SHAP Dataset 1

Nam Huynh

9/7/2023

```
kpres <- readRDS("~/OneDrive - UMP/RHM/2021-2022/Bài báo 30 years of water fluoridation/Results/Try7/kp

dataset1 <- cbind(kpres[["data"]],kpres[["cluster"]])
colnames(dataset1)[8] <- "cluster"


x_var_cat <- c("gender", "DT", "DMFT", "DS", "DMFS", "year", "fluoride_concentration")
y_var <- "cluster"


x <- split(dataset1, sample(rep(1:2, times=c(100,4673))))

x_test_cat <- x$`1`
x_test_cat <- x_test_cat[,-8]
x_train_cat <- x$`2`
y_train <- x_train_cat[,-1:-7]
x_train_cat <- x_train_cat[,-8]

# -- special function when using categorical data + xgboost
dummylist <- make_dummies(traindata = x_train_cat, testdata = x_test_cat)

x_train_dummy <- dummylist$train_dummies
x_test_dummy <- dummylist$test_dummies

# Fitting a basic xgboost model to the training data
model_cat <- xgboost::xgboost(
  data = x_train_dummy,
  label = y_train,
  nround = 20,
  verbose = FALSE
)
model_cat$feature_list <- dummylist$feature_list

explainer_cat <- shapr(dummylist$traindata_new, model_cat)

p <- mean(y_train)

explanation_cat <- explain(
  dummylist$testdata_new,
  approach = "ctree",
  explainer = explainer_cat,
  prediction_zero = p
```

```
)

# Plot the resulting explanations for observations 1 and 6, excluding
# the no-covariate effect
plot(explanation_cat, plot_phi0 = FALSE, index_x_test = c(1, 20, 40, 60, 80, 100))
```