# Model-agnostic: Partial Dependency Plot (PDP)

Yusef Ahsini Ouariaghli, Eva Cantín Larumbe and Pablo Díaz-Masa Valencia

2023-04-26

## 1. One dimensional Partial Dependence Plot.

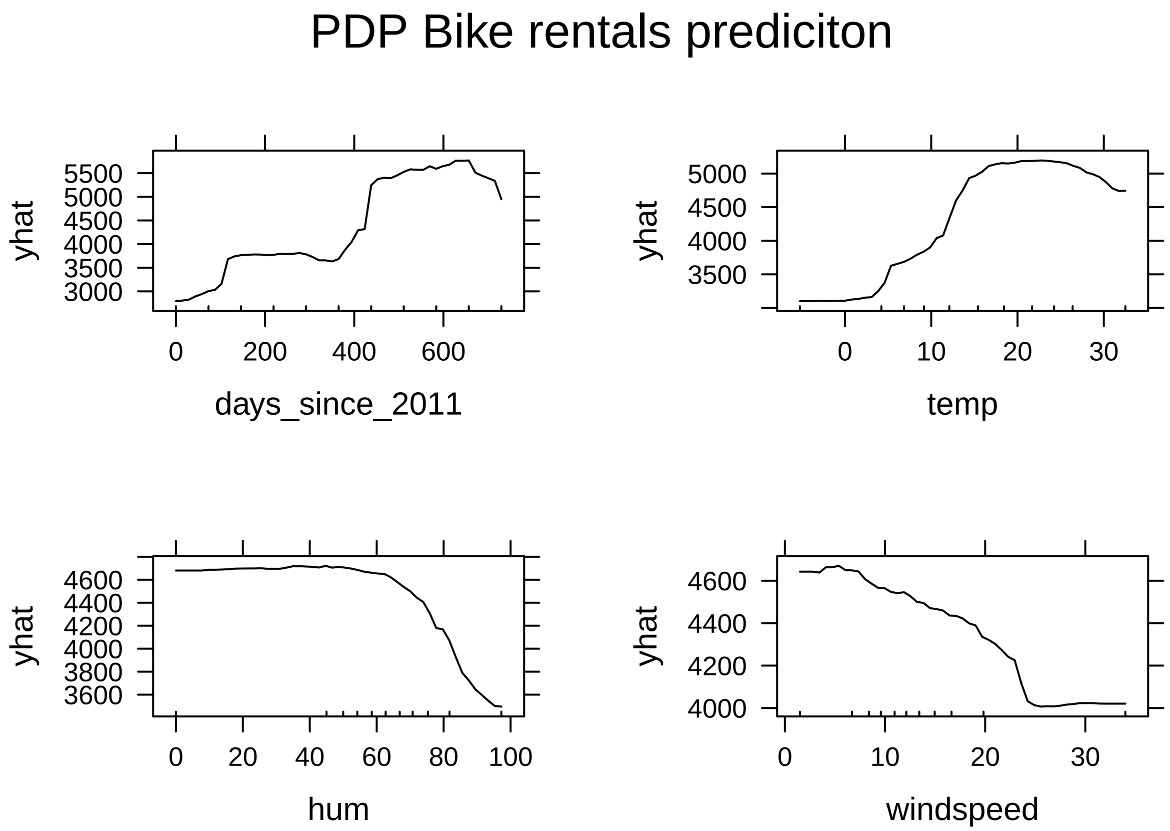
The partial dependence plot shows the marginal effect of a feature on the predicted outcome of a previously fit model.

### Exercise

**Apply PDP to the regression example of predicting bike rentals. Fit a random forest approximation for the prediction of bike rentals (cnt). Use the partial dependence plot to visualize the relationships the model learned. Use the slides shown in class as model.**

First and foremost, we use the perform a Random Forest to predict the number of bike rentals (cnt).

Then, we do a Partial Dependence Plot to explain the results from the Random Forest, because it can show if the relationship between the target and a feature is linear, monotonic or more complex.



The rugs of the X-axis indicate the data distribution.

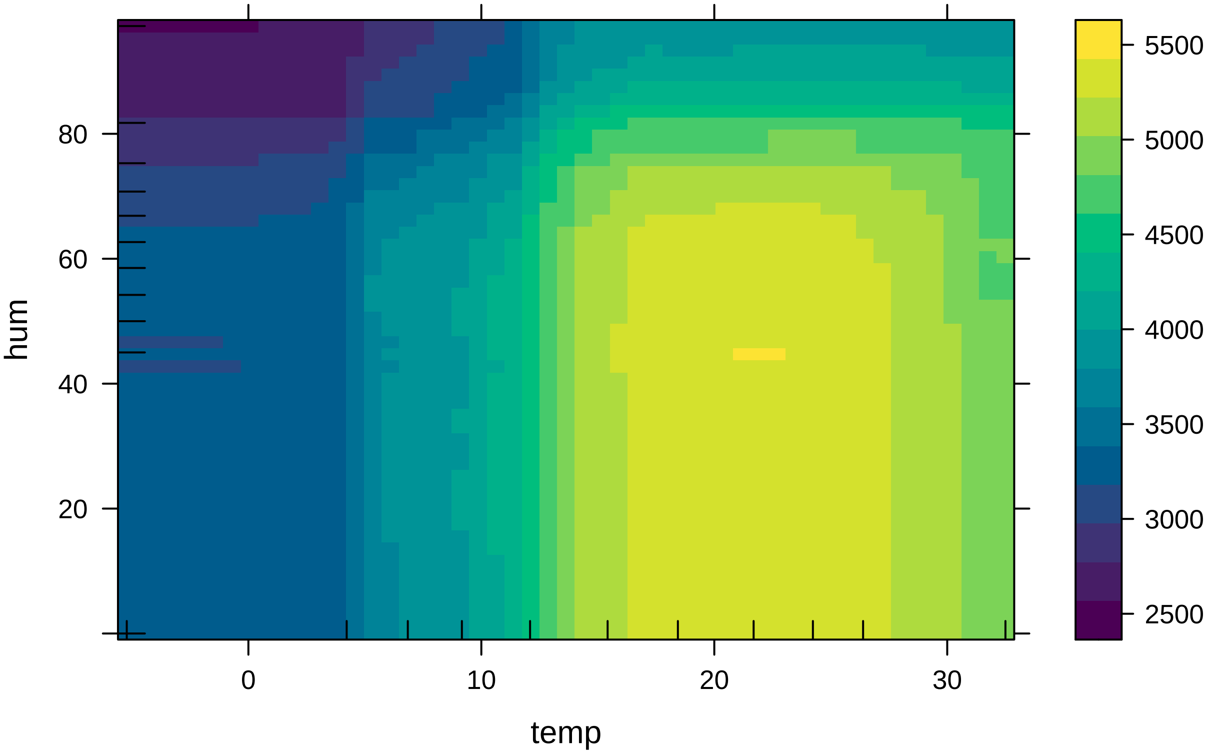
- PDP days_since_2011 and cnt: It is shown that in the first 100 days, bike rentals barely increase. However, after these 100 days, there is a tipping point that rises to approximately 3700. Then, there is a flat tendency until 400 days, when it soars more than 1000 bikes in a few days. Then, we can observe again a tiny increase in the number of bikes rented. Finally, it drops slightly at day 650. On average, the more days elapsed, the more bikes are predicted to be rented.
- PDP temp and cnt: the hotter days, more bikes are predicted to be rented. However, more bikes could be rented on average when the temperature is neither too cold nor too hot.
- PDP hum and cnt: People are predicted to rent fewer bikes when the humidity is above 60%. The more humidity there is, the fewer bikes are rented, so it makes sense.
- PDP windspeed and cnt: When the windspeed exceeds 24 km/h, suddenly the number of bikes rented is plummets. An interesting aspect is that when windspeed is between 25km/h and 30km/h, there is no difference in the number of bikes rented. It could be that there is not much training data, and the Random Forest is not predicting well extreme (anomalous) values.

Finally, we can see that the most significant differences can be seen in days_since_2011 and temperature. Therefore, it can be inferred that a change in these two values contributes more to predicting the number of bikes.

# 2. Bidimensional Partial Dependency Plot.

## Exercise

**Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending on those parameters.**



The generated graphic relates humidity and temperature to the predicted amount of bikes rented.

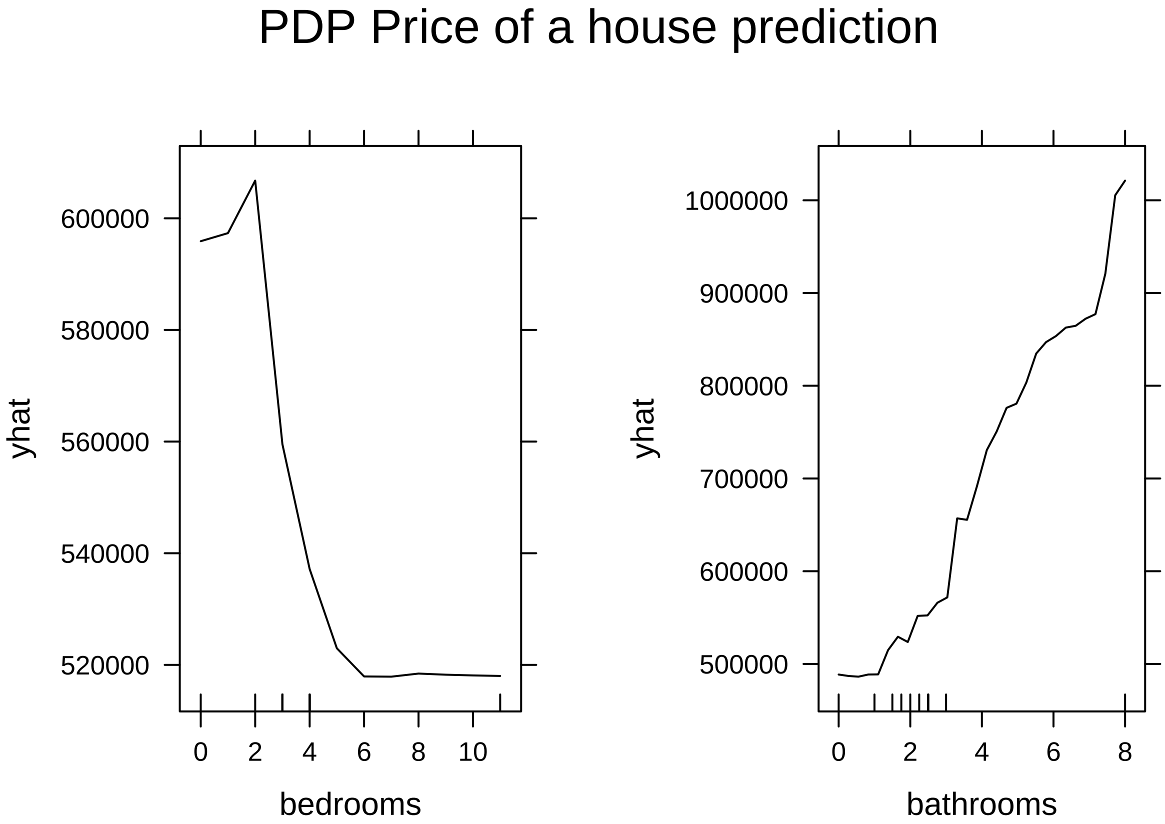
The first thing that catches our attention is the large yellowish zone on the right side of the figure, indicating the ideal temperature range where people rent bikes. However, when we look at the area above this yellow zone, we realize a sharp drop in the number of rented bikes when humidity exceeds approximately 70%. This vertical parallelism is maintained throughout the entire graph. That is, humidity only causes users to stop renting bikes once it exceeds the threshold of 70%. Regarding the other variable, low temperatures persuade people to rent bikes. This effect is also observed at high temperatures, although to a much lesser extent.

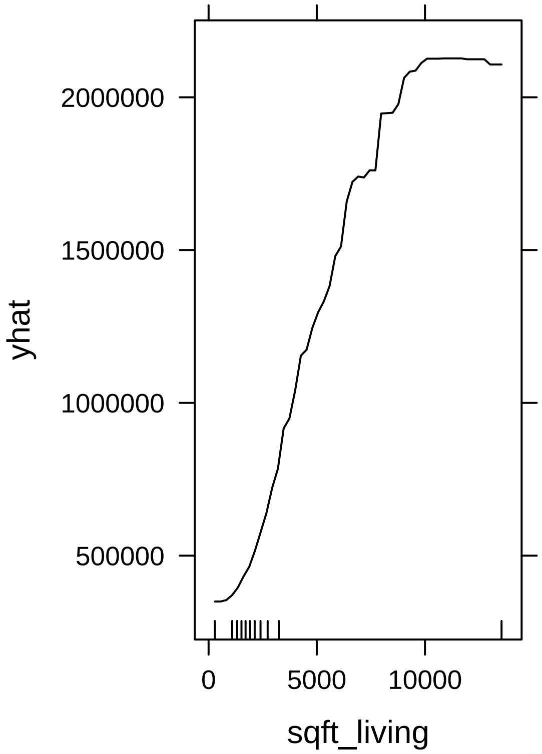
Naturally, when the temperature is very low, and humidity is very high, the number of rented bikes reaches its minimum. However, the rugs in both variables indicate few cases in the database. The approximate ranges of temperature and humidity that maximize the number of rented bikes are 20-23 ºC and 40-50 %, respectively.

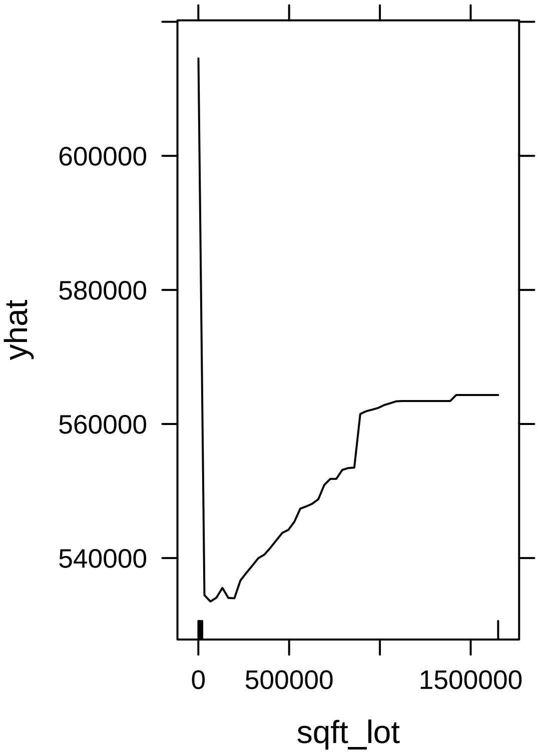
## 3.- PDP to explain the price of a house.

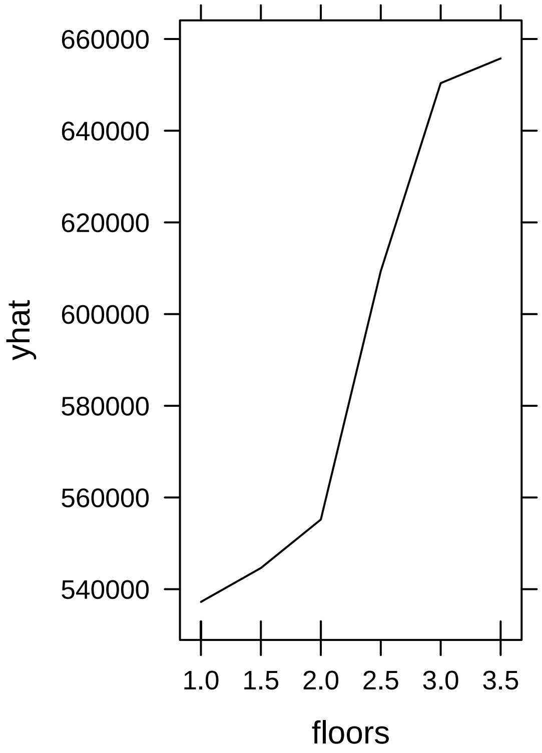
### EXERCISE

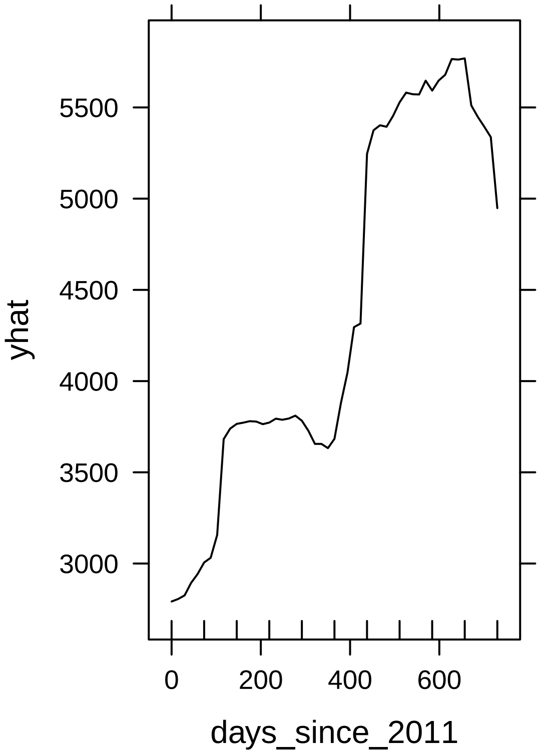
**Apply the previous concepts to predict the price of a house from the database kc_house_data.csv. In this case, use again a random forest approximation for the prediction based on the features bedrooms, bathrooms, sqft_living, sqft_lot, floors and yr_built. Use the partial dependence plot to visualize the relationships the model learned.**




yhat
2000000
1500000
1000000
500000
0
5000
10000
sqft_living


yhat
600000
580000
560000
540000
0
500000
1500000
sqft_lot





In the bathrooms chart, we can see a practically linear relationship between the number of bathrooms and the house price, as it grows more or less constantly as the number of bathrooms increases.

In the PDP chart of sqft_living, something similar happens, as the price increases depending on the number of living sqft. However, this relationship no longer fits as adequately into a straight line as the previous one. However, upon reaching a certain upper threshold of sqft, the price remains stable despite the increase in sqft. In any case, the behavior on the right side of the chart cannot be considered representative as that on the left side since the rugs indicate very few occurrences when living_sqft is very high.

With the sqft_lot variable, something similar happens, although this effect of little representation on the right side is even more accentuated.

In the PDP chart of floors, the price gradually increases as the number of floors increases. The most significant price jump occurs when moving from two- to three-floor houses.

Finally, in the PDP chart of yr_built, the rugs indicate that occurrences are relatively evenly distributed over the years (although it is noticeable that there are more new homes than old ones). The price reaches its minimum around the year 2000. Because the oldest houses are larger and more stately or located near the center, they reach the highest prices. At the same time, newly built homes cost more than those built between the 1980s and 2000s.