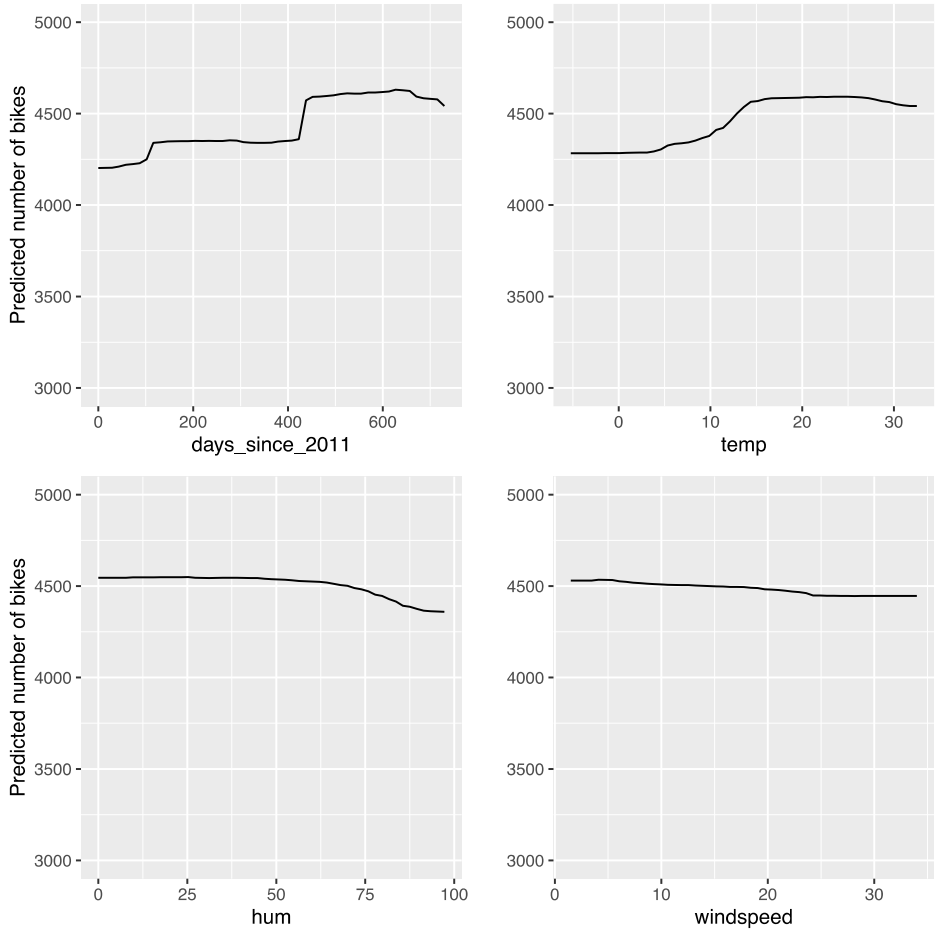
Deployment: Model-agnostic methods

Aina Magraner Rincón and Alba Martínez López

Exercise 5.- Model-agnostic: Partial Dependency Plot (PDP).

1.- One dimensional Partial Dependence Plot.

Ex1:: 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.



Question 1: Analyse the influence of days since 2011, temperature, humidity and wind speed on the predicted bike counts.

The influence of the variable `days_since_2011` has 3 different levels. Firstly, when the variable is lower than 100, the number of bikes is around 4250, once the variable is in

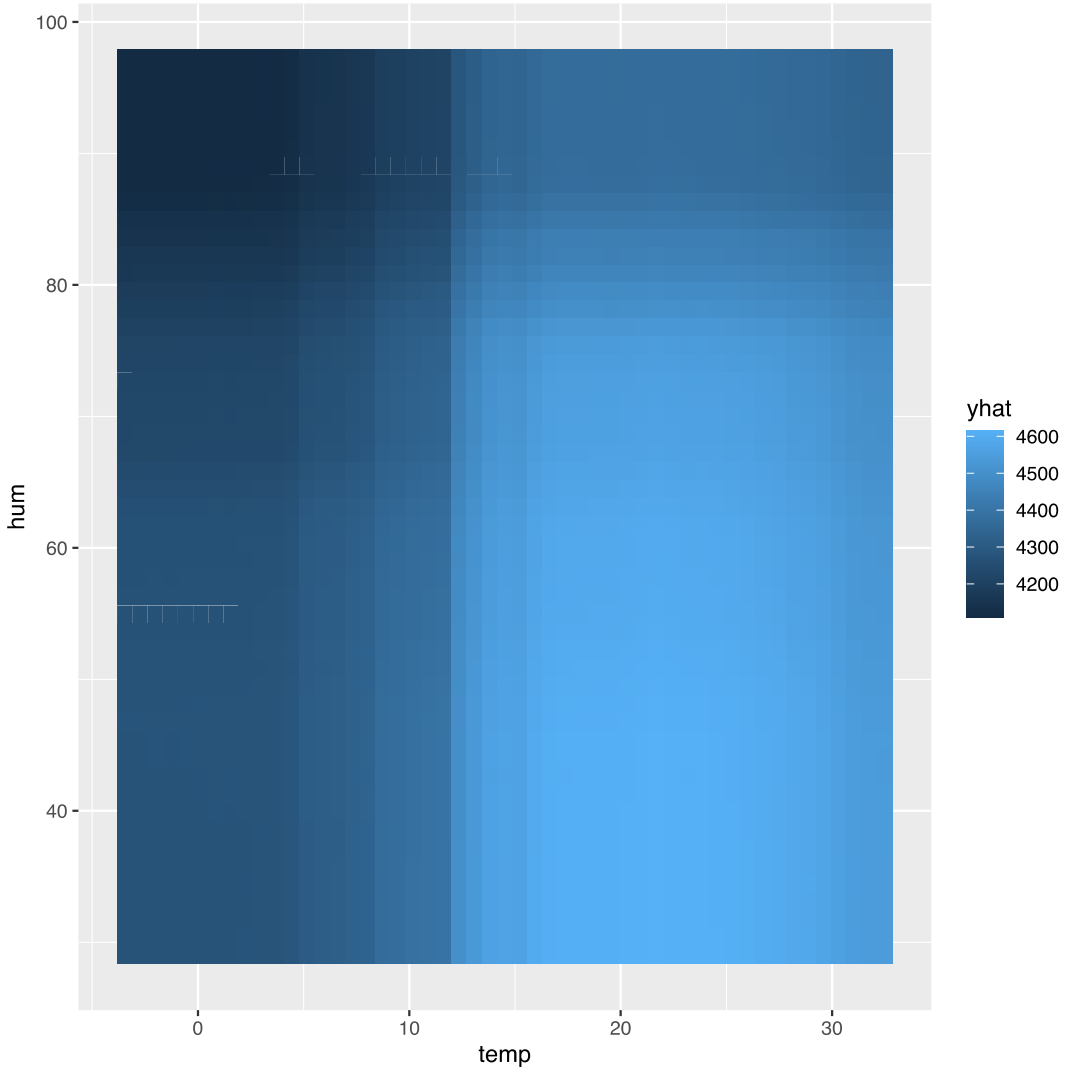
between a 100 and 425, the number of bikes is around 4.350. Lastly when the variable is higher than 425, the number of rented bikes increases up to 4750.

The influence of temperature is positives, since the variable to be predicted increases when the temperature does, mostly from 5 to 15, in this interval is when the influence is stronger.

As of the variables humidity and windspeed, we can see how their influence is much lower and has a negative effect.

2.- Bidimensionl Partial Dependency Plot

Ex2: Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending on those parameters.

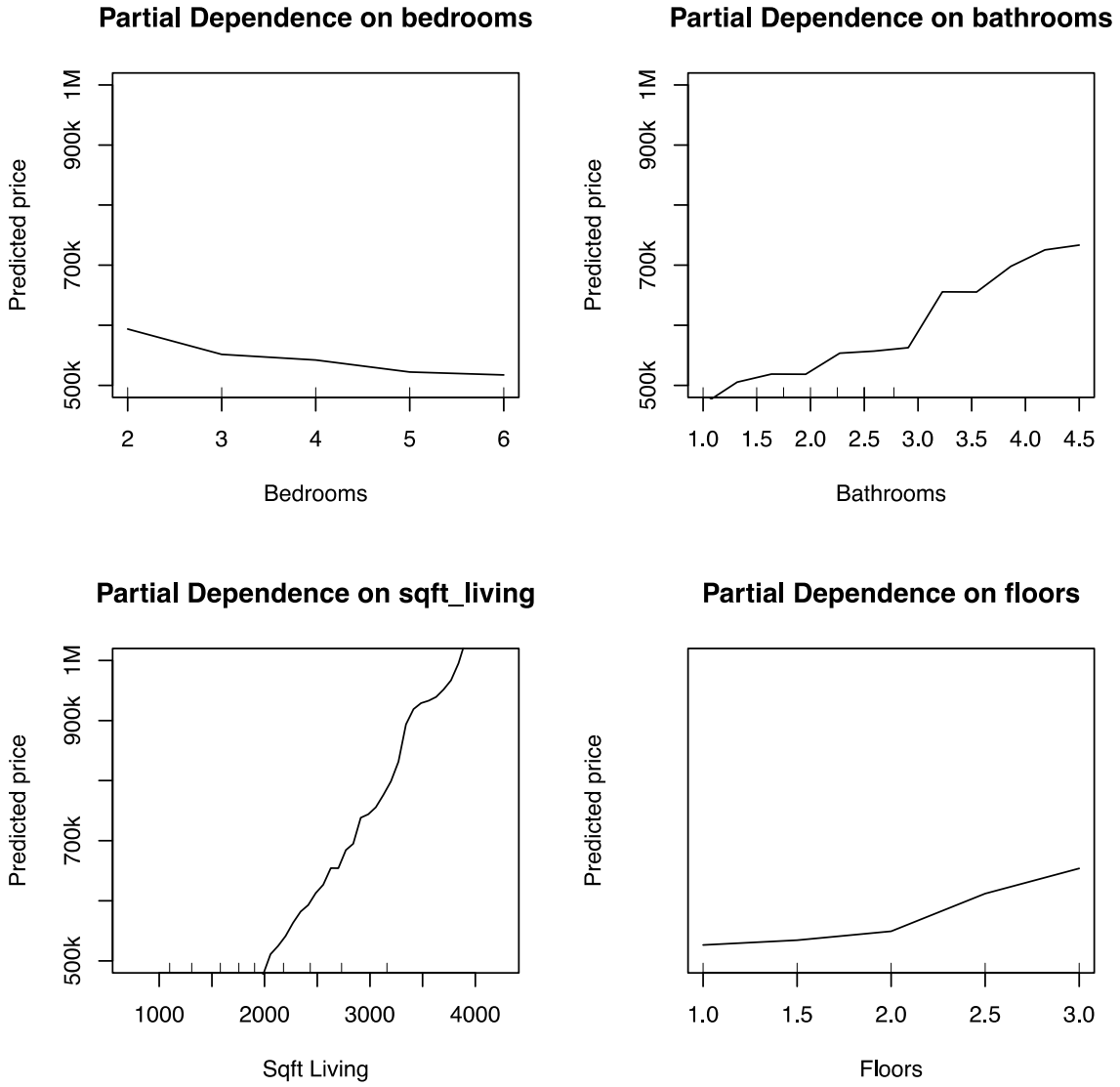


Question 2: Interpret the results.

The following graph comes to show how the effect that the humidity has over the number of rented bikes is not comparable to the one that the temperature has, meaning that we clearly see a significant change in the number of bikes when changing temperature, this change most significant when the temperature goes from 13 to 15. Whereas there is no significant increase or decrease in the rented bikes has the humidity is higher or lower.

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

Ex3: 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.



Question 3: Analyse the influence of bedrooms, bathrooms, sqft_living and floors on the predicted price.

In order to obtain the plots, we decide to make them in the same scale, in order to do a better analysis. As we can see, the characteristic of the house with major influence is the sqft_living. We can say that more sqft_living indicates major prices and viceversa. However, the number of bathrooms in the house is also an important characteristic to take into account when talking about prices. A house with 3.5 to 4.5 bathrooms is significantly more expensive than a house with 1 to 3 bathrooms. In the other hand, it seems that the number of bedrooms or the number of floors do not increase or decrease the price as much as the other characteristics. If we have 2 to 4 bedrooms, the price is slighlty higher than if we have 4 to 6 bedrooms, while if we talk about floors, having 1 to 2 is cheaper than 2 to 3.

To sum up, the number of bedrooms and floors are the characteristics with less influence in the price of the house (the range of variation is from 500k to 600k for the first one and 500k to 680k for the second), while the sqft_living and number of bathrooms are the ones which contribute more to the final price (the first ranges between 500k and 1M and the second from 500k to 750k).