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 ktorym roku pierwszy raz spotkales 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

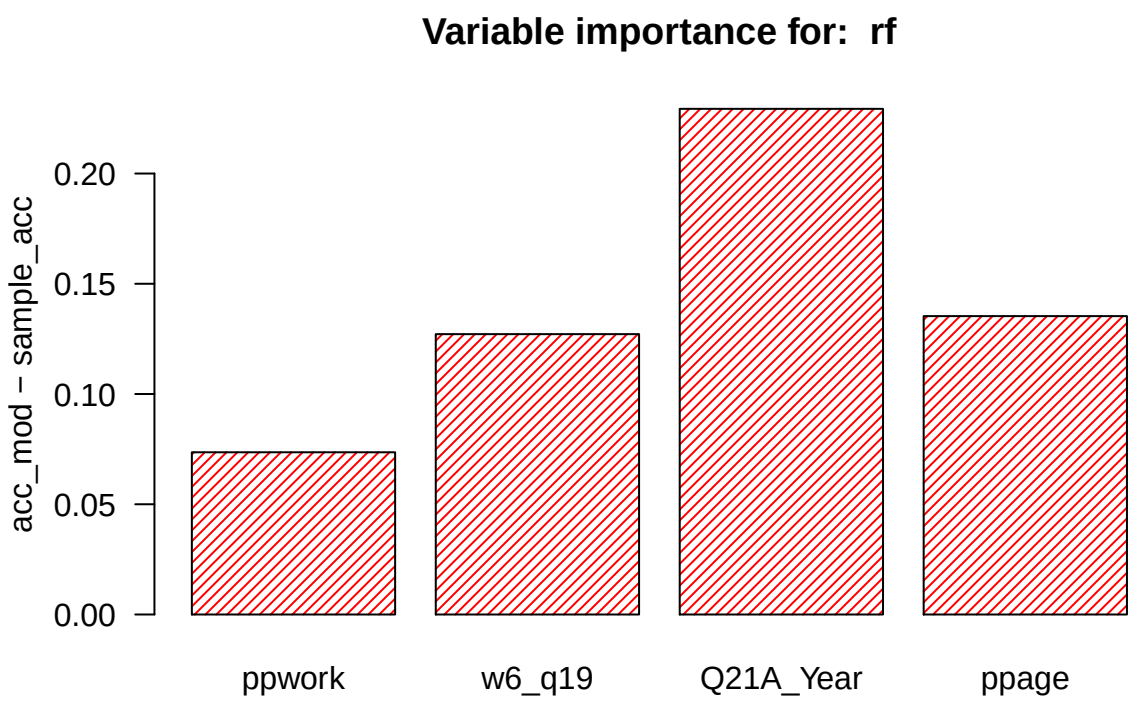
Uzyje dwoch modeli, rozniacych sie struktura. Pierwszym modelem, bedzie las losowy uzywany w poprzednich pracach domowych, natomiast drugim modelem bedzie regresja logistyczna.

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

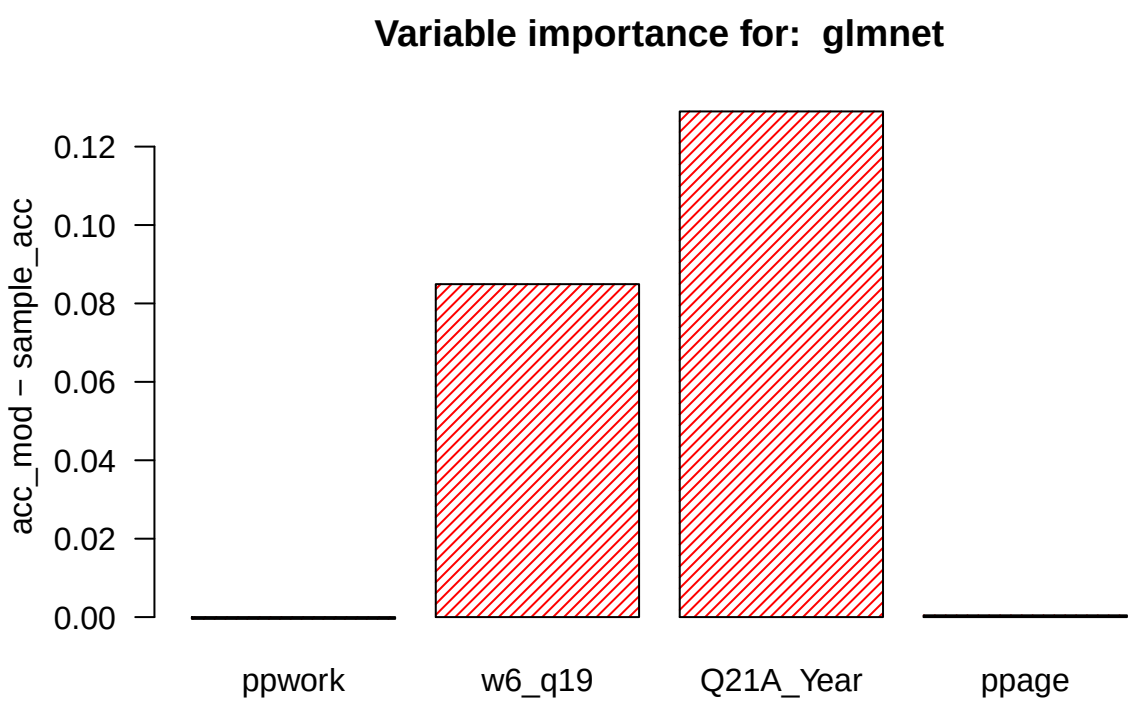
```
Accuracy <- function(y_pred,y){
  mat <- table(y_pred,y)
  sum(diag(mat))/sum(mat)
}
variable_importance <- function(model,data_X,Y){
  importance <- list()
  data <- data_X
  basic_accuracy <- Accuracy(predict(model,data_X),Y)
  for (zmienna in colnames(data_X)){
    data[[zmienna]] <- sample(data[[zmienna]])
    accuracy_zmienna <- Accuracy(predict(model,data),Y)
    importance[[zmienna]] <- basic_accuracy - accuracy_zmienna
    data <- data_X
  }
  importance <- unlist(importance)
  barplot(importance,las=1,col = "red",angle = 45,density = 25,border = "black",
          main = paste("Variable importance for: ",model$method),
          ylab = "acc_mod - sample_acc")
  }
```

waznosc zmiennych dla modeli:

```
variable_importance(RF,df[,-1],df$S1)
```




```
variable_importance(logit,df[,-1],df[,1])
```

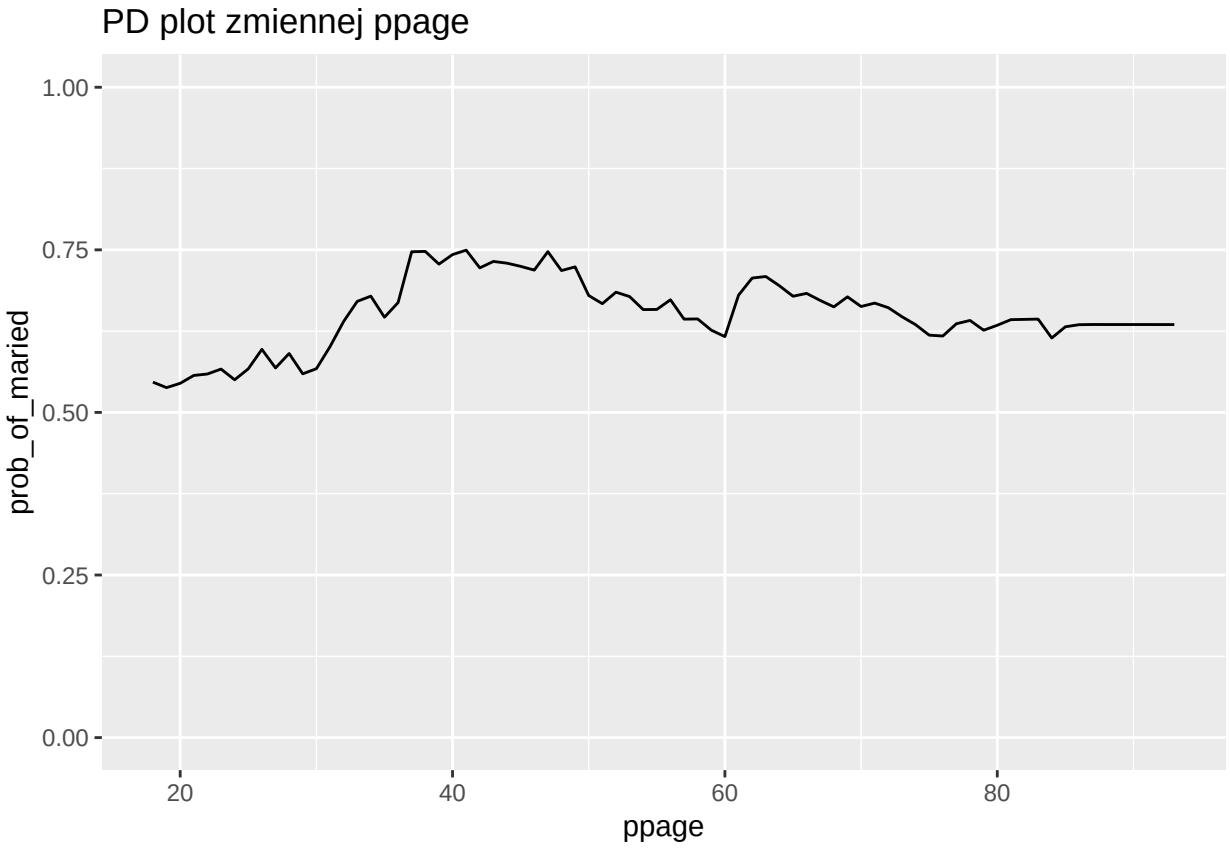


PD plot

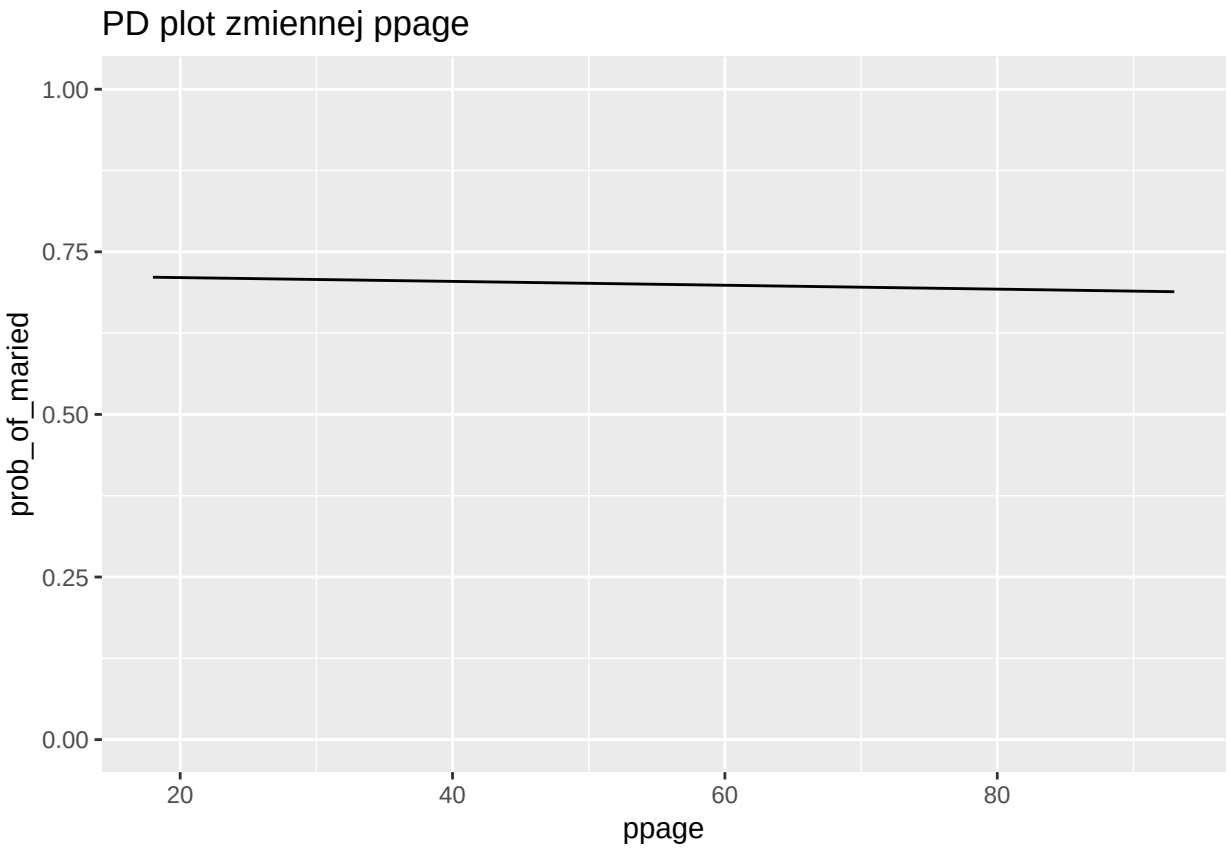
Przedstawie PD ploty zmiennej ppage (wiek), ktora dla lasu losowego jest wazna zmienna (druga co do waznosci), a regresja logistyczna calkowicie ignoruje ta zmienna (roznica pomiedzy accuracy bazowym i accuracy po zmianie tej zmiennej jest bardzo sobie bliskie).

```
PD_plot <- 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_maried = apply(result,1,mean)))
  prob
  ggplot(prob,aes(x=ppage,y=prob_of_maried)) + geom_line() +
  ggtitle(paste("PD plot zmiennej",zmienna)) + ylim(0,1)
}
```

PD plot las losowy



PD plot regresja logistyczna



Komentarz

Wpływ zmiennej *ppage* (wiek) nie jest wychwytywany przez model regresji logistycznej, moze to wynikac z ograniczenia modelu (zaloenie liniowej zaleznosci), las losowy jest w stanie wychwycić zależności nieliniowe. Roznica w skrajnych punktach PD plotu, dla regresji liniowej, wynosi ok. 0.02, wiec niewiele. PD plot dla lasu losowego pokazuje wieksze wahania, mozemy to interpretowac nastepujaco, dla lasu losowego odpowiedz modelu zalezy od wartosci tej zmiennej, gdzie w przypadku regresji liniowej wartosc tej zmiennej w niewielkim stopniu wplywa na odpowiedz modelu.

Najwazniejsza zmienna dla obu modeli okazala sie zmienna *Q21A_Year* (w ktorym roku pierwszy raz spotkales partnera?).