Présentation des données

Thomas Fernandes

2023-11-09

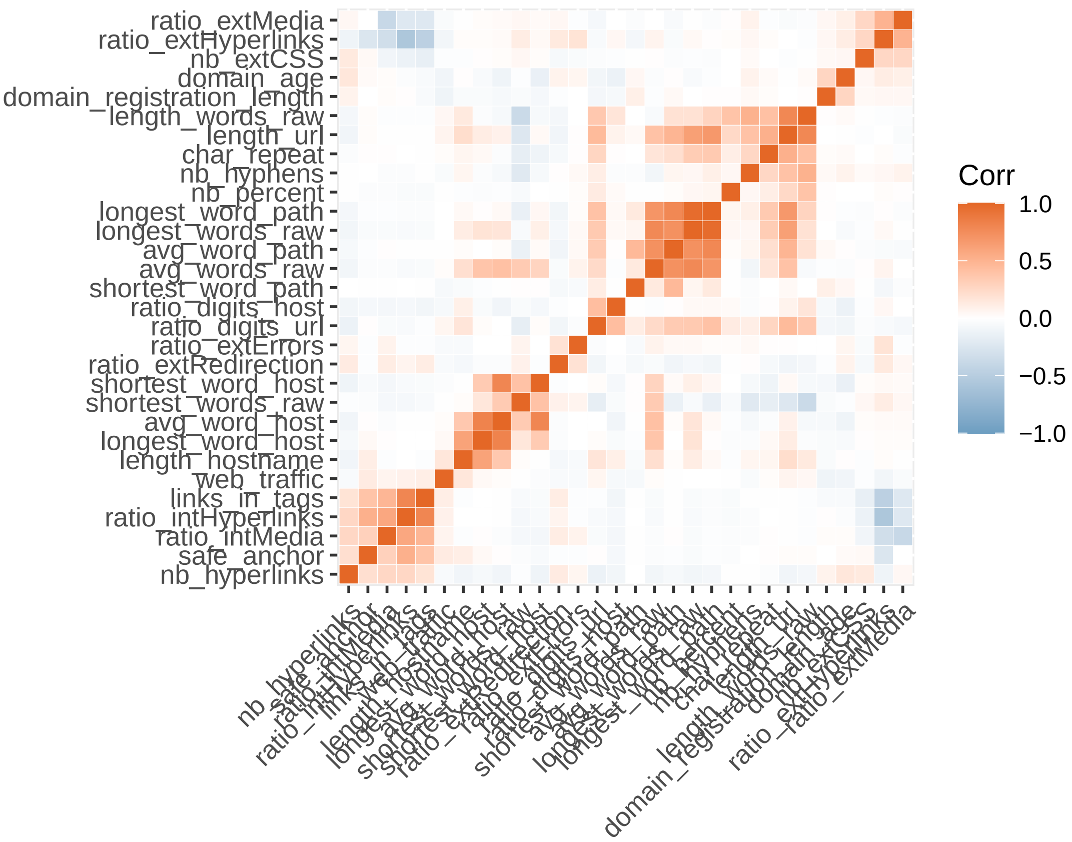
Présenation des données

1.

```r
df_data$target <- as.integer(df_data$status == 'legitimate')
df_data <- df_data[, !names(df_data) %in% c('status')]
```

```r
tmp <- data.frame(missing_val = colSums(is.na(df_data)))
tmp <- tmp[tmp$missing_val != 0, ]

likely_cat <- vector("logical", length = ncol(df_data) - 1)
for (i in 2:ncol(df_data)) {
  likely_cat[[i]] <- (length(unique(df_data[[i]])) / sum(!is.na(df_data[[i]]))) < 0.002
}

num_cols <- character()
cat_cols <- character()

for (i in 1:length(likely_cat)) {
  if (!likely_cat[[i]]) {
    num_cols <- c(num_cols, names(df_data)[i])
  } else {
    cat_cols <- c(cat_cols, names(df_data)[i])
  }
}
```

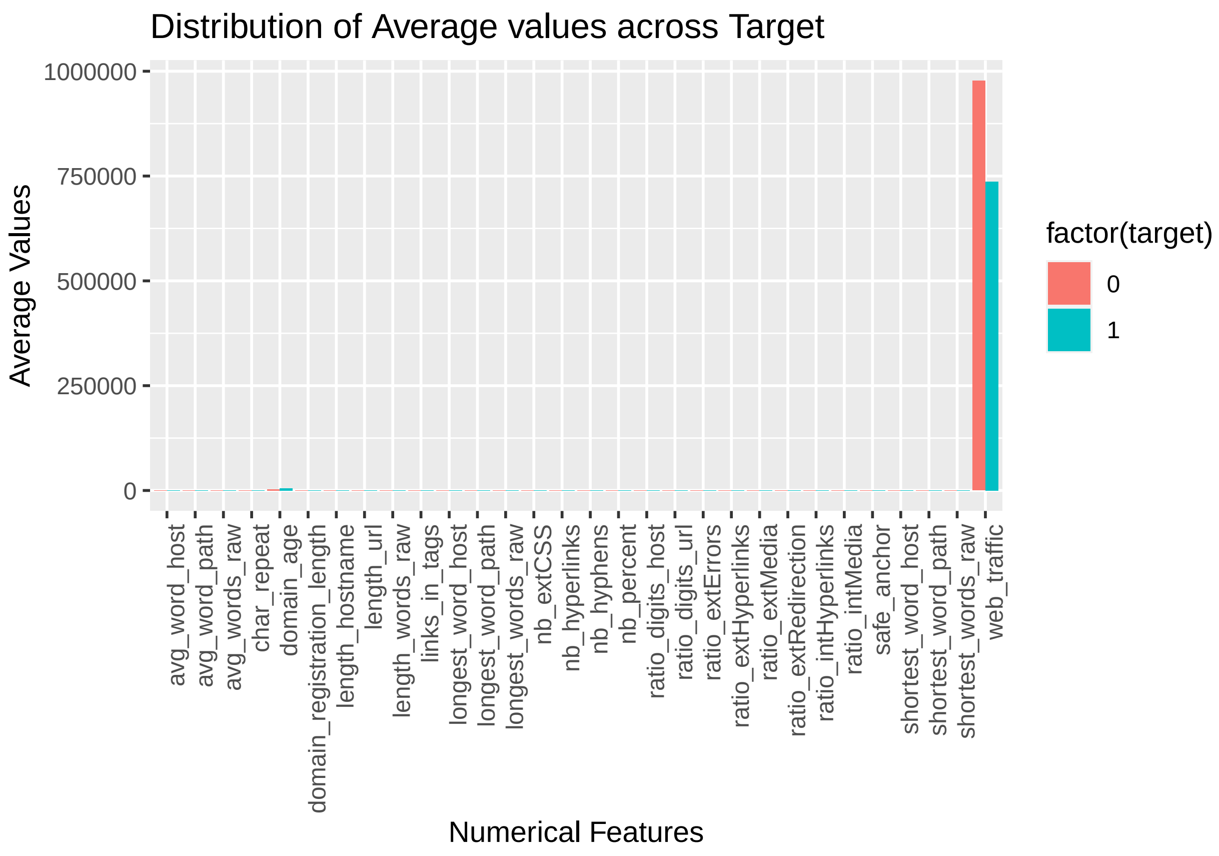
```r
corr <- cor(df_data[num_cols])

ggcorrplot(
  corr,
  hc.order = TRUE,
  type = "full",
  outline.color = "white",
  ggtheme = ggplot2::theme_gray,
  colors = c("#6D9EC1", "white", "#E46726"),
  show.diag = TRUE,
  tl.cex = 10
)
```



```r
df_distr <- df_data %>%
  group_by(target) %>%
  summarise(across(num_cols, mean)) %>%
  pivot_longer(cols = -target, names_to = "feature", values_to = "value")
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(num_cols, mean)`.
## Caused by warning:
## ! Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(num_cols)
##
##   # Now:
##   data %>% select(all_of(num_cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```r
ggplot(df_distr, aes(x = feature, y = value, fill = factor(target))) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Numerical Features", y = "Average Values", title = "Distribution of Average values across Ta
```


Distribution of Average values across Target
Average Values
1000000
750000
500000
250000
0
avg_word_host
avg_word_path
avg_words_raw
char_repeat
domain_age
domain_registration_length
length_hostname
length_url
length_words_raw
links_in_tags
longest_word_host
longest_word_path
longest_words_raw
nb_extCSS
nb_hyperlinks
nb_hyphens
nb_percent
ratio_digits_host
ratio_digits_url
ratio_extErrors
ratio_extHyperlinks
ratio_extMedia
ratio_extRedirection
ratio_intHyperlinks
ratio_intMedia
safe_anchor
shortest_word_host
shortest_word_path
shortest_words_raw
web_traffic
Numerical Features
factor(target)
0
1