

# Interpretable Machine Learning PD5

*Daniel Ponikowski*

*14 kwietnia 2019*

## Wybrane zmienne :

1. ppwork - aktualny status zatrudnienia
2. w6\_q20 - czy obecnie mieszkasz z partnerem?
3. Q21A\_Year - w którym roku pierwszy raz spotkałeś partnera?
4. ppage - wiek

## Wczytanie danych:

```
data <- read.dta13(file = "../PD1/HCMST 2017 fresh sample for public sharing draft v1.1.dta")
df <- data[,c("S1", "ppwork", "w6_q19", "Q21A_Year", "ppage")]
df <- df %>% mutate(Q21A_Year = as.numeric(as.character(Q21A_Year))
                    , ppwork = factor(ppwork)
                    , w6_q19 = factor(w6_q19)
                    , ppage = as.numeric(ppage)
                    , S1 = factor(S1)) %>%
  na.omit() %>% unique() %>% as.data.frame()
row.names(df) <- 1:nrow(df)
```

## Modele

Użyj dwóch modeli, różniących się strukturą. Pierwszym modelem, będzie las losowy używany w poprzednich pracach domowych, natomiast drugim modelem będzie regresja logistyczna.

```
RF <- readRDS("../PD3/randomForestPD1.rds")
logit <- train(S1~., df, "glmnet", family = "binomial")
```

## PD5

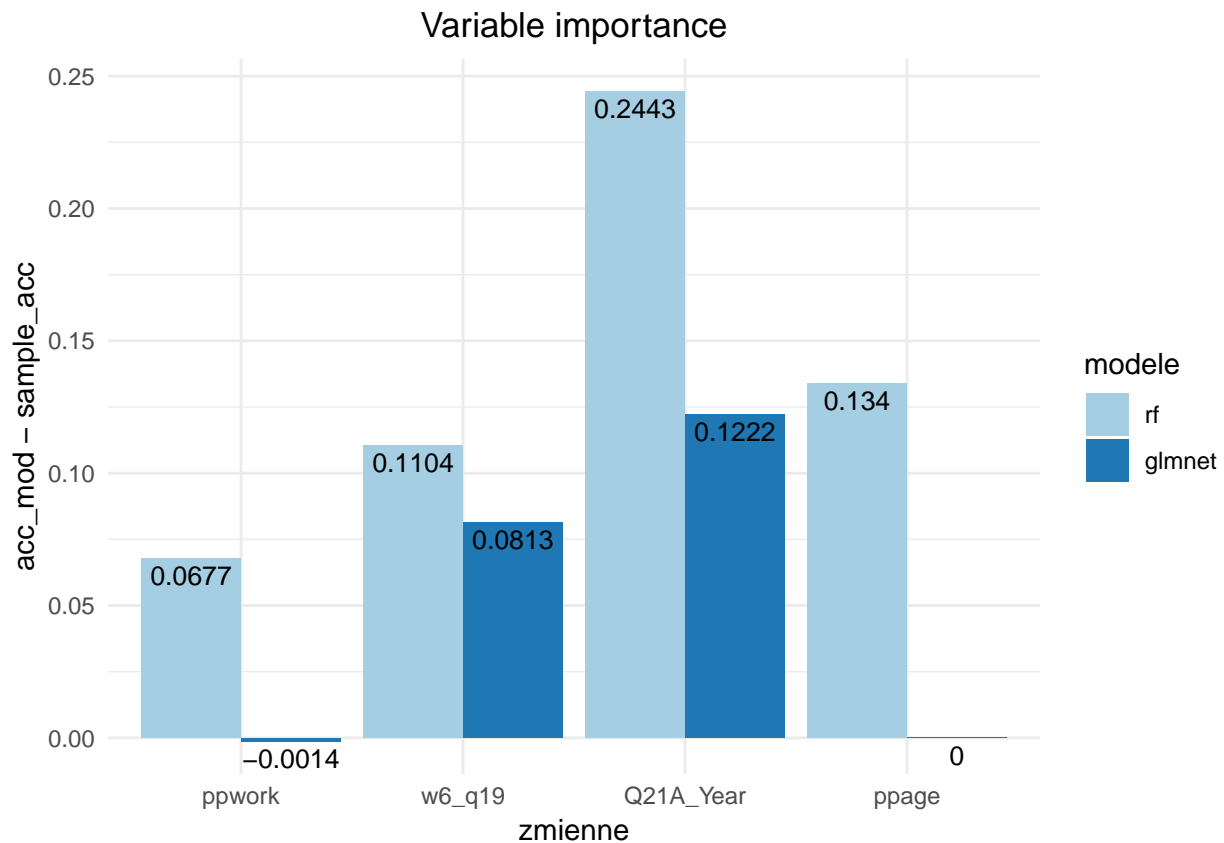
```
Accuracy <- function(y_pred, y){
  mat <- table(y_pred, y)
  sum(diag(mat))/sum(mat)
}
variable_importance <- function(model, data_X, Y){
  importance <- list()
  data <- data_X
  basic_accuracy <- Accuracy(predict(model, data_X), Y)
  for (zmienna in colnames(data_X)){
    data[[zmienna]] <- sample(data[[zmienna]])
    accuracy_zmienna <- Accuracy(predict(model, data), Y)
    importance[[zmienna]] <- basic_accuracy - accuracy_zmienna
    data <- data_X
  }
  importance <- unlist(importance)
  importance
}
```

```
variable_importance_plot <- function(lista_modeli,data_X,Y){
  wyniki <- lapply(lista_modeli,function(x) variable_importance(model = x,data_X = data_X,
                                                                Y = Y)) %>%
    unlist() %>% matrix(nrow = ncol(data_X))
  colnames(wyniki) <- lapply(lista_modeli,function(x) x$method)
  rownames(wyniki) <- colnames(data_X)
  wyniki <- melt(wyniki,varnames = c("zmienne","modele")) %>%
    mutate(value = round(value,4))

  ggplot(wyniki,aes(x = zmienne,y = value, fill = modele)) +
    geom_bar(stat = "identity", position=position_dodge()) +
    geom_text(aes(label=value), position = position_dodge(0.9),vjust = 1.4,
              color="black", size=3.5) +
    theme_minimal() + scale_fill_brewer(palette="Paired") +
    ggtitle(label = "Variable importance") +
    theme(plot.title = element_text(hjust = 0.5)) +
    ylab(label = "acc_mod - sample_acc")
}
```

Waznosc zmiennych dla modeli:

```
variable_importance_plot(list(RF,logit),df[, -1],df$S1)
```



## PD plot

Przedstawie PD ploty zmiennej ppage (wiek), która dla lasu losowego jest ważną zmienną (druga co do ważności), a regresja logistyczna całkowicie ignoruje tę zmienną (różnica pomiędzy accuracy bazowym i accuracy po zmianie tej zmiennej jest bardzo sobie bliska).

```
PD <- function(model,data_X,zmienna){
  result <- data.frame(zmienna = min(data_X[[zmienna]]):max(data_X[[zmienna]]))
  for (i in 1:nrow(data_X)){
    n <- nrow(result)
    Q21A_Year <- rep(data_X$Q21A_Year[i], n)
    ppwork <- rep(data_X$ppwork[i], n)
    w6_q19 <- rep(data_X$w6_q19[i], n)
    df <- data.frame(ppage = result$zmienna, ppwork = ppwork,w6_q19=w6_q19,
                     Q21A_Year=Q21A_Year)
    result[[i+1]] <- predict(model,df,"prob")[,1]
  }
  ppage <- result$zmienna
  result$zmienna <- NULL
  prob <- as.data.frame(cbind(ppage,prob_of_married = apply(result,1,mean)))
  prob
}

PD_plot <- function(lista_modeli,data_X,zmienna){
  wynik <- lapply(lista_modeli,function(x) PD(model = x,data_X = data_X
                                              ,zmienna = zmienna)) %>%
    bind_rows(.id = "model")
  colnames(wynik) <- c("model","zmienna","prob_of_married")
  ggplot(wynik,aes(x=zmienna,y=prob_of_married)) +
    geom_line(aes(colour = model)) +
    ggtitle(paste("PD plot zmiennej",zmienna)) +
    theme(plot.title = element_text(hjust = 0.5)) + xlab(zmienna)
}
```

## PD plot



## Komentarz

Wpływ zmiennej *ppage* (wiek) nie jest wychwytywany przez model regresji logistycznej, może to wynikać z ograniczenia modelu (założenie liniowej zależności), las losowy jest w stanie wychwycić zależności nieliniowe. Różnica w skrajnych punktach PD plotu, dla regresji liniowej, wynosi ok. 0.02, więc niewiele. PD plot dla lasu losowego pokazuje większe wahania, możemy to interpretować następująco, dla lasu losowego odpowiedź modelu zależy od wartości tej zmiennej, gdzie w przypadku regresji liniowej wartość tej zmiennej w niewielkim stopniu wpływa na odpowiedź modelu.

Najważniejsza zmienna dla obu modeli okazała się zmienna *Q21A\_Year* (w którym roku pierwszy raz spotkałeś partnera?).