury (50.29%), which may reflect that individuals drive slower speeds through these areas, but these odds are not statistically significant (p=.0624). Other yielded a p=.4547 implying great uncertainty, but this was to be expected with this category representing a grouping of underrepresented occurrences. For low visibility, there are statistically significantly lower odds of injury (57.02%; p<.05). This may reflect more cautious driving at

night. Finally, there is a 34.92% higher chance of injury on wet roads, but this is not significant (p=.1527).

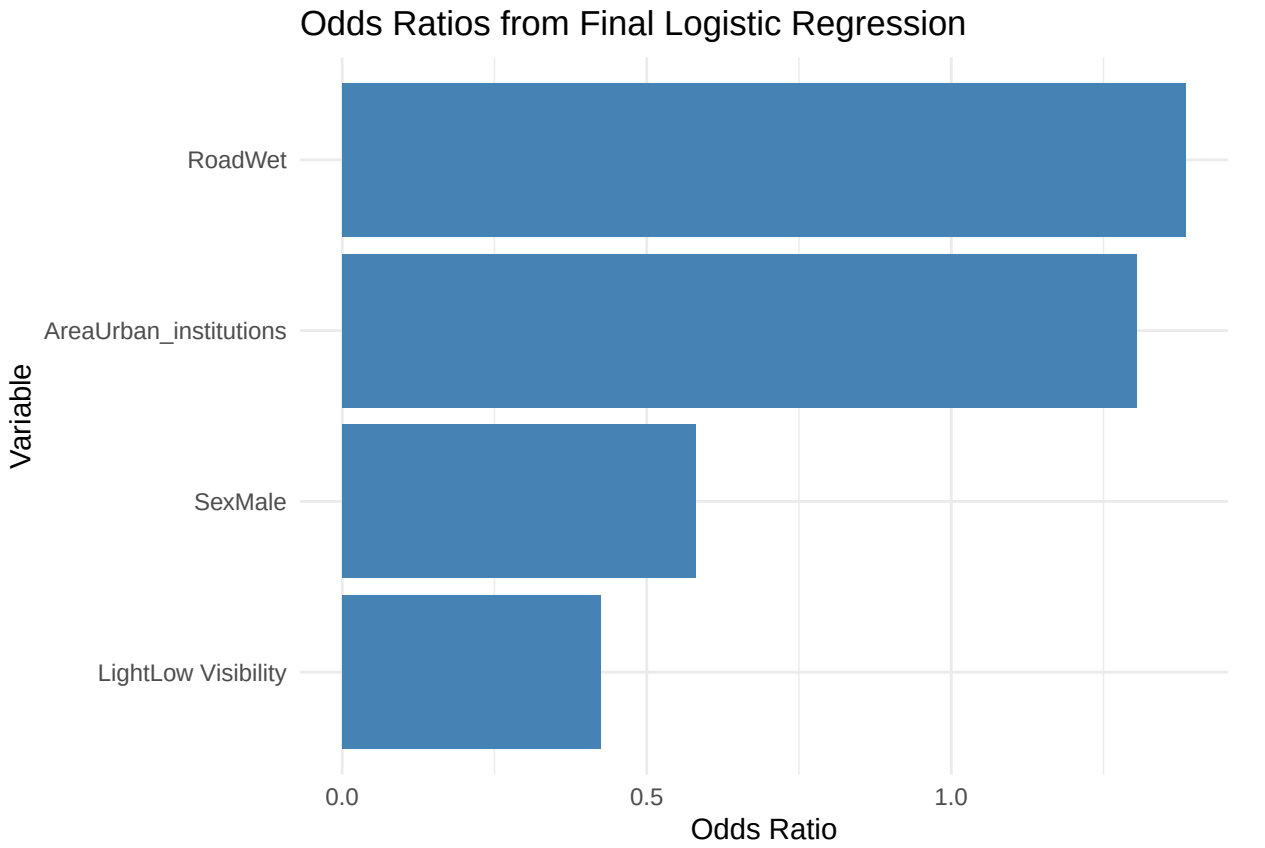
```
Data = Data %>%
  filter(Area %in% c("Residential/Market", "Urban_institutions")) %>%
  droplevels()

logit_model1 = glm(Injury ~ Sex + Area + Light + Road, family = binomial, data = Data)
summary(logit_model1)
```

```
##
## Call:
## glm(formula = Injury ~ Sex + Area + Light + Road, family = binomial,
##     data = Data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.0476     0.4666  10.819  < 2e-16 ***
## SexMale                  -0.5453     0.4592  -1.188    0.235
## AreaUrban_institutions    0.2660     0.1889   1.409    0.159
## LightLow Visibility      -0.8574     0.1810  -4.737 2.17e-06 ***
## RoadWet                   0.3259     0.2193   1.486    0.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1399.1  on 10455  degrees of freedom
## Residual deviance: 1374.0  on 10451  degrees of freedom
## AIC: 1384
##
## Number of Fisher Scoring iterations: 7
```

```
Odds = data.frame(Variable = names(coef(logit_model1)), OR = exp(coef(logit_model1)))

ggplot(Odds[-1,], aes(x = reorder(Variable, OR), y = OR)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Odds Ratios from Final Logistic Regression", x = "Variable", y = "Odds Ratio") +
  theme_minimal()
```




```
vif_values = vif(logit_model)
print(vif_values)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## Sex   1.003437  1        1.001717
## Area  1.001359  3        1.000226
## Light 1.044547  1        1.022031
## Road  1.042553  1        1.021055
```

Variance Inflation Factors were all near 1, indicating no evidence of multicollinearity among predictors.

Overall Findings

In this project, I analyzed over 12,000 road traffic incidents to identify key risk factors for injury using logistic regression. I cleaned and recoded categorical data, merged sparse categories, and ensured model stability through exploratory data checks. My analysis revealed that injury risk was significantly reduced in low-visibility conditions, highlighting the importance of driver behavior over raw environmental danger. The results were visualized using odds ratio plots and interpreted for both practical meaning and model reliability. Its also important to note that recreational and "Other" areas were removed entirely. While this improves model stability, this makes Area strictly binary, not often a common occurrence in real life. Similarly, the original severity variable contained ordinal levels, but I collapsed them into a binary outcome (method I am educated in). This discards potentially valuable distinctions between minor and fatal injuries, which might affect conclusions. Recall that some sparse levels had inflated standard errors that were removed and/or merged, which could obscure true effects. While there are currently limitations, I look forward to continuing my journey of learning R and practicing my data wrangling skills.