NYPD Shooting Incident Data Project : Child Victims

2022-07-03

Overview

The focus of this project will be to understand how attributes of the known perpetrator of a shooting impacts the likelihood of the victim being a child (under the age of 18) via logistic regression. Historical data of shootings from the the New York Police Department (NYPD) is utilized to address this question. The data is located here https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8.

Note code snippets have not been suppressed throughout to showcase reproducibility.

Tidyverse is needed for the `read_csv()` function. Ensure the package is installed via `install.packages("tidyverse")`. Lubridate is needed for converting date data to the appropriate form. ROCR is used for calculating the Receiving Operator Characteristic (ROC) curve and the corresponding area under the curve (AUC) in logistic regression and must also be installed via `install.packages("ROCR")`.

```
library(tidyverse)
library(lubridate)
library(ROCR)
```

The data is retrieved and converted into a tidyverse tibble.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd <- read_csv(url)
```

The data is inspected for a high-level overview.

```
dim(nypd)
```

```
## [1] 25596    19
```

```
summary(nypd)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
```

```
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

In total the data contains records of 25,596 shootings, between 2006 and 2021 inclusive. Up to 19 recorded variables are associated with each shooting. These inform when and where the shooting took place as well as demographics of both the perpetrator and victim. The demographic fields involving the perpetrator and the "VIC_AGE_GROUP" will be the ones of interest for this project.

Data Preprocessing

Some tidying of the data is required. Not all cleaned fields will ultimately be used in the subsequent analysis but are included to highlight best practices. Fields of character type need to be converted to factors. Additionally, some fields that are perceived to be numeric also need to be converted to factors. These include: "INCIDENT_KEY", "JURISDICTION CODE" and "PRECINCT". Finally, the "OCCUR_DATE" field should be converted to a date type field. The following subset of columns are selected as they are the ones relevant for the current analysis (the remaining columns are dropped) : "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE", "VIC_AGE_GROUP". The below code chunk performs what has been described above.

```r
nypd <- nypd %>%
  mutate_if(is.character, as.factor) %>%
  mutate(across(c(INCIDENT_KEY, JURISDICTION_CODE, PRECINCT), as.factor)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(PERP_AGE_GROUP:VIC_AGE_GROUP)
```

Now we take a second look at the summary information.

```
summary(nypd)
```

```
##  PERP_AGE_GROUP PERP_SEX              PERP_RACE     VIC_AGE_GROUP
##  18-24   :5844   F   :  371   BLACK          :10668   <18    : 2681
##  25-44   :5202   M   :14416   WHITE HISPANIC: 2164   18-24  : 9604
##  UNKNOWN:3148   U   : 1499   UNKNOWN        : 1836   25-44  :11386
##  <18     :1463   NA's: 9310   BLACK HISPANIC: 1203   45-64  : 1698
##  45-64   : 535                WHITE          :  272   65+    :  167
##  (Other):  60                (Other)        :  143   UNKNOWN:   60
##  NA's    :9344                NA's           : 9310
```

There are missing values in our fields of interest ("NA's"). The similar number of missing values across fields suggests the missing values may be concentrated among the same row records.

```
missing <- rowSums(is.na(nypd))
paste('Rows missing three values: ',
       sum(missing == 3))
```

```
## [1] "Rows missing three values:  9310"
```

This confirms that the missing values primarily come from shooting records with missing values in all three perpetrator demographic fields. Perhaps the perpetrator is unknown in these instances. Note this is different from a demographic attribute being characterized as "UNKNOWN", where there are mismatches in field values. A reduced data set, with rows containing missing values removed, is constructed below and then summarized. The data set should still be sufficiently large but the systematic removal of rows with potentially unknown perpetrators may bias the analysis in some way. Sources of bias will be discussed at a later stage when interpreting results.

```
nypd_red <- nypd %>%
  na.omit

summary(nypd_red)
```

```
##  PERP_AGE_GROUP PERP_SEX                              PERP_RACE     VIC_AGE_GROUP
##  18-24   :5844   F:  371   AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    :1869
##  25-44   :5202   M:14416   ASIAN / PACIFIC ISLANDER      :  141   18-24  :6036
##  UNKNOWN:3148   U: 1465   BLACK                         :10668   25-44  :7044
##  <18     :1463             BLACK HISPANIC                : 1203   45-64  :1125
##  45-64   : 535             UNKNOWN                       : 1802   65+    : 123
##  65+     :  57             WHITE                         :  272   UNKNOWN:  55
##  (Other):   3             WHITE HISPANIC                : 2164
```

The "PERP_AGE_GROUP" field needs further investigation for the "(Other)" category.

```
summary(nypd_red$PERP_AGE_GROUP)
```

```
##     <18    1020   18-24     224   25-44   45-64     65+     940 UNKNOWN
##    1463       1    5844       1    5202     535      57       1    3148
```

There appear to be some typos for the age. As there are only a few, they can be removed as well. The typo levels are also removed from the "PERP_AGE_GROUP" factor.

```
nypd_red <- nypd_red %>%
  subset(PERP_AGE_GROUP != 1020 &
         PERP_AGE_GROUP != 224 &
         PERP_AGE_GROUP != 940)

nypd_red$PERP_AGE_GROUP <- nypd_red$PERP_AGE_GROUP %>%
  droplevels()
```

A binary variable "child is constructed and appended based on the"VIC_AGE_GROUP" category being "<18".

```
nypd_red["child"] = (nypd_red["VIC_AGE_GROUP"] == '<18')
```

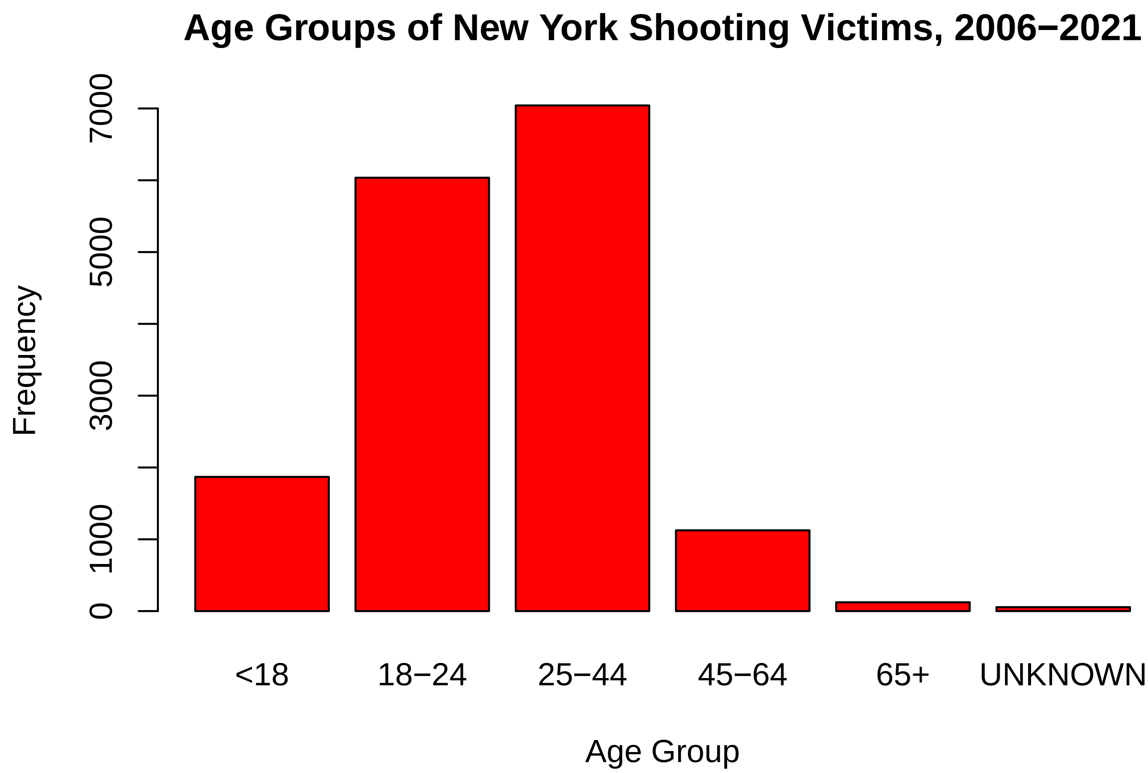
Finally to better interpret coefficients outputted from logistic regression, the predictor factor variables concerning the perpetrator are all re-leveled with baselines corresponding to their respective most common factor.

```
nypd_red <-  within(nypd_red,
  PERP_AGE_GROUP <- relevel(PERP_AGE_GROUP, ref = '18-24'))

nypd_red <- within(nypd_red,
  PERP_SEX  <- relevel(PERP_SEX, ref = 'M'))

nypd_red <- within(nypd_red,
  PERP_RACE <- relevel(PERP_RACE, ref = 'BLACK'))
```

Visualization

When plotting the distribution of "VIC_AGE_GROUP" it is easy to see the size of the target class (child victims) is unbalanced,

```
plot(nypd_red$VIC_AGE_GROUP,
  main = "Age Groups of New York Shooting Victims, 2006-2021",
  xlab = "Age Group",
  ylab = "Frequency",
  col  = "Red")
```



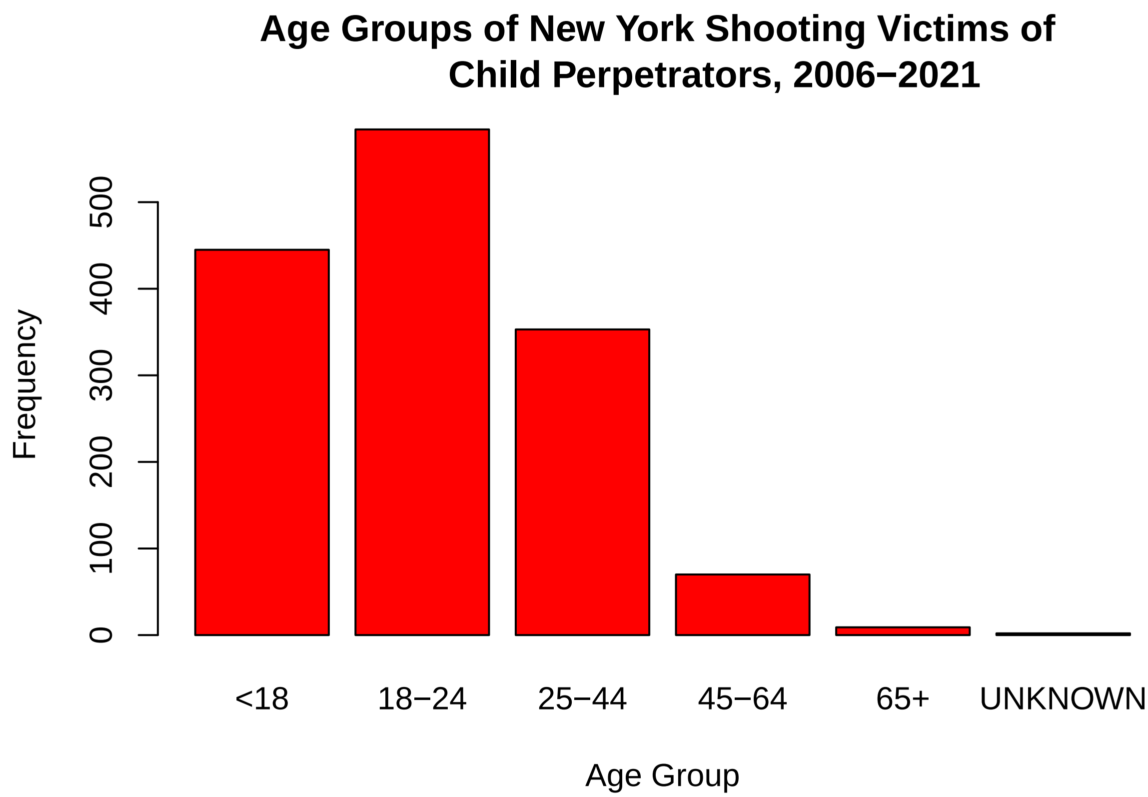
The proportion of the target class is determined below. This is important to keep in mind in subsequent analysis.

```
paste('Proportion of child victims: ',
       round(sum(nypd_red$child)/nrow(nypd_red), 2))
```

```
## [1] "Proportion of child victims:  0.12"
```

Some insights may be gained by looking only at instances where the perpetrator is known to be a child (under the age of 18).

```
plot(nypd_red[nypd_red$PERP_AGE_GROUP == '<18', ]$VIC_AGE_GROUP,
  main = "Age Groups of New York Shooting Victims of
          Child Perpetrators, 2006-2021",
  xlab = "Age Group",
  ylab = "Frequency",
  col  = "Red")
```



Quite clearly the proportion of child victims has increased substantially. The visualizations suggests that perhaps a child perpetrator of a shooting may be a strong predictor of a child victim. Subsequent modelling can provide additional evidence for this.

Modelling

To avoid overfitting, the data will be split 80% for training the model and 20% for testing model performance. The `set.seed(0)` is used for reproducibility due to randomness in sampling.

```
sample_size <- floor(0.8 * nrow(nypd_red))

set.seed(0)
train_ind <- sample(1:nrow(nypd_red), size = sample_size)

train <- nypd_red[train_ind, ]
test <- nypd_red[-train_ind, ]
```

The model is fit on the training data with the target variable and perpetrator demographic predictor variables. "Family" is set to binomial as the type of generalized liner model (glm) is logistic regression for binary classification. The factor variables are converted to indicator variables at each level with the exception of the respective baselines.

```
logr_fit <- glm(child ~ PERP_AGE_GROUP + PERP_SEX + PERP_RACE,
  data = train,  family = binomial)
```

```r
summary(logr_fit)
```

```
##
## Call:
## glm(formula = child ~ PERP_AGE_GROUP + PERP_SEX + PERP_RACE,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9162  -0.5234  -0.5026  -0.3101   2.6992
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.91853    0.04799 -39.978  < 2e-16
## PERP_AGE_GROUP<18                        1.04169    0.07802  13.351  < 2e-16
## PERP_AGE_GROUP25-44                     -1.09194    0.08555 -12.764  < 2e-16
## PERP_AGE_GROUP45-64                     -1.34657    0.26771  -5.030 4.91e-07
## PERP_AGE_GROUP65+                      -12.53943  128.42399  -0.098   0.9222
## PERP_AGE_GROUPUNKNOWN                   -0.08658    0.09790  -0.884   0.3765
## PERP_SEXF                               -0.60579    0.25024  -2.421   0.0155
## PERP_SEXU                                0.28793    0.19999   1.440   0.1499
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -11.55560  882.74338  -0.013   0.9896
## PERP_RACEASIAN / PACIFIC ISLANDER       -0.33013    0.37652  -0.877   0.3806
## PERP_RACEBLACK HISPANIC                  0.15884    0.10436   1.522   0.1280
## PERP_RACEUNKNOWN                        -0.06211    0.19270  -0.322   0.7472
## PERP_RACEWHITE                          -0.36931    0.35077  -1.053   0.2924
## PERP_RACEWHITE HISPANIC                  0.01298    0.08624   0.151   0.8804
##
## (Intercept)                             ***
## PERP_AGE_GROUP<18                       ***
## PERP_AGE_GROUP25-44                     ***
## PERP_AGE_GROUP45-64                     ***
## PERP_AGE_GROUP65+
## PERP_AGE_GROUPUNKNOWN
## PERP_SEXF                               *
## PERP_SEXU
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER
## PERP_RACEBLACK HISPANIC
## PERP_RACEUNKNOWN
## PERP_RACEWHITE
## PERP_RACEWHITE HISPANIC
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9195.7  on 12998  degrees of freedom
## Residual deviance: 8607.9  on 12985  degrees of freedom
## AIC: 8635.9
##
## Number of Fisher Scoring iterations: 13
```

Many coefficients are not statistically significant with the exception of a few levels of perpetrator age and also when the perpetrator is female. For the latter of these the likelihood of a child victim decreases compared to the baseline of a male perpetrator. With respect to the perpetrator's age, if the perpetrator is also under 18, the likelihood of a child victim increase compared to the 18-24 baseline. This coincides with what was highlighted by the earlier visualization. Also the likelihood decreases when the perpetrator falls into the 25-44 or 45-64 age groups.

Predictions are generated by the model on the test data. First the probabilities of each instance in the test data is determined. As a first pass, the standard 0.5 threshold is used to generate predictions.

```
logr_probs <- predict(logr_fit, newdata = test, type = "response")

logr_pred <- rep(FALSE, nrow(test))
logr_pred[logr_probs > .5] = TRUE

paste('Number of child victims predicted: ',
       sum(logr_pred))
```

```
## [1] "Number of child victims predicted:  0"
```

The model has predicted every shooting on the test set as a non-child victim using a 0.5 threshold. This appears problematic in comparison to the actual number of child victims in the test set, as shown below. However the expected number of child victims predicted by the model (sum of the probabilities) appears to reasonable.

```
paste('Actual number of child victims in test set: ',
       sum(test$child))
```

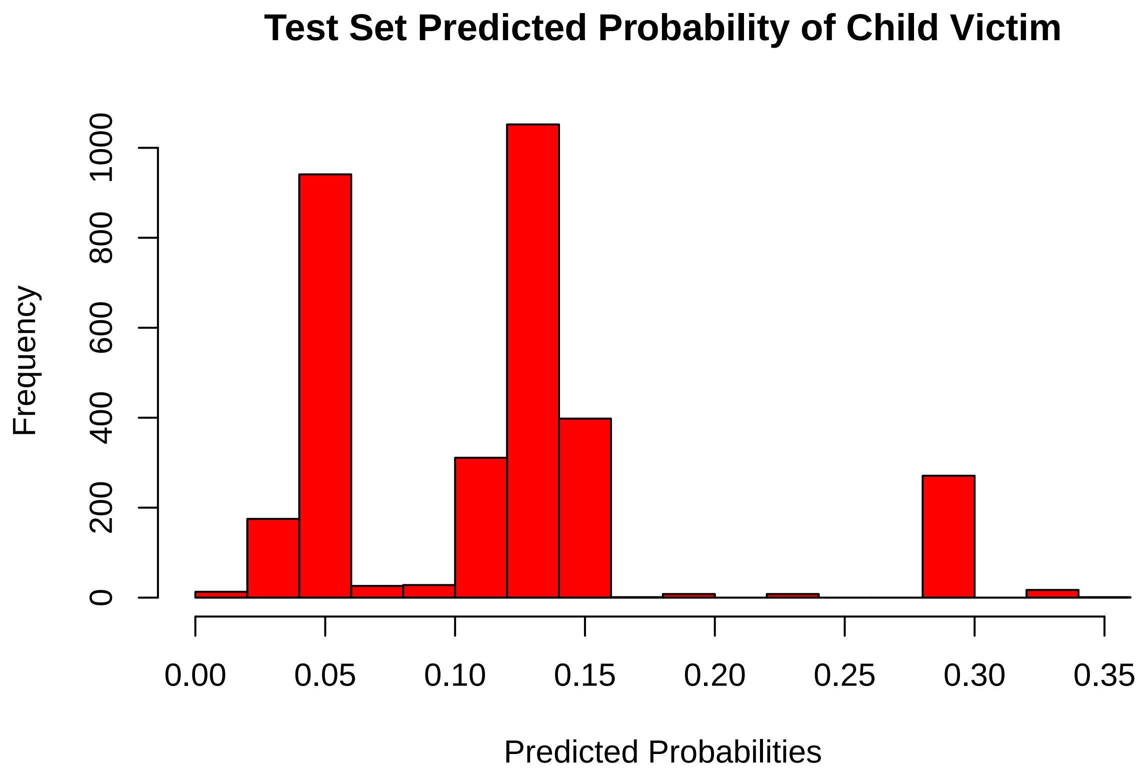
```
## [1] "Actual number of child victims in test set:  394"
```

```
paste('Predicted expected number of child victims in test set: ',
       round(sum(logr_probs), 2))
```

```
## [1] "Predicted expected number of child victims in test set:  373.5"
```

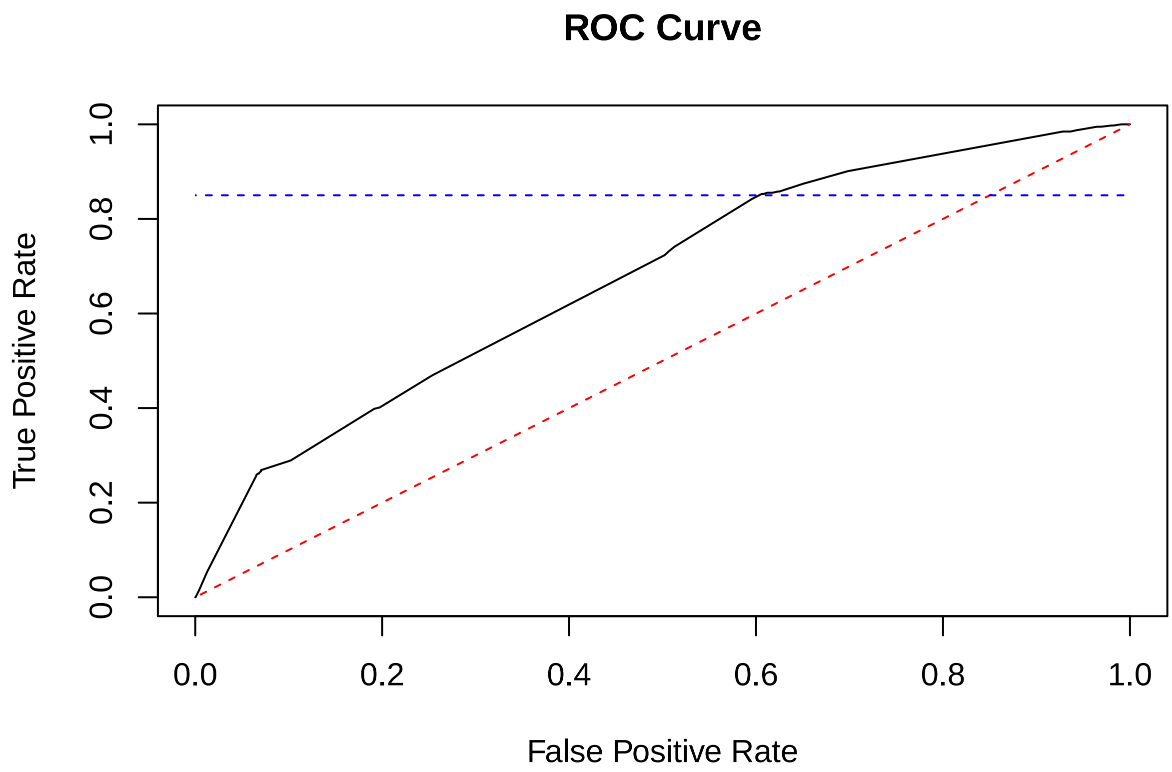
Observe the plot of the distribution of model predicted probabilities on the test set.

```
hist(logr_probs,
  main = "Test Set Predicted Probability of Child Victim",
  xlab = "Predicted Probabilities",
  ylab = "Frequency",
  col  =  "Red")
```



No shooting in the test set has the model predicting a probability above 0.35 of the victim being a child. Between the unbalanced data, a limited number of predictors (all categorical) and levels within each predictor, higher predicted probabilities would be difficult to obtain. This is not to say the model we have constructed is without use. Thresholds other than 0.5 may be reasonable depending on the application. The model could be utilized as a screen for child victims. For this purpose, accuracy on the test as a performance metric is less important than the true positive rate (also called sensitivity or recall) via minimizing false negatives (actual child victims that the model predicts as adult victims). The receiving operator characteristic (ROC) curve for the model can be constructed to highlight model performance with respect to true positive rate versus false positive rate, at various thresholds.

```
pred <- prediction(logr_probs, test$child)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")

plot(perf,
  main = "ROC Curve",
  xlab = "False Positive Rate",
  ylab = "True Positive Rate")

clip(0, 1, 0, 1)
abline(0, 1, col = "red", lty = 2)
abline(0.85, 0, col = "blue", lty = 2)
```



The ROC curve shows that slightly above a true positive rate of 0.8, further changes in the threshold do not lead to as significant an increase. The area under the ROC curve (AUC) is calculated to show how the model performs against a baseline AUC of 0.5 (dotted red line).

```
auc = performance(pred, "auc")
paste('Area under the ROC curve: ',
       round(auc@y.values[[1]],2))
```

```
## [1] "Area under the ROC curve:  0.68"
```

An AUC of 0.68 (2 s.f.) for a model this simple is reasonable.

The threshold for which the true positive rate does not meaningfully improve is next determined.

```
thresh <- data.frame(threshold = perf@alpha.values[[1]],
                     tpr = perf@y.values[[1]])
thresh[c(15:25),]
```

```
##     threshold       tpr
## 15 0.14439078 0.3984772
## 16 0.13631210 0.4010152
## 17 0.13564526 0.4010152
## 18 0.12948206 0.4695431
## 19 0.12802597 0.7233503
## 20 0.12125089 0.7309645
```

```
## 21 0.12003229 0.7411168
## 22 0.11866798 0.8426396
## 23 0.11232432 0.8527919
## 24 0.09546583 0.8527919
## 25 0.09213556 0.8553299
```

The table shows that the optimal threshold is around 0.12 (2 s.f.) as there are less significant changes to the true positive rate (tpr) when the threshold is further decreased . This coincides to two significant figures with the proportion of child victims in the data set. At this threshold, our model performs with a true positive rate of 0.84 (2 s.f.).

Conlcusion

A logistic regression model was constructed to predict child victims of shootings based on demographics (age group, sex and race) of the perpetrator. Younger age groups and in particular child perpetrators were found to be good predictors of a shooting involving a child victim. Furthermore female perpetrators decreased the likelihood of a child victim in comparison to the male baseline.

While the model could not predict any instances out of sample having an especially high probability of a child victim (due to unbalanced data and the simplicity of the model, it had reasonable efficacy as a screen for shootings with child victims. The AUC of the ROC curve was 0.68 (2.s.f.) and with a threshold of 0.12 (2 s.f.) the model was able to achieve a true positive rate of 0.84 (2 s.f.). Further reductions of the threshold did not lead to meaningful improvements in the true positive rate.

Bias

Conclusions regarding perpetrator demographics in shootings can reinforce one's preconceived notions of whom is perceived to be predisposed to violent crime. Young males are identified here (with no meaningful distinction across racial groups) as those most likely to be involved in shootings where the victim is a child. From a personal perspective, the conclusion does appear to coincide with what is generally presented in media. However, there may systematic bias in which shootings are investigated, and subsequently in both which shootings contain missing values and which shooting would have perpetrator attributes that are unknown. From this, the process by which data was removed for missing values, potentially for unknown perpetrators, would pass the bias onto the interpretation of the model results. For example it may be more likely that for child victims, child perpetrators are more likely to be known in comparison to adult victims. Note also that the model is only predictive, not causal, and inference cannot be extended to the general population. Those who perpetrated shootings of a child may not be representative of the general population and nothing can be said from the results about the demographics of an individual that would lead them to more likely be involved in such a shooting. Having awareness of these sources of bias in data gathering and handling, as well as understanding the limited scope of the ensuing results can mitigate the effects of the various biases.

Appendix

These are the packages and respective versions utilized.

```r
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
```

```
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Singapore.utf8  LC_CTYPE=English_Singapore.utf8
## [3] LC_MONETARY=English_Singapore.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_Singapore.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ROCR_1.0-11     lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0
##  [5] dplyr_1.0.9     purrr_0.3.4     readr_2.1.2     tidyr_1.2.0
##  [9] tibble_3.1.7    ggplot2_3.3.6   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2 xfun_0.31        haven_2.5.0      colorspace_2.0-3
##  [5] vctrs_0.4.1      generics_0.1.2   htmltools_0.5.2  yaml_2.3.5
##  [9] utf8_1.2.2       rlang_1.0.2      pillar_1.7.0     glue_1.6.2
## [13] withr_2.5.0      DBI_1.1.3        bit64_4.0.5      dbplyr_2.2.1
## [17] modelr_0.1.8     readxl_1.4.0     lifecycle_1.0.1  munsell_0.5.0
## [21] gtable_0.3.0     cellranger_1.1.0 rvest_1.0.2      evaluate_0.15
## [25] knitr_1.39       tzdb_0.3.0       fastmap_1.1.0    curl_4.3.2
## [29] parallel_4.2.1   fansi_1.0.3      highr_0.9        broom_0.8.0
## [33] backports_1.4.1  scales_1.2.0     vroom_1.5.7      jsonlite_1.8.0
## [37] bit_4.0.4        fs_1.5.2         hms_1.1.1        digest_0.6.29
## [41] stringi_1.7.6    grid_4.2.1       cli_3.3.0        tools_4.2.1
## [45] magrittr_2.0.3   crayon_1.5.1     pkgconfig_2.0.3  ellipsis_0.3.2
## [49] xml2_1.3.3       reprex_2.0.1     assertthat_0.2.1 rmarkdown_2.14
## [53] httr_1.4.3       rstudioapi_0.13  R6_2.5.1         compiler_4.2.1
```