EDM

2023-04-26

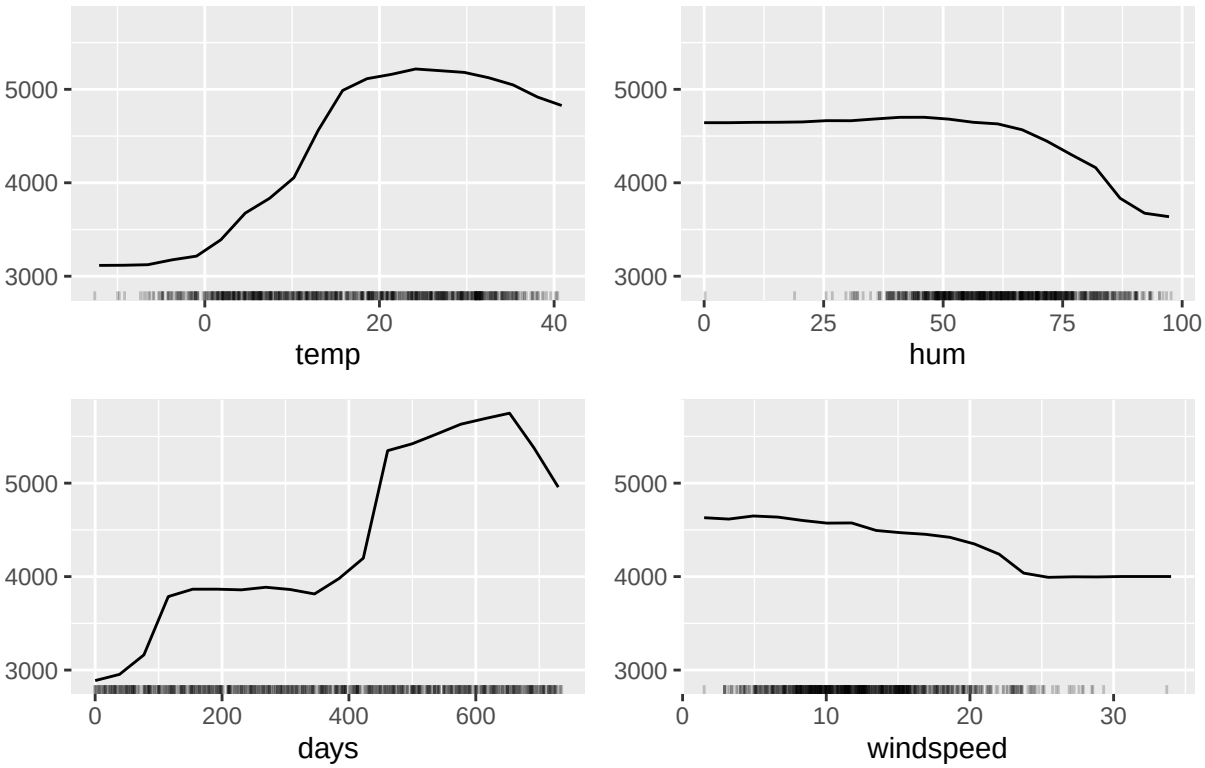
1.- One dimensional Partial Dependence Plot.

```
bici = import("day.csv")

bici$MISTY = ifelse(bici$weathersit == 2, 1, 0)
bici$RAIN = ifelse(bici$weathersit %in% c(3,4), 1, 0)

bici$windspeed2 = bici$windspeed * 67
bici$hum2 = bici$hum * 100
bici$temp2 <- (50 - (-16)) * bici$temp + (-16)

bici$dteday = as.Date(bici$dteday)
reference_date <- as.Date("2011-01-01")
bici$days_since_2011 <- as.numeric(bici$dteday - reference_date)

pred_df = data.frame(workingday = bici$workingday, holiday = bici$holiday, misty = bici$MISTY,
                     rain = bici$RAIN, windspeed = bici$windspeed2, temp = bici$temp2,
                     hum = bici$hum2, days = bici$days_since_2011, cnt = bici$cnt)

bici$season = as.factor(bici$season)

seasons = model.matrix(~ bici$season , data = bici)[,-1]
colnames(seasons) <- c('season2','season3','season4')
pred_matrix = cbind(pred_df, seasons)
```

```
set.seed(1234)
rf <- randomForest(formula = cnt ~ .,
                   data = pred_matrix,
                   ntree = 100)
rf
```

```
##
## Call:
##  randomForest(formula = cnt ~ ., data = pred_matrix, ntree = 100)
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 462713.1
##                     % Var explained: 87.65
```

```
predictor <- Predictor$new(rf,data=pred_matrix,y=pred_matrix$cnt)
effs <- FeatureEffects$new(predictor,feature = c("temp","hum","days","windspeed"),method = "pdp")
plot(effs)
```



On this PDP plots we can see various things about the correlation between the predictor features and the amount of bikes that are predicted to get withdrawn.

Firstly, the temperature is has an interesting behavior, as lower values of the prediction are expected when the temperature is lower, specially below 19 degrees more or less. It appears that the optimal temperature to go on a bike ride is between 20 and 30 degrees and if the temperature continues to rise, a smaller number of bikes is expected to get rented.

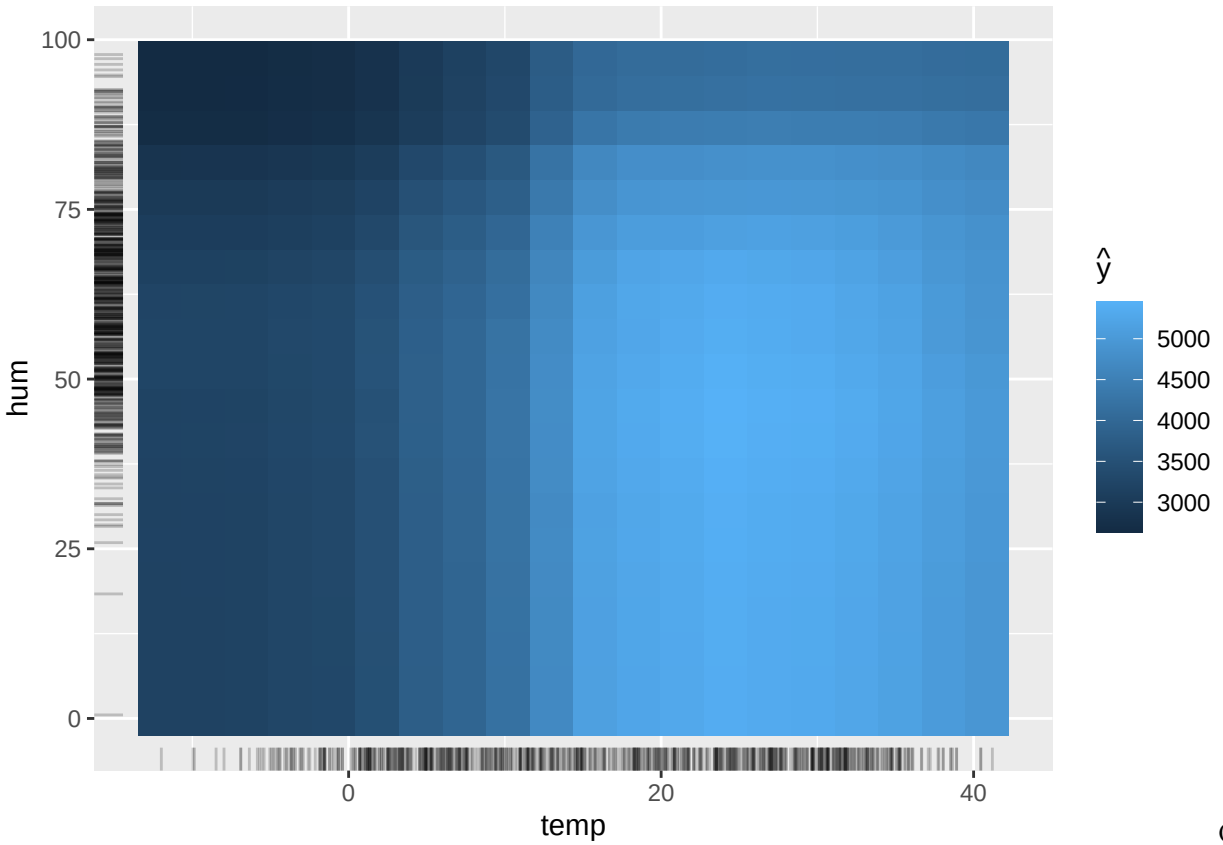
Humidity is different from temperature as it pretty much does not matter to the prediction (because of being constant) as long as it is lower than 75 %, where higher values of humidity lower the prediction.

About the days since 2011, mostly the trend is increasing the values of the prediction, in particular there were two main big increases, one around the 100th day and the second one on the 400th day (more or less). However, despite having a positive trend all throughout time, on the last 100 days or so the trend was decreasing and the amount of bikes rented was less the on the 600th day.

Finally, wind speed's behavior was much like the humidity, as it does not affect to the prediction up to higher values of it.

2.- Bidimensional Partial Dependency Plot.

```
pred1 <- Predictor$new(rf,data=pred_matrix)
effs1 <- FeatureEffect$new(pred1,feature = c("temp","hum"),method = "pdp")
plot(effs1)
```



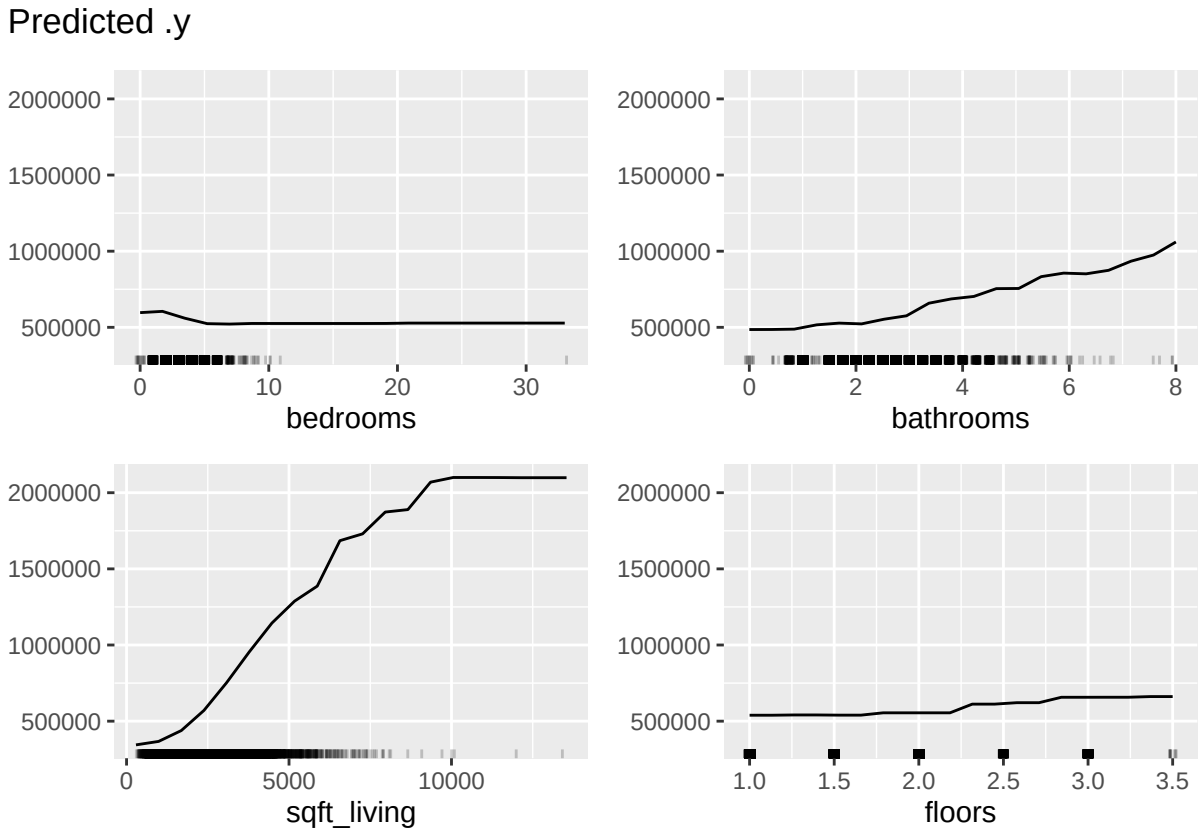
On this graph we can obtain similar conclusions, although we now can see the correlation between two variables and the prediction of cnt. (bikes rented) As we mentioned before, the optimal point for temperature is between 20 and 30 degrees and the humidity does not affect the prediction as long as it isn't high (or at least higher than around 75%). We can see that on the plot above, where the lighter areas are a few columns on the optimal temperature and below some high humidity level.

3.- PDP to explain the price of a house.

```
house = read.csv("kc_house_data.csv")

house <- house %>% select("price","bedrooms","bathrooms","sqft_living","sqft_lot","floors","yr_built")

rf2 <- randomForest(formula = price ~ .,
                    data = house,
                    ntree = 100)
rf2
```

```
##
## Call:
##  randomForest(formula = price ~ ., data = house, ntree = 100)
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 51810303277
##                     % Var explained: 61.56
```

```
pred3 <- Predictor$new(rf2,data=house,y=house$price)
effs2 <- FeatureEffects$new(pred3,feature = c("bedrooms","bathrooms","sqft_living","floors"),method = "
plot(effs2)
```



As we can see in the plots, the number of bedrooms and floors does not matter much, in the bedrooms feature it seems lesser bedrooms makes the price go up, although this could be happening because of the amount of data and the variability of house pricing with the same number of bedrooms. Also, more floors implies higher price, although the correlation between floors and pricing is not as notable as the correlation with the number of bathrooms and the square feet of living. On those variables, it is easy to see that higher values means higher pricing.