VI_

Sai

2025-07-06

```
library(tidyverse)
library(skimr)
library(ranger)
library(ggplot2)
library(ranger)
```

```
crime <- read_csv("./data/crime_sample.csv")
crime
```

```
## # A tibble: 1,174,559 x 21
##    highest_offense_description    highest_offense_code family_violence
##    <chr>                                         <dbl> <chr>
##  1 ASSAULT W/INJURY-FAM/DATE VIOL                  900 Y
##  2 POSS OF ALCOHOL - AGE 17 TO 20                 2209 N
##  3 POSS OF FIREARM BY FELON                       1502 N
##  4 CRIMINAL MISCHIEF                              1400 N
##  5 HARASSMENT                                     2703 N
##  6 CRIMINAL TRESPASS                              2716 N
##  7 THEFT                                           600 N
##  8 THEFT                                           600 N
##  9 MISAPPLY FIDUCIARY PROP                        1201 N
## 10 CRIMINAL TRESPASS                              2716 N
## # i 1,174,549 more rows
## # i 18 more variables: occurred_date_time <dttm>,
## #   occurred_date_time_year <dbl>, occurred_date_time_month <chr>,
## #   occurred_date_time_week_of_year <dbl>, occurred_date_time_day <dbl>,
## #   occurred_date_time_day_of_week <chr>, occurred_date_time_hour <dbl>,
## #   occurred_date_time_minute <dbl>, occurred_date <dttm>, occurred_time <dbl>,
## #   report_date_time <dttm>, report_date <dttm>, report_time <dbl>, ...
```

```
crime <- crime %>%
  mutate(
    highest_offense_description = as.factor(highest_offense_description),
    highest_offense_code        = as.factor(highest_offense_code),
    family_violence             = as.factor(family_violence),
    occurred_date_time_month    = as.ordered(occurred_date_time_month),
    occurred_date_time_day_of_week = as.ordered(occurred_date_time_day_of_week),
    location_type               = as.factor(location_type),
    apd_sector                  = as.factor(apd_sector),
    apd_district                = as.factor(apd_district)
  )
```

```
crime
```

```
## # A tibble: 1,174,559 x 21
##    highest_offense_description    highest_offense_code family_violence
##    <fct>                          <fct>                <fct>
##  1 ASSAULT W/INJURY-FAM/DATE VIOL 900                  Y
##  2 POSS OF ALCOHOL - AGE 17 TO 20 2209                 N
##  3 POSS OF FIREARM BY FELON       1502                 N
##  4 CRIMINAL MISCHIEF              1400                 N
##  5 HARASSMENT                     2703                 N
##  6 CRIMINAL TRESPASS              2716                 N
##  7 THEFT                          600                  N
##  8 THEFT                          600                  N
##  9 MISAPPLY FIDUCIARY PROP        1201                 N
## 10 CRIMINAL TRESPASS              2716                 N
## # i 1,174,549 more rows
## # i 18 more variables: occurred_date_time <dttm>,
## #   occurred_date_time_year <dbl>, occurred_date_time_month <ord>,
## #   occurred_date_time_week_of_year <dbl>, occurred_date_time_day <dbl>,
## #   occurred_date_time_day_of_week <ord>, occurred_date_time_hour <dbl>,
## #   occurred_date_time_minute <dbl>, occurred_date <dttm>, occurred_time <dbl>,
## #   report_date_time <dttm>, report_date <dttm>, report_time <dbl>, ...
```

```
rf_data <- crime %>%
  select(-occurred_date_time, -occurred_date,
         -report_date_time,   -report_date,
         -clearance_date) %>%
  mutate(family_violence = factor(family_violence)) %>% mutate(clearance_status = factor(clearance_statu

rf_data
```

```
## # A tibble: 1,174,559 x 16
##    highest_offense_description    highest_offense_code family_violence
##    <fct>                          <fct>                <fct>
##  1 ASSAULT W/INJURY-FAM/DATE VIOL 900                  Y
##  2 POSS OF ALCOHOL - AGE 17 TO 20 2209                 N
##  3 POSS OF FIREARM BY FELON       1502                 N
##  4 CRIMINAL MISCHIEF              1400                 N
##  5 HARASSMENT                     2703                 N
##  6 CRIMINAL TRESPASS              2716                 N
##  7 THEFT                          600                  N
##  8 THEFT                          600                  N
##  9 MISAPPLY FIDUCIARY PROP        1201                 N
## 10 CRIMINAL TRESPASS              2716                 N
## # i 1,174,549 more rows
## # i 13 more variables: occurred_date_time_year <dbl>,
## #   occurred_date_time_month <ord>, occurred_date_time_week_of_year <dbl>,
## #   occurred_date_time_day <dbl>, occurred_date_time_day_of_week <ord>,
## #   occurred_date_time_hour <dbl>, occurred_date_time_minute <dbl>,
## #   occurred_time <dbl>, report_time <dbl>, location_type <fct>,
## #   apd_sector <fct>, apd_district <fct>, clearance_status <fct>
```

Sampling 50,000 rows.

```
set.seed(1906525)

# 1. compute the overall fraction we need
frac <- 50e3 / nrow(rf_data)

# 2. do a proportional (stratified) draw within each clearance_status
sampled <- rf_data %>%
  group_by(clearance_status) %>%
  slice_sample(prop = frac) %>%
  ungroup() %>%
  # 3. in case of rounding you might get  50k, so force exactly 50k
  slice_sample(n = 50e3)

sampled
```

```
## # A tibble: 49,999 x 16
##    highest_offense_description    highest_offense_code family_violence
##    <fct>                          <fct>                <fct>
##  1 CUSTODY ARREST TRAFFIC WARR    3722                 N
##  2 BURGLARY OF VEHICLE            601                  N
##  3 THEFT OF TRAILER               613                  N
##  4 FAMILY DISTURBANCE             3400                 N
##  5 INTER EMERG PHONECALL FAM/DATE 2712                 N
##  6 THEFT FROM AUTO                603                  N
##  7 POSS CONTROLLED SUB/NARCOTIC   1800                 N
##  8 ASSAULT W/INJURY-FAM/DATE VIOL 900                  Y
##  9 CRIMINAL TRESPASS              2716                 N
## 10 POSS CONTROLLED SUB/NARCOTIC   1800                 N
## # i 49,989 more rows
## # i 13 more variables: occurred_date_time_year <dbl>,
## #   occurred_date_time_month <ord>, occurred_date_time_week_of_year <dbl>,
## #   occurred_date_time_day <dbl>, occurred_date_time_day_of_week <ord>,
## #   occurred_date_time_hour <dbl>, occurred_date_time_minute <dbl>,
## #   occurred_time <dbl>, report_time <dbl>, location_type <fct>,
## #   apd_sector <fct>, apd_district <fct>, clearance_status <fct>
```

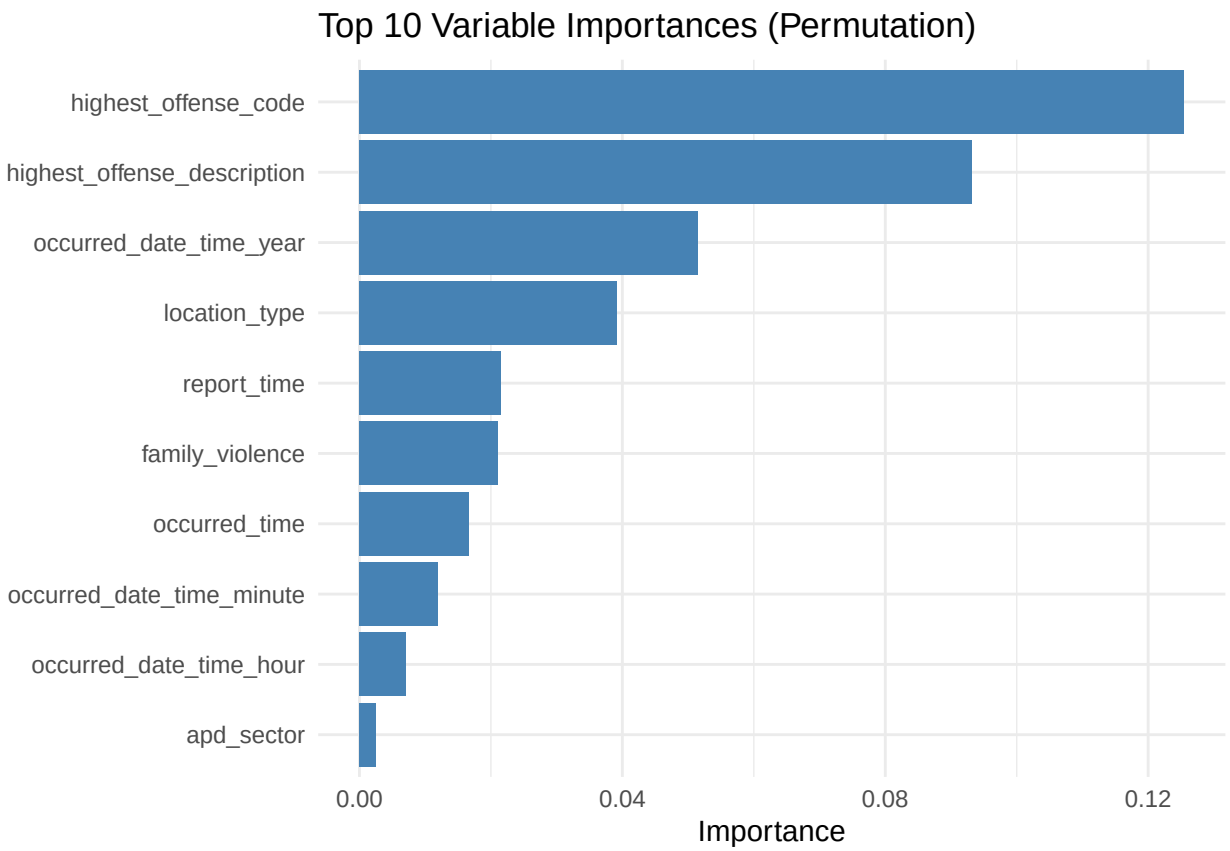
Variable Importance Plot

```
# 1. Fit a ranger with permutation importance
library(vip)
library(tibble)
library(dplyr)
library(ggplot2)

set.seed(1906525)
rf_mod_perm <- ranger(
  formula      = clearance_status ~ .,
```

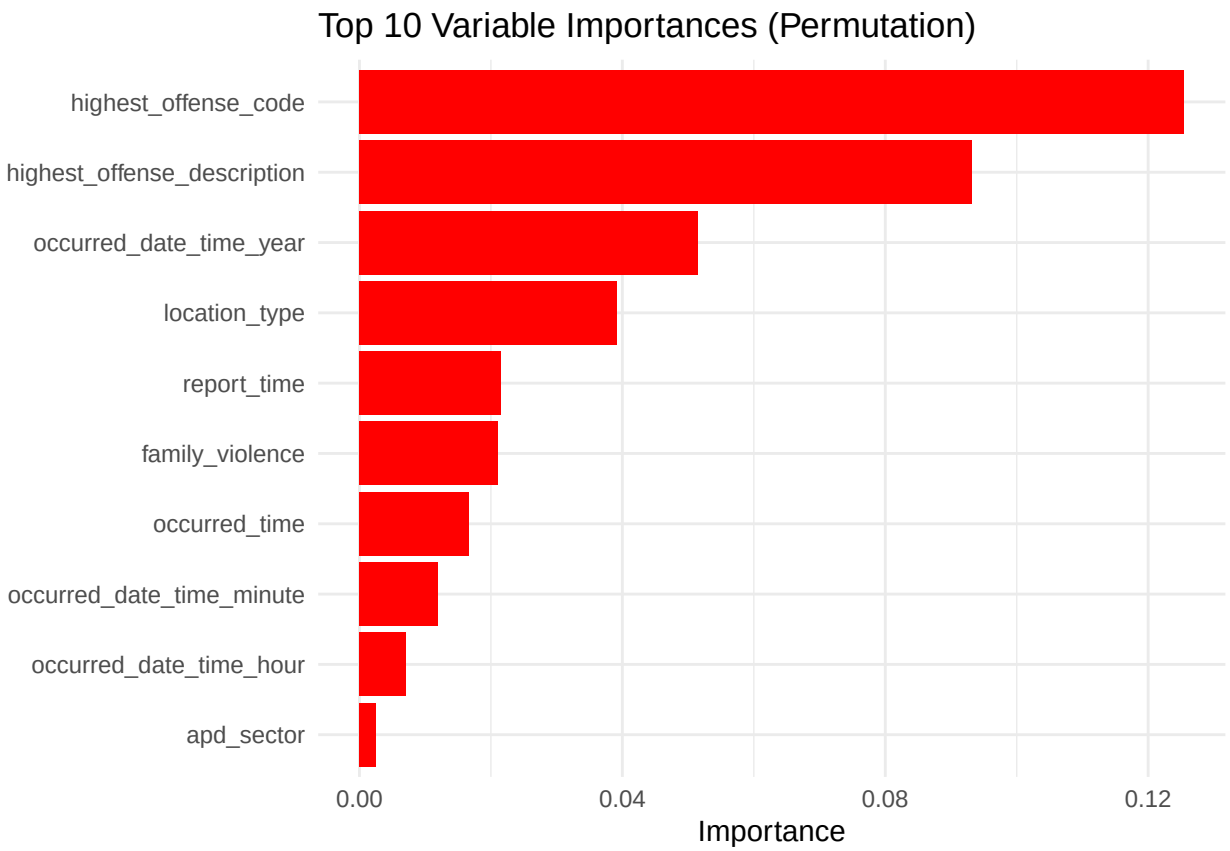
```
  data         = sampled,
  importance   = "permutation",
  num.trees    = 200,
  write.forest = TRUE,              # needed for OOB permutations
  num.threads  = parallel::detectCores()
)

# 2. Extract the top 10 importances into a tibble
imp_tibble <- rf_mod_perm$variable.importance %>%
  enframe(name = "Variable", value = "Importance") %>%
  arrange(desc(Importance)) %>%
  slice_head(n = 10)

# View the tibble
print(imp_tibble)
```

```
## # A tibble: 10 x 2
##    Variable                     Importance
##    <chr>                             <dbl>
##  1 highest_offense_code            0.125
##  2 highest_offense_description     0.0932
##  3 occurred_date_time_year         0.0516
##  4 location_type                   0.0391
##  5 report_time                     0.0216
##  6 family_violence                 0.0210
##  7 occurred_time                   0.0167
##  8 occurred_date_time_minute       0.0120
##  9 occurred_date_time_hour         0.00712
## 10 apd_sector                      0.00249
```

```
#> # A tibble: 10 × 2
#>    Variable           Importance
#>    <chr>                   <dbl>
#>  1 some_top_feature        1.23
#>  2 next_feature            0.87
#>  …                       …

# 3a. Plot with vip (drop-in for your existing code)
vip(
  rf_mod_perm,
  num_features = 10,
  geom         = "col",
  aesthetics   = list(fill = "steelblue")
) +
  coord_flip() +
  labs(
    title = "Top 10 Variable Importances (Permutation)",
    x     = NULL,
    y     = "Importance"
  ) +
  theme_minimal()
```




```r
# 3b. Alternative: ggplot2 using the tibble you just made
imp_tibble %>%
  ggplot(aes(x = reorder(Variable, Importance), y = Importance)) +
  geom_col(fill = "red") +
  coord_flip() +
  labs(
    title = "Top 10 Variable Importances (Permutation)",
    x     = NULL,
    y     = "Importance"
  ) +
  theme_minimal()
```


Top 10 Variable Importances (Permutation)
highest_offense_code
highest_offense_description
occurred_date_time_year
location_type
report_time
family_violence
occurred_time
occurred_date_time_minute
occurred_date_time_hour
apd_sector
0.00
0.04
0.08
0.12
Importance