UNIVERSIDAD POLITÉCNICA DE VALENCIA

PRACTICAL EXERCISE 5: DEPLOYMENT

Evaluation, deployment and monitoring of models
Data Science Degree

AUTHORS:
Raquel Chaves Martinez
Constantino Martínez de la Rosa
Ana Rubio García

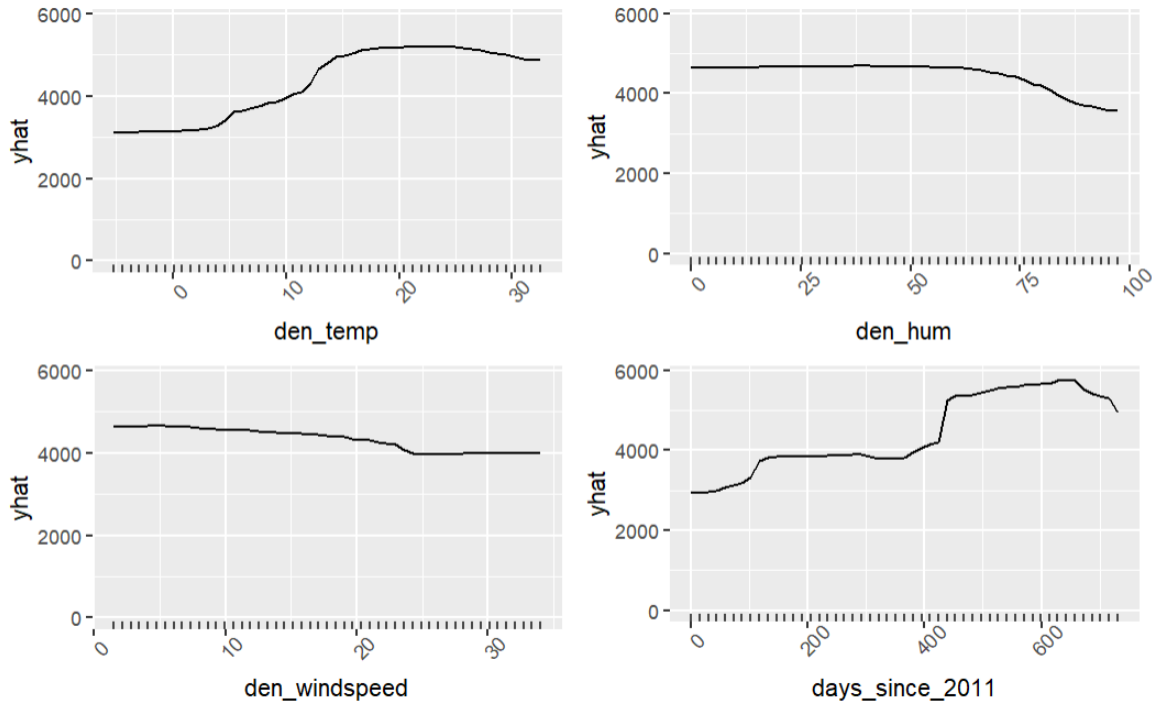
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Content table

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. In our case, we build a Random Forest model to predict the number of bikes rented using the features workingday, holiday, spring, summer, fall, misty, rain, den_temp, den_hum, den_windspeed and days since 2011.

After building the model, we can use the partial() function to learn and then visualize the relationships the model learned. Specifically, we are going to study the influence of the number of days since 2011, the temperature (den_temp), the humidity (den_hum) and the wind speed. To do so, we build the following graph:



Thanks to this figure we can compare the influence of each of this four variables on the predicted bike counts.

For instance, if we look at the top left graph we can see that the number of bikes rented increases with the temperature until the temperature is around 15. Then it remains constant until the temperature is above 30 degrees and it starts decreasing.

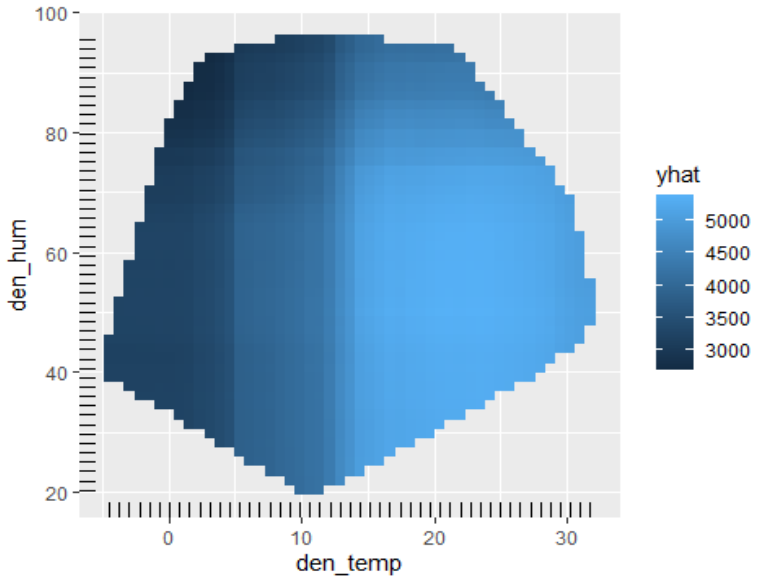
If we look at the top right, we can see that the humidity has almost no impact on the bike count as the number of bikes is constant except when the humidity is over 75, that makes it decrease.

The wind speed follows a similar pattern but, in this case, the number of bikes rented decreases gradually until the wind speed is above 20 when it remains constant.

Finally, we can study the influence of the number of days since 2011. In the correspondent graph we can see that the number of bikes rented increases with the number of days.

2.- Bidimensional Partial Dependency Plot

For this section, we want to analyze the influence in the prediction of the number of bikes rented of both humidity and temperature. To do so, we have built a partial dependency plot that includes these two features.



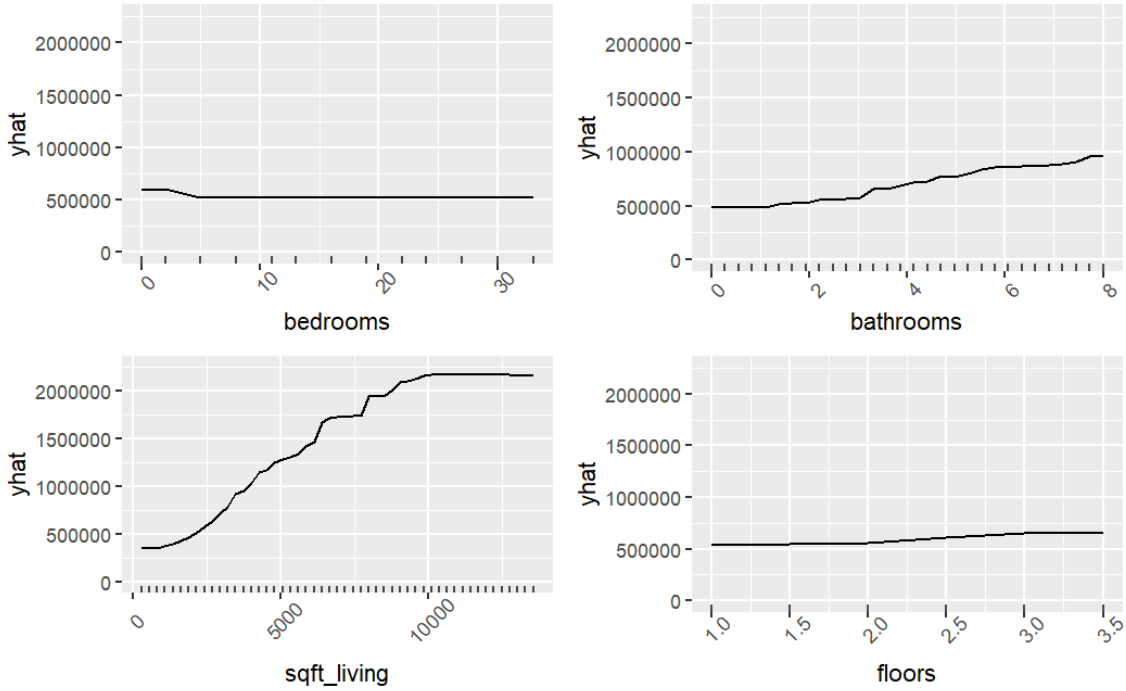
To interpret this plot it is important to take into account the plots generated in the previous section. This means that at first, we could get an idea of what the plot would look like in case there was no interaction between the variables. That is, in the previous exercise we have observed in the plot of the variable "den_temp" that the value of the variable explained (cnt, which is the number of bicycles rented) increases as the value of the variable "den_temp" increases. It is also observed in the graph of the variable "den_hum" that the value of the explained variable decreases as the value of the variable "den_hum" increases.

In short, what we could expect in advance from the 2D partial dependency plot, in case there was no interaction between the variables "den_temp" and "den_hum", is that the highest values of the explained variable are in the area where the values of the variable that measures humidity are low and the values of the variable that measures temperature are high. We would also expect the lowest values of the explained variable to be on the opposite side of the plot (where humidity is high and temperature is low).

As we observe that, indeed, the plot has the shape we expected, we can conclude that if there is some kind of interaction between the variables "den_temp" and "den_hum", we are unable to detect it through this dependency plot.

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

For this last section we are working with a different dataset. In this case we use a random forest approximation for the prediction of the price of a house based on features bedrooms, bathrooms, sqft living, sqft lot, floors and year built. To visualize the relationships the model has learned, we have built the following plot



With these four partial depency plots we can analyze some variables such as the number of rooms, the number of bathrooms, the number of habitable square meters or the number of floors affect the price of a home.

Of the four graphs, the first that catches our attention is that of the variable "sqft_living" since it is in which the greatest change occurs throughout the plot. It can be perfectly appreciated that the value of a home increases significantly the more habitable square meters it has, even homes with more square meters quintuple the value of those with fewer meters.

In the case of the variables "floors" and "bathrooms", we can observe how there is also an increase in the price of housing as the value of these variables increases, although this is not as large as in the case of the variable "sqft_living".

Finally, surprisingly the plot of the variable "bedrooms" seems to show us something totally unthinkable at first, since it is observed that the value of homes does not increase as the number of rooms in these increases. Moreover, it seems that homes with few rooms are even a little more expensive than the rest.

In short, the conclusions we can draw is that the variable that most increases the price of housing is the number of square meters habitable.