Interpretable Machine Learning PD6

Daniel Ponikowski

27 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 modelu regresji logistycznej

```
logit <- train(S1~.,df,"glmnet",family = "binomial")
```

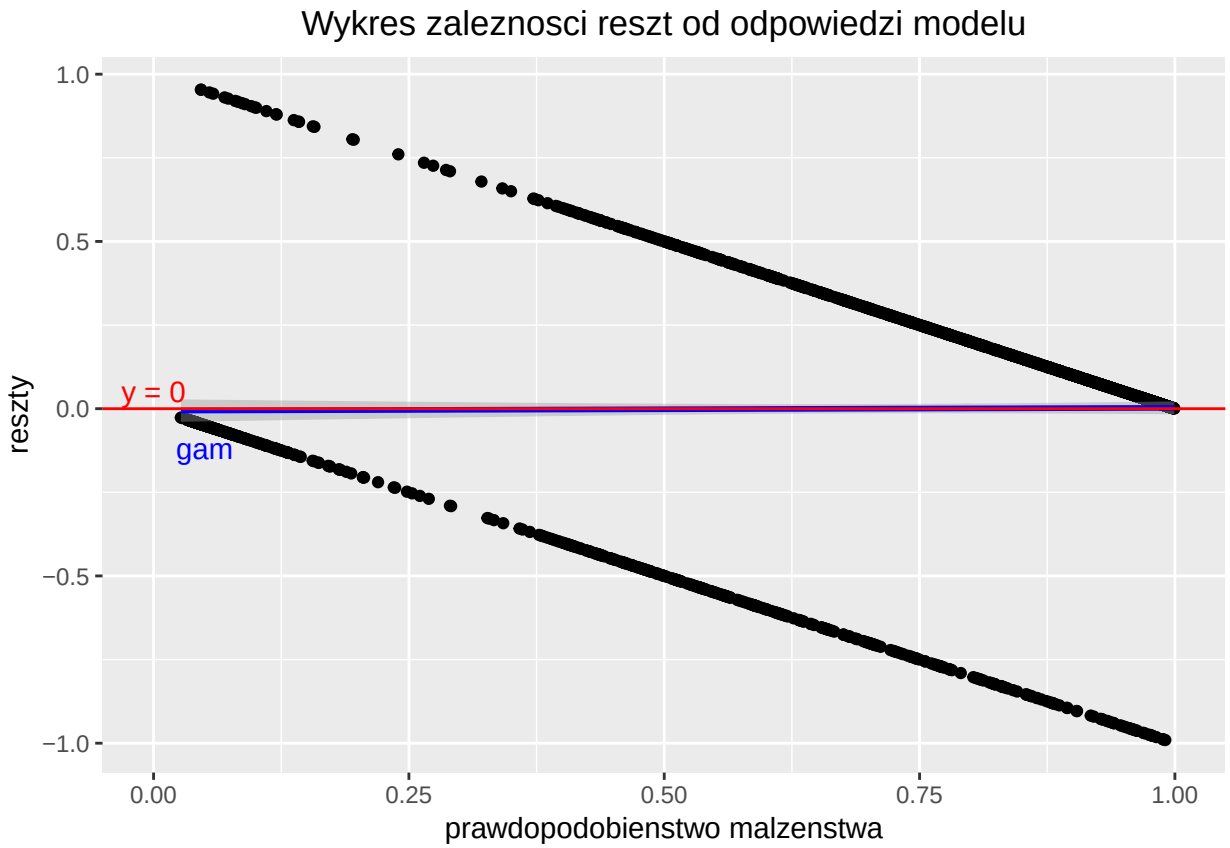
PD6

Wyznaczanie reszt

```
predykcja <- predict(logit,df,"prob")[,1]
y <- ifelse(df$S1 == "Yes, I am Married",yes = 1,no = 0)

reszty <- y - predykcja
```

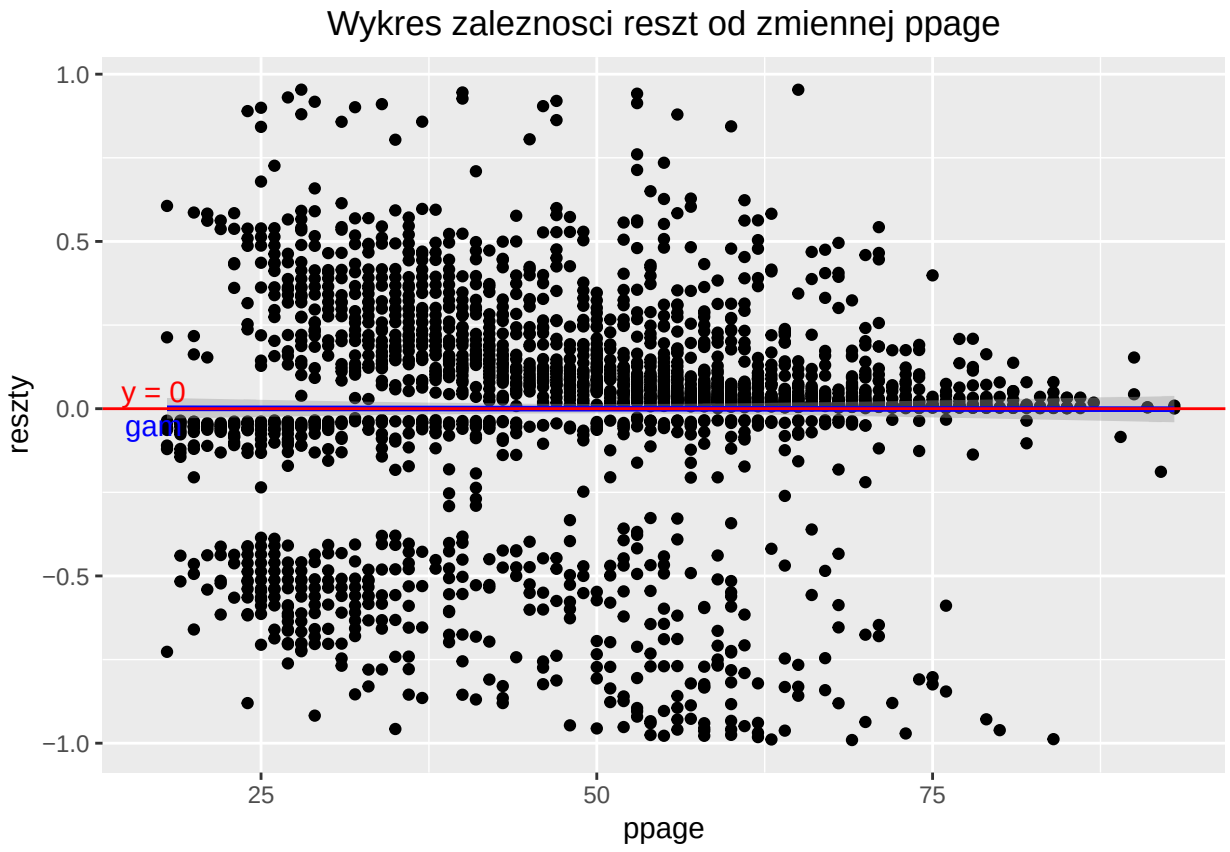
Wykresy

```
df_reszty <- data.frame(predykcja,y,reszty)

ggplot(data = df_reszty,aes(x = predykcja, y = reszty)) + geom_point(colour = "black") +
  geom_smooth(method = "gam",colour = "blue") + geom_hline(yintercept = 0,colour = "red") +
  annotate(geom="text", x=0, y=0.05, label="y = 0",color="red") +
  annotate(geom = "text",x = 0.05,y = -0.12,label = "gam",color = "blue") +
  ggtitle(label = "Wykres zaleznosci reszt od odpowiedzi modelu") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab(label = "prawdopodobienstwo malzenstwa")
```



Mozna zauwazyc ze model regresji logistycznej myli sie calkowicie dla niektorych wartosc tzn. zwraca prawdopodobienstwo malzenstwa bliskie 1 dla osob ktore nie sa w malzenstwie, a takze dla osob nie bedacych w malzenstwie przewiduje prawdopodobienstwo malzenstwa bliskie 1. Jednak krzywa lokalnego trendu (**gam**), praktycznie pokrywa się z prosta stale rowna 0, czyli model myli się podobnie w "jedna jak i druga strone".

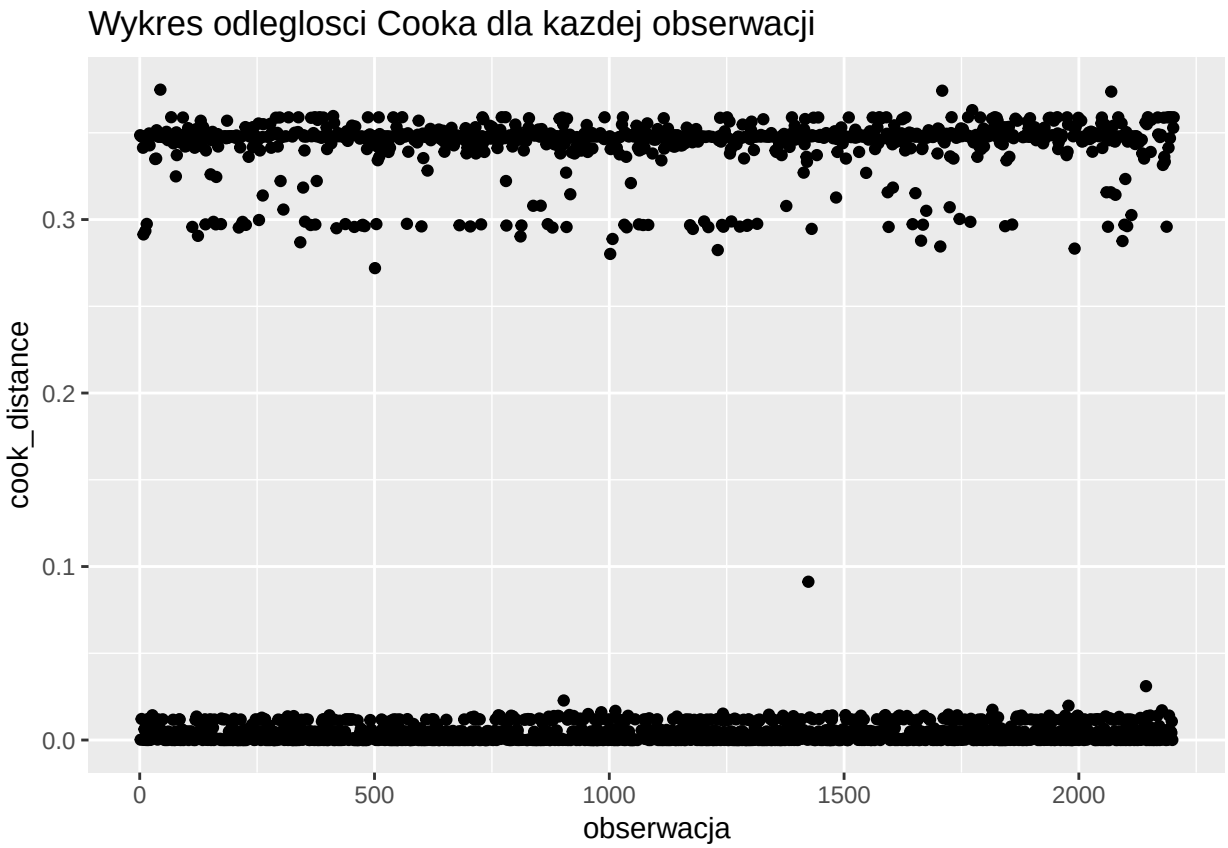
```
df_reszty$ppage <- df$ppage

ggplot(data = df_reszty,aes(x = ppage, y = reszty)) + geom_point(colour = "black") +
  geom_smooth(method = "gam",colour = "blue") + geom_hline(yintercept = 0,colour = "red") +
  annotate(geom = "text",x = min(df_reszty$ppage)-1, y = 0.05, label = "y = 0",color="red") +
  annotate(geom = "text",x = min(df_reszty$ppage)-1,y = -0.05, label = "gam",color = "blue") +
  ggtitle(label = "Wykres zaleznosci reszt od zmiennej ppage") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab(label = "ppage")
```



Tutaj zauwazamy, ze wartosci reszt dla wiekszej ilosci obserwacji są ponad prosta $\mathbf{y = 0}$, jednak krzywa lokalnego trendu jest bardzo zblizona do prostej o rownaniu $\mathbf{y = 0}$.

Odleglosci Cooka

```
y_pred <- df_reszty$predykcja
cook_distance <- data.frame(obserwacja = 1:nrow(df),cook_distance = numeric(nrow(df)))

for (i in 1:nrow(df)){
  reg_log <- train(S1~.,data = df[-i,],method = "glmnet", family = "binomial")
  y_pred_bez_i <- predict(reg_log,df,"prob")[,1]
  cook_distance$cook_distance[i] <- sum((y_pred-y_pred_bez_i)^2)
   }

ggplot(cook_distance,aes(x = obserwacja, y = cook_distance)) + geom_point() +
  ggtitle(label = "Wykres odleglosci Cooka dla kazdej obserwacji")
```



Zadna z obserwacji nie wybija się w zdecydowany sposob sposrod pozostalych, wiec mozemy uznac, ze zadna z obserwacji nie jest bardzo wplywowa. Utworzenie sie grup obserwacji o podobnej wartosci odleglosci Cooka, pokazuje ze obserwacje te maja podobny wplyw na estymowane parametry. Moga to byc obserwacje o podobnych wartosciach zmiennych (np. rozniace sie wartoscia tylko jednej zmiennej).