hm5

Robert Benke

22 kwietnia 2019

Homework V

Zadanie 1.1

Wybrane zmienne:

- time_from_rel_to_cohab - czas pomiedzy poznaniem a rozpoczęciem relacji
- hcm2017q24_college - poznali sie na uniwersytecie

- hcm2017q24_bar_restaurant - poznali się w barze/restauracji/itp.
- partner_yrsed - liczba lat jaką pertner spędził na edukcji

Zadanie 1 - random forest i regresja logistyczna

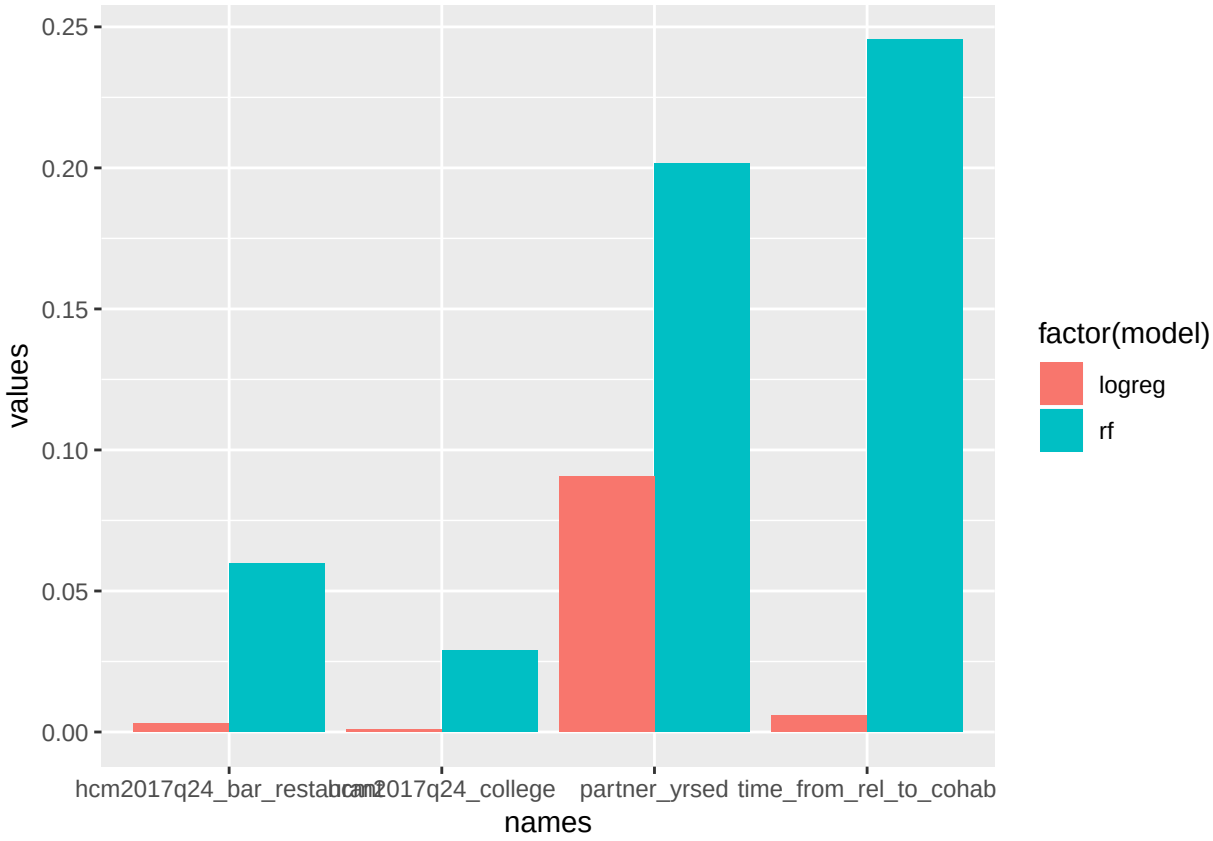
Zadanie 2 - spadek funkcji loss

Do porównania istotności zmiennych wykorzystana została miara jakości modelu 'AUC'. Jest to pole pod krzywą ROC, która jest zależnością TPR (true positive rate) od FPR (false positive rate) dla różnych punktów odcięcia. Wyższa wartość AUC oznacza lepszy model (chociaż mogą występować punkty odcięcia dla których model z niższym AUC zachowuje się lepiej).

```
perturbationImportance <- function(data = data_dfr, target_ind, model, lrn, title){

  preditction_vec <- predictLearner(lrn,model,data)[,2]
  prediction_ROCR <- ROCR::prediction(preditction_vec, data[,target_ind])
  global_auc <- ROCR::performance(prediction_ROCR, measure = "auc")@y.values[[1]]

  perturb_auc <- numeric(ncol(data)-1)
  for (i in 1:(ncol(data)-1)){
    data_pert <- data[,-target_ind]
    data_pert[,i] <- sample(data_pert[,i], nrow(data_pert), replace = FALSE)

    preditction_vec <- predictLearner(lrn,model,data_pert)[,2]
    prediction_ROCR <- ROCR::prediction(preditction_vec, data[,target_ind])
    perturb_auc[i] <-global_auc - ROCR::performance(prediction_ROCR, measure = "auc")@y.values[[1]]
  }

  results_dfr <- data.frame(names = names(data[,-target_ind]), values = perturb_auc)


}
```

Zadanie 3 - porównanie modeli

```
perturbation_logreg_dfr <- perturbationImportance(data = data_dfr, target_ind = 5, model = model_logreg
perturbation_rf_dfr <- perturbationImportance(data = data_dfr, target_ind = 5, model = model_rf, lrn = l

perturbation_logreg_dfr <- perturbation_logreg_dfr %>% mutate(model := "logreg")
perturbation_rf_dfr<- perturbation_rf_dfr %>% mutate(model := "rf")
plot_dfr = rbind(perturbation_rf_dfr, perturbation_logreg_dfr)

   ggplot(plot_dfr) + geom_col(aes(x = names, y = values,
                                   fill = factor(model)),
                               position = "dodge")
```



W obu modelach 'partner_yrsed' odgrywa ważną rolę, natomiast zmienne binarne: 'hcm2017q24_bar_restaurant' oraz 'hcm2017q24_college', mają znacznie mniejsze znaczenie. Największą różnicę widać przy istotności zmiennej 'time_from_rel_to_cohab' która okazała się najważniejszą zmienna przy modelu Random Forest i prawie nieistotną przy regresji logistycznej.

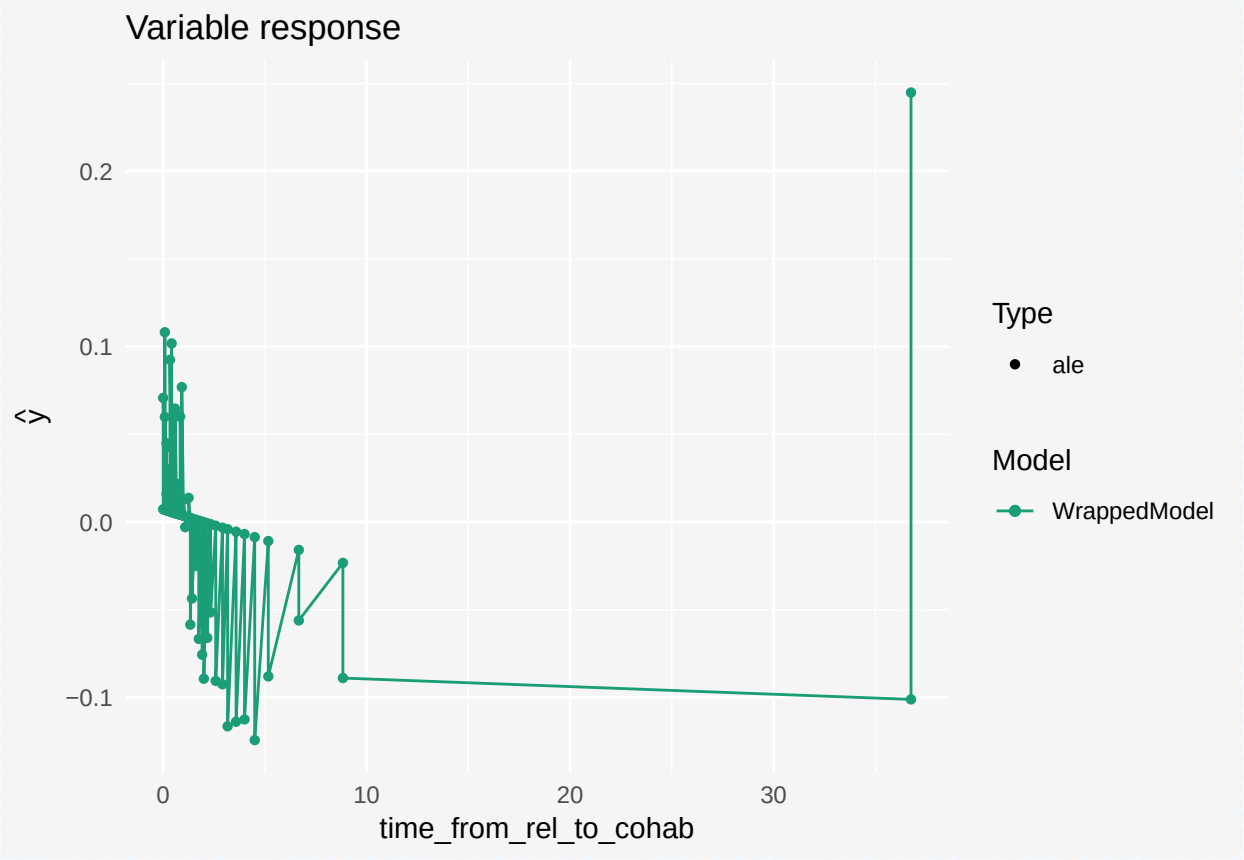
Znacznie mniejsze wartosci na wykersie dotyczącącym regresji logistycznej w porównaniu do Random Forest wynikają ze znacznie mniejszej wartości pola pod krzywą ROC dla pierwszego modelu (okolo 0.60) w stosunku do drugiego z nich (0.80).

Zadanie 4

```
library(DALEX)
```

```
## Welcome to DALEX (version: 0.2.7).
##
## Attaching package: 'DALEX'
## The following object is masked from 'package:dplyr':
##
##     explain
```

```
custom_predict_rf <- function(object, newdata) {pred <- predictLearner(lrn_rf,object,newdata)
                                                response <- pred[,2]
                                                return(response)}
explainer_rf <- explain(model_rf, data = data_dfr, predict_function = custom_predict_rf)

custom_predict_logreg <- function(object, newdata) {pred <- predictLearner(lrn_logreg,object,newdata)
                                                response <- pred[,2]
                                                return(response)}
explainer_logreg <- explain(model_logreg, data = data_dfr, predict_function = custom_predict_logreg)

expl_rf <- single_variable(explainer_rf, "time_from_rel_to_cohab", "ale")
expl_logreg <- single_variable(explainer_logreg, "time_from_rel_to_cohab", "ale")
plot(expl_logreg, expl_rf)
```



ALE plot dla random forest pokazuje wyraźnie nieliniową zależność odpowiedzi modelu względem zmiany ‘time_from_rel_to_cohab’. Znaczna część danych posiada wartość ‘time_from_rel_to_cohab’ poniżej 5, gdzie trend jest silnie spadkowy. Ta część miała największy wpływ na współczynnik modelu liniowego jakim jest regresja logistyczna. Drugi wykres przedstawia ALE plot dla regresji logistycznej. Widzimy tutaj liniowy wpływ zmiennej na odpowiedź modelu. Zmienna ta ma niski wskaźnik istotności uzyskany metodą

perturbacji, głownie dlatego, że współczynnik dla tej zmiennej powinien byc ujemny dla niskich wartości ‘time_from_rel_to_cohab’ oraz dodatni gdy zmienna ta przyjmuje wartości powyżej 10. Poprawę modelu logistycznego można uzyskać dodając nieliwą funkcję zmiennej ‘time_from_rel_to_cohab’, na przykład podnosząc ją do kwadratu.