Présentation des données

Thomas Fernandes

2023-11-09

L'étude vise à prédire la légitimité des sites web en utilisant diverses techniques de machine learning. Le phénomène du phishing consiste en des tentatives de fraude en ligne par le biais de sites web frauduleux imitant des sites légitimes.

La variable que nous cherchons à prédire est "status", qui indique si un site web est légitime ou potentiellement frauduleux (phishing). Pour ce faire, nous disposons d'un ensemble de données équilibré de 88 variables explicatives différentes, chacune fournissant des informations sur divers aspects de 11430 sites web différents. Ces données incluent 56 variables basées sur la structure, 24 extraites du contenu des pages web correspondantes, 7 obtenues par des requêtes auprès de services externes.

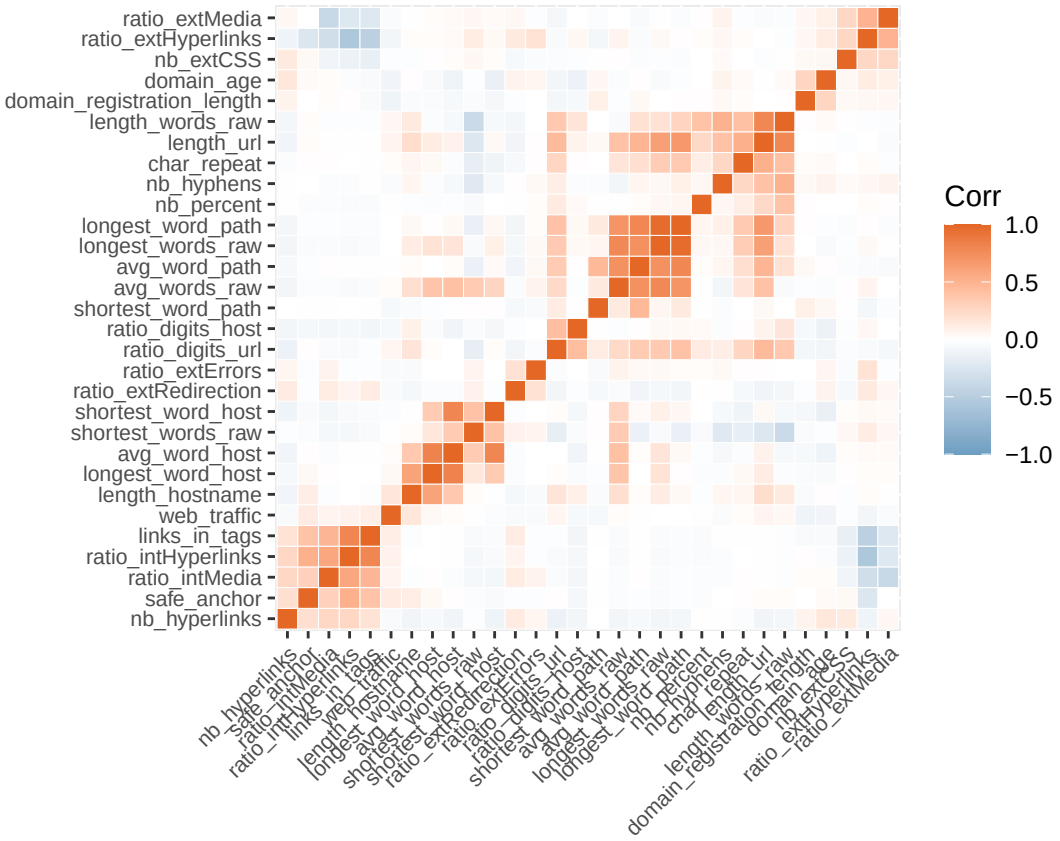
Présentation des données

Avant de commencer les différentes modélisations, nous allons regarder comment se structurent nos données.

1. Corrélation entre les variables quantitatives

```
df_present <- df

#Extraire les variables qualitatives
v_quali <- vector("logical", length = ncol(df_present) - 1)
for (i in 2:ncol(df_present)) {
  v_quali[[i]] <- (length(unique(df_present[[i]])) / sum(!is.na(df_present[[i]]))) < 0.002
}

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

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

corr <- cor(df_present[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 = 8
)
```



On remarque que de nombreuses variables sont corrélées entre elles.

```
mean_by_status <- function(df, col_name) {
  df %>%
    group_by(status) %>%
    summarise(mean_value = mean(.data[[col_name]], na.rm = TRUE))
}
# Liste pour stocker les résultats des variables qualitatives
mean_values_list_cat <- list()

# Calculer la moyenne pour chaque colonne qualitative
for (col in cat_cols) {
  mean_values_list_cat[[col]] <- mean_by_status(df_present, col)
}
```

```
## Warning: There were 2 warnings in `summarise()`.
## The first warning was:
## i In argument: `mean_value = mean(.data[["status"]], na.rm = TRUE)`.
## i In group 1: `status = legitimate`.
## Caused by warning in `mean.default()`:
## ! l'argument n'est ni numérique, ni logique : renvoi de NA
```

```
## i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```r
# Convertir les résultats en un dataframe utilisable pour ggplot
mean_values_df_cat <- do.call(rbind, mean_values_list_cat)
mean_values_df_cat$col_names <- rownames(mean_values_df_cat)

# Tracer le graphique à barres pour les variables qualitatives
ggplot(mean_values_df_cat, aes(x = col_names, y = mean_value, fill = status)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Variables", y = "Moyenne", title = "Moyenne des variables qualitatives par statut") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("#E46726", "#6D9EC1"))
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```