DEPLOYMENT

Marc Sánchez Gil, Samuel Lozano Gómez, Dylan Lanza Méndez

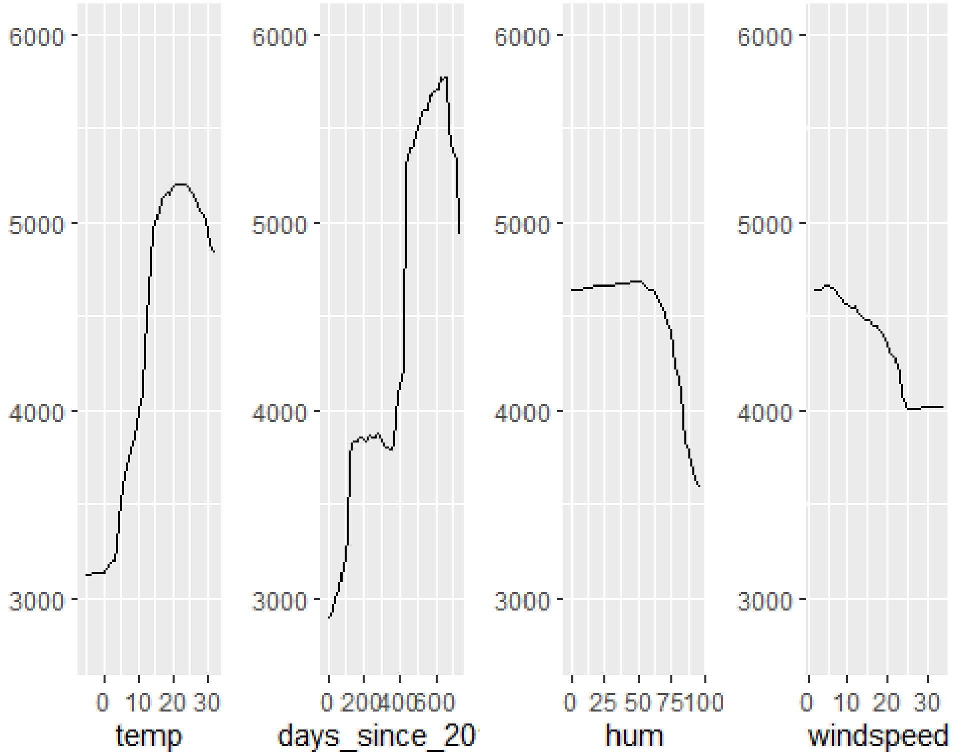
1.- One dimensional Partial Dependence Plot

In the first section, our aim is to predict the number of rented bikes in a day using variables such as temperature, humidity, rainy day … etc.

We have implemented a random forest approximation for the prediction.

Our next objective is to observe the relationships the model has learned, so we are going to generate several Partial Dependence Plots.

These plots show how a variable or several variables influence in the prediction of the model.



Once we have generated the plots for the variables temperature, days_since_2011, humidity and windspeed, we can observe that each of these variables have a great influence in the predicted value of rented bikes.

On the one hand, we can see that while the temperature raises, the rented bikes also do, but not for the entire range of temperatures. When the temperature is higher than 21 or 22 degrees, the number of rented bikes starts to decrease. This means that, when it is hot, people do not like going by bike.

On the other hand, we observe that in days_since_2011 there is not a clear relationship with the rented bikes, but what we observe is that there is a huge difference in different values, probably it is because the different season, but we cannot assume that, it is only an idea.

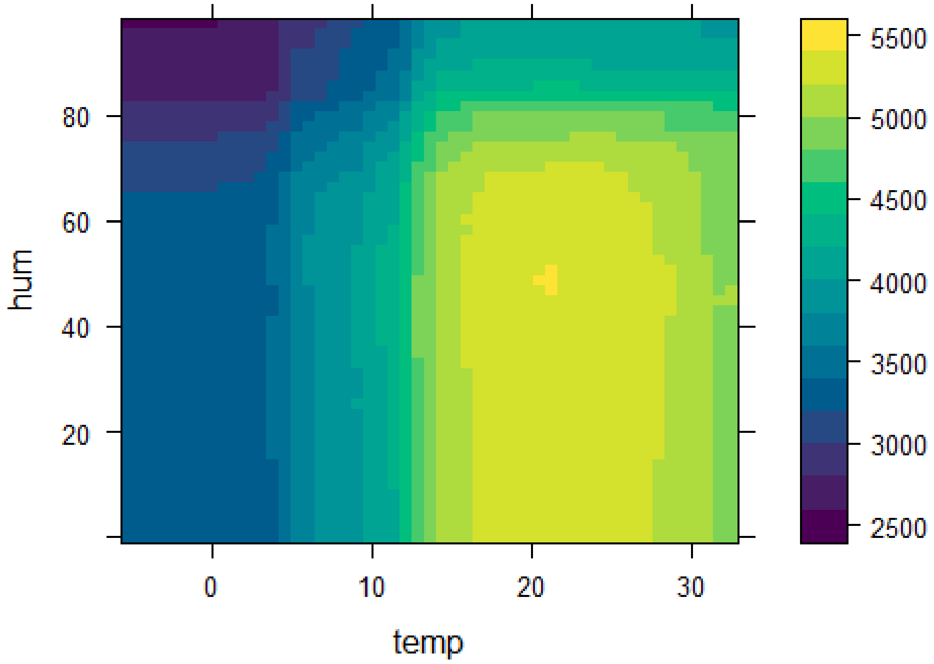
Regarding to the humidity, we observe that it does not have any influence until when it is higher than 60, when the number of rented bikes starts to decrease. This means that a high relative humidity reduces the number of rented bikes and people do not like riding bikes when the humidity is high.

In the last plot, we can observe that while the windspeed increases, the number of rented bikes decrease. We can assume that it is because it is difficult and uncomfortable to ride a bike when the wind is strong, so it is normal that the number of bikes decreases.

In conclusion, all of the four variables that we have analysed have a great influence in the prediction of bikes rented.

2.- Bidimensional Partial Dependency Plot

In this section, in order to predict the number of bikes rented depending on humidity and temperature parameters, we have generated a 2D Partial Dependency Plot with these two and the result was as follows:



We can see that the plot we have obtained is similar to a heatmap, where the x-axis corresponds to the temperature, the y-axis corresponds to the humidity and the color gradient corresponds to the number of bikes rented. In order to interpret the plot, we must take into account what said before and also the color bar. That said, let's first focus on the marginal effect of each feature on the number of bikes rented.

Starting with the humidity, we can notice that, as it increases, the predicted number of bikes rented remains constant, until humidity reaches a value at which the predicted variable starts to decrease continuously.

Continuing with the temperature, we can notice that, as the temperature increases, the predicted number of bikes rented also increases, until the temperature reaches a point where the predicted variable stops increasing and remains constant, until the temperature reaches another point from which it starts to decrease.

Having said this, we realize that what we have just discussed coincides with what we have seen in temperature's PD plot and humidity's PD plot of the previous exercise.

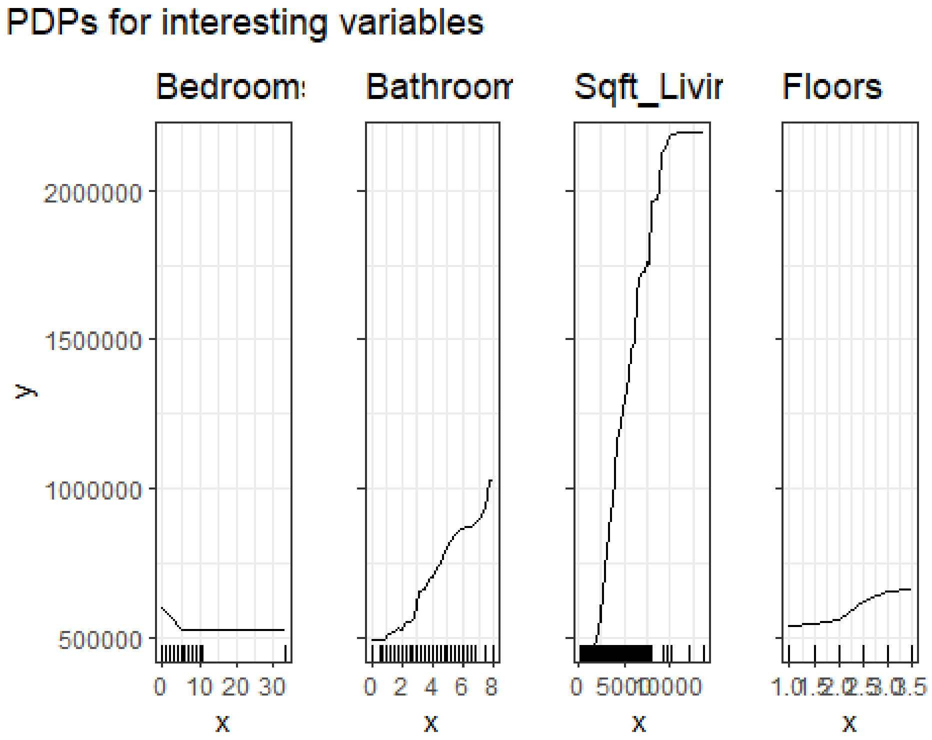
Having seen the relationship between each feature separately and the number of bikes rented, we can say that, according to the model, in general the number of bikes rented will be higher the higher the temperature and the lower the humidity. However, this relationship is causal for the model but not necessarily for the real world.

3.- PDP to explain the price of a house

In this section, we try to predict the price of a house and analyze some interesting variables using Partial Dependence Plots (PDPs). For this, we use the dataset of *kc_house_data.csv* which contains historical data of houses sold between May 2014 and May 2015 in Washington, USA.

For this purpose, a Random Forest model has been developed to predict the price of a house as a function of bedrooms, bathrooms, lot area, house area, number of floors, and year of construction.

For a better representation, about one-third of the observations have been randomly removed. Through the following figure with four PDPs, we can study the influence of bedrooms, bathrooms, house area, and the number of floors.



Looking at the figure we see several things. The price of a house is not very conditioned by the number of bedrooms, we could even say that as they increase the price is maintained or decreases, this may be because there are other types of rooms in a house that are more expensive, having more bedrooms means fewer rooms of other types, so to make a house more expensive it will be important not to have many. This can also be because it is normal for a house to be inhabited by few people, one, two, or three, it is usually the norm, adding more bedrooms than people means that there are bedrooms without much utility so they do not make the house more expensive. Curiously, a value of 33 bedrooms is observed, which could be studied in depth if it is an incorrect value, or see the characteristics of that house, as it could be a shelter for homeless people. If this were the case, it would be normal that the price would not be affected since the characteristics of this “house” would probably not be of great luxury and would not have many more rooms that could make the house more expensive.

Regarding the bathrooms, there is a linear relationship as the number of bathrooms increases, adjusting a straight line we would see approximately a relationship of an increase of 62,500€ extra for each bathroom. This is an important room in the house, most are between 0 and 3, an increase in this value also means an increase in many other attributes which generate a significant price increase.

The area of the house (sqft_living) is the feature that has more influence, on those we are studying in the price of housing. The size of a house has to do with more comfort, more space, and more rooms… so this positive relationship is understandable. The usual size of the dataset is between 0 and 6000 square meters, where most of the observations are concentrated. Undoubtedly it is a very important factor, the price per meter is expensive and each increase directly affects the price of housing. There comes a point where it remains unchanged up to 10,000 square meters, probably because in that case it is about luxurious houses where other factors come into play rather than the size of the house itself, but for inhabitants of a normal purchasing power, it is a key factor in the price.

Finally, the number of floors. It is normal that as the number of floors increases so does the price, there is not a big increase, probably because it is not something so decisive, it would be necessary to see what is the surface of each floor as there are houses with only one floor that can cover more area than one with three floors, and as we have seen before the big key that determines the price is the size of the house. This fact, having one or more floors sometimes depends on the location or the type of building where the houses are located, also it is not always preferable to have more floors for a house, sometimes with less is more, because many floors can become uncomfortable.