Chicago Crime

Xin Guan, Vera Hudak, Yuqi Zhang

2023-11-24

```r
knitr::opts_chunk$set(cache = T)
# Make the whole document reproducible
# Packages required
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```r
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
```

```r
library(magrittr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.0
## v ggplot2 3.4.3     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.0
## v readr   2.1.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
##
## The following objects are masked from 'package:caret':
##
##     MAE, RMSE
##
## The following object is masked from 'package:base':
##
##     Recall
```

```
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
library(knitr)
```

Data Processing

```
# Read data
Crime_data <- read.csv("OriginalData.csv")
```

Processing variable 'Date'

```
# We want to firstly convert the format of date for the `lubridate` package, then extract information f
Crime_data$newDate <- mdy_hms(Crime_data$Date)
Crime_data$Hour <- hour(Crime_data$newDate)
Crime_data$WeekDay <- weekdays(Crime_data$newDate)
Crime_data$DayOfMonth <- day(Crime_data$newDate)
Crime_data$DayOfYear <- yday(Crime_data$newDate)
Crime_data$Month <- month(Crime_data$newDate, label = TRUE, abbr = FALSE)
Crime_data$Time <- hour(Crime_data$newDate)*100 + minute(Crime_data$newDate) #This format of `Time` var
Crime_data$TimeOfDay <- cut(
  hour(Crime_data$newDate),
  breaks= c(-Inf, 5, 12, 17, 20, Inf),
  labels = c("Night", "Early Morning", "Morning", "Afternoon", "Evening"),
  include.lowest = TRUE
)
rnorm(1)
```

```
## [1] 1.36677
```

Processing missing data

```
colSums(is.na(Crime_data))
```

```
##                   ID          Case.Number                 Date
##                    0                    0                    0
##                Block                 IUCR         Primary.Type
##                    0                    0                    0
##          Description Location.Description               Arrest
##                    0                    0                    0
##             Domestic                 Beat             District
##                    0                    0                    0
##                 Ward       Community.Area             FBI.Code
##                   15                    0                    0
##         X.Coordinate         Y.Coordinate                 Year
##                 2205                 2205                    0
##           Updated.On             Latitude            Longitude
##                    0                 2205                 2205
##             Location              newDate                 Hour
##                    0                    0                    0
##              WeekDay           DayOfMonth            DayOfYear
##                    0                    0                    0
##                Month                 Time            TimeOfDay
##                    0                    0                    0
```

```
rnorm(1)
```

```
## [1] 0.238726
```

We can see that the number of missing values for the variables `X.Coordinate`, `Y.Coordinate`, `Latitude`, and `Longtitude` are exactly the same, we can deduce that the coordinates are calculated from the latitude and longitude, so we don't need to include both pairs of location information. Also since the number of missing value is small compare to the number of data size, we will just eliminate the rows with missing values.

```
Crime_data %<>%  na.omit()
```

Expalnatory Data Analysis

Feature Selection

Our first task is to predict whether arrest or not given time, location, and the crime type.

```
Binary_pred_df <- Crime_data %>% select(c("ID","X.Coordinate", "Y.Coordinate","Hour","Time","WeekDay","
#???location variables
```