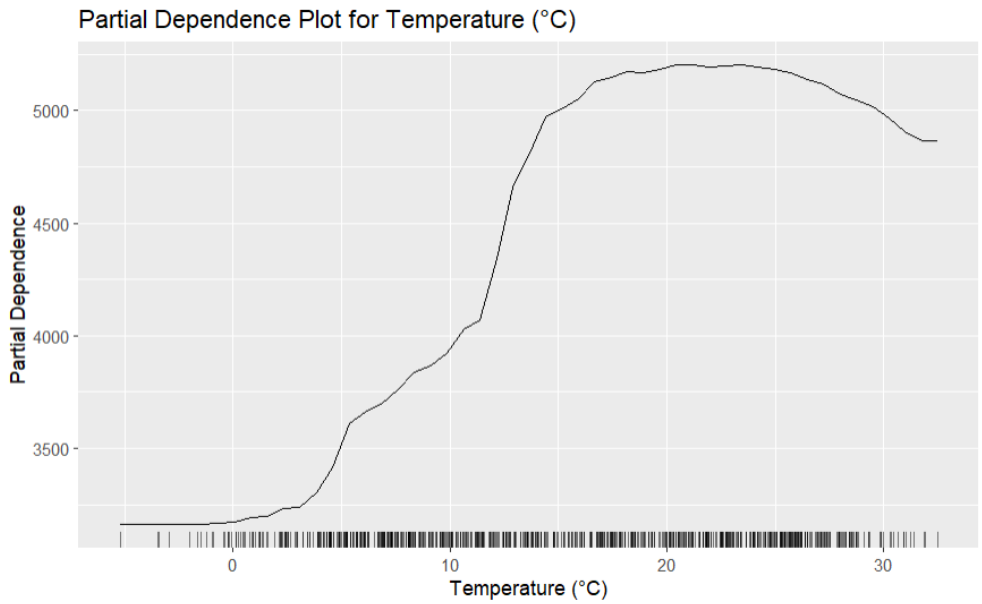
XAI3: Model-agnostic methods

In this report, we will create a random forest model to predict the number of bicycles rented. We will use the bike rental dataset from our previous work. The purpose of this report is to interpret black-box models, such as random forests, using model-agnostic methods. Specifically, we will perform Partial Dependence Plots (PDP) on the variables in the dataset. We will also apply the same methodology to a new dataset related to housing prices, which includes variables such as the number of bathrooms, bedrooms, and more.

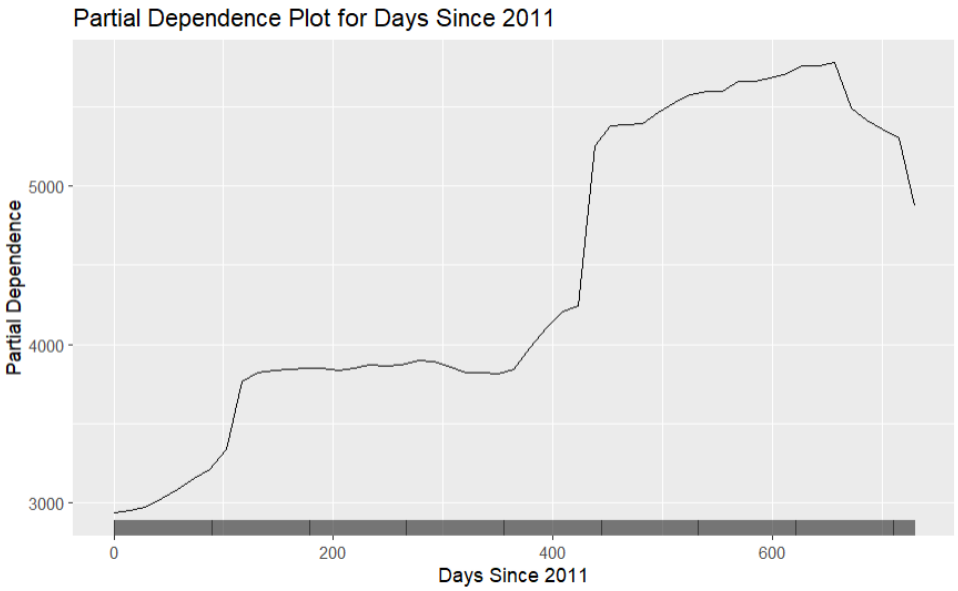
One dimensional Partial Dependence Plot

Now we will examine the importance of each variable for the constructed model using PDP plots:

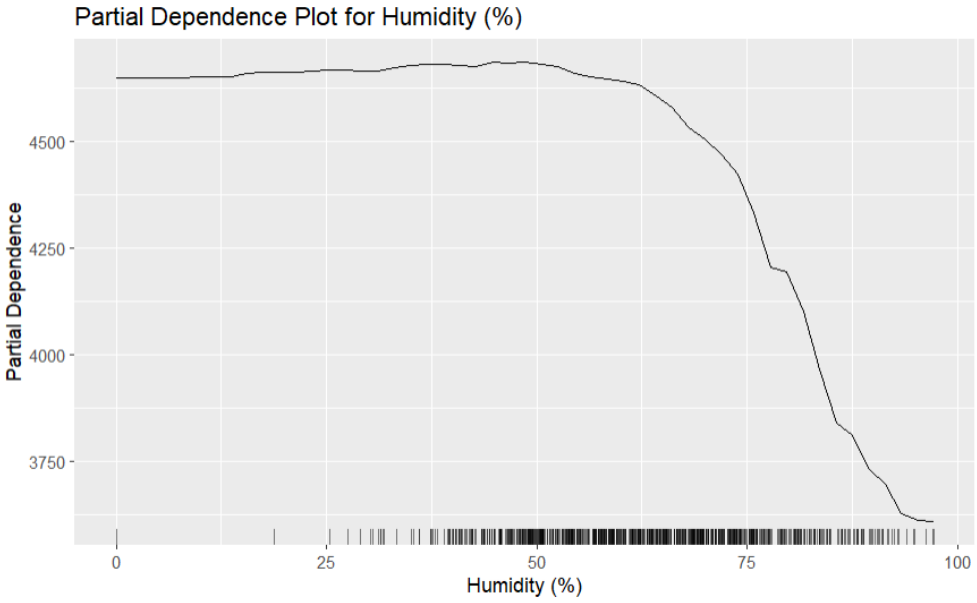


In PDP plots, it is important not only to look at the line indicating how the price changes with the temperature variable, but also to consider the distribution or number of records at each value of the variable whose impact we want to see. This is illustrated by the black lines on the x-axis.

Now that we understand this, we can begin to analyze the plots. By focusing on the values with the most records, we can say that it seems when the temperature is low, bicycle rentals are fewer. When the temperature rises from 5 to 10 degrees, rentals appear to increase slightly, while the most significant change occurs between 12 and 18 degrees, where it stabilizes at higher values. Additionally, when the temperature rises above 25 degrees, bicycle rentals start to decrease slightly, likely due to the conditions being too hot for biking. Although it seems to continue decreasing as the temperature rises further, there are few cases, and this cannot be confirmed. However, logic suggests that it would indeed continue to decrease.

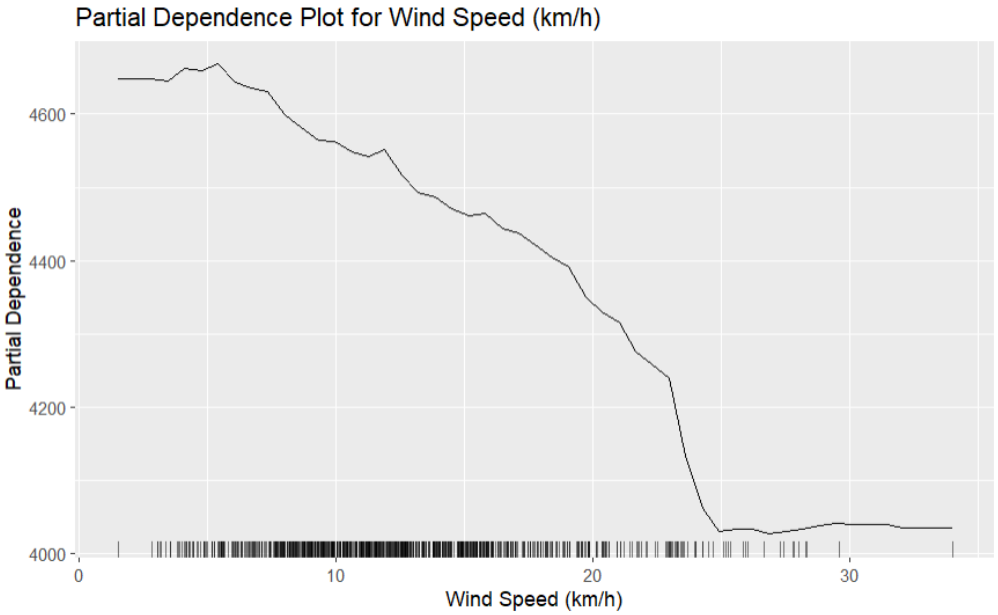


In the case of days, we do not have the problem of distribution since we have one record per day value, allowing us to observe the passage of time correctly. In this case, we see a gradual increase during the first 110 days from 2011. Then, between 110 and approximately 350 days, it remains stable. After that, it grows slightly until day 410, where it suddenly experiences a significant increase in a short period. Following this spike, it continues to grow slightly until it starts to decline around 650 days. This initial growth could be due to the company becoming more well-known. The significant spike might be due to an increase in stock and the number of bicycles available for rent.



Regarding the humidity variable, although it remains practically constant and very high until 50, we will not consider this because there are very few values below 50% humidity. What we can observe is that as humidity increases, the number of bicycles rented decreases, which suggests that higher humidity results in fewer bicycles being rented, or

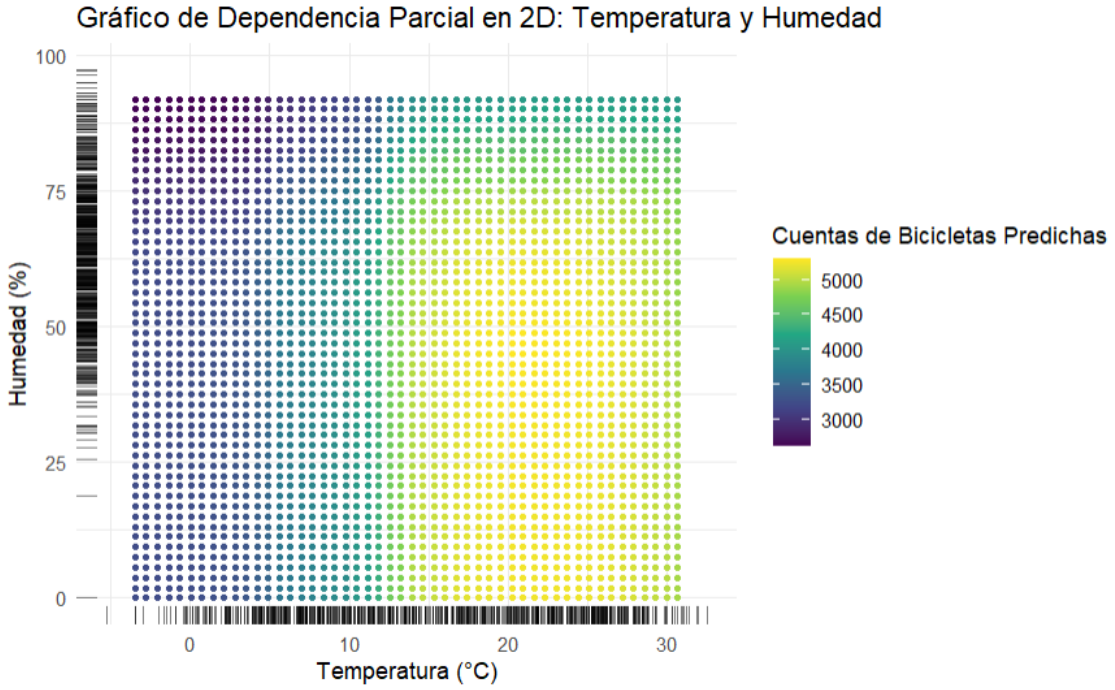
at least this is the case from 50% humidity onwards.



We will interpret wind speed similarly to humidity. It seems that when there is less wind, bicycle rentals are higher, and like with humidity, there is a clear and consistent decrease. It is true that values above 24 km/h are less frequent, so we cannot make accurate assessments beyond that point. Nonetheless, it seems we can also say that higher wind speeds result in fewer bicycle rentals.

Bidimensional Partial Dependency Plot

Now we will create a two-dimensional PDP with the variables temperature and humidity, which helps us see the interaction between these two variables.



Looking at the graph and considering the distribution of humidity, we can see that we should focus mainly on the upper part of the graph (humidity greater than 50) and draw conclusions from there.

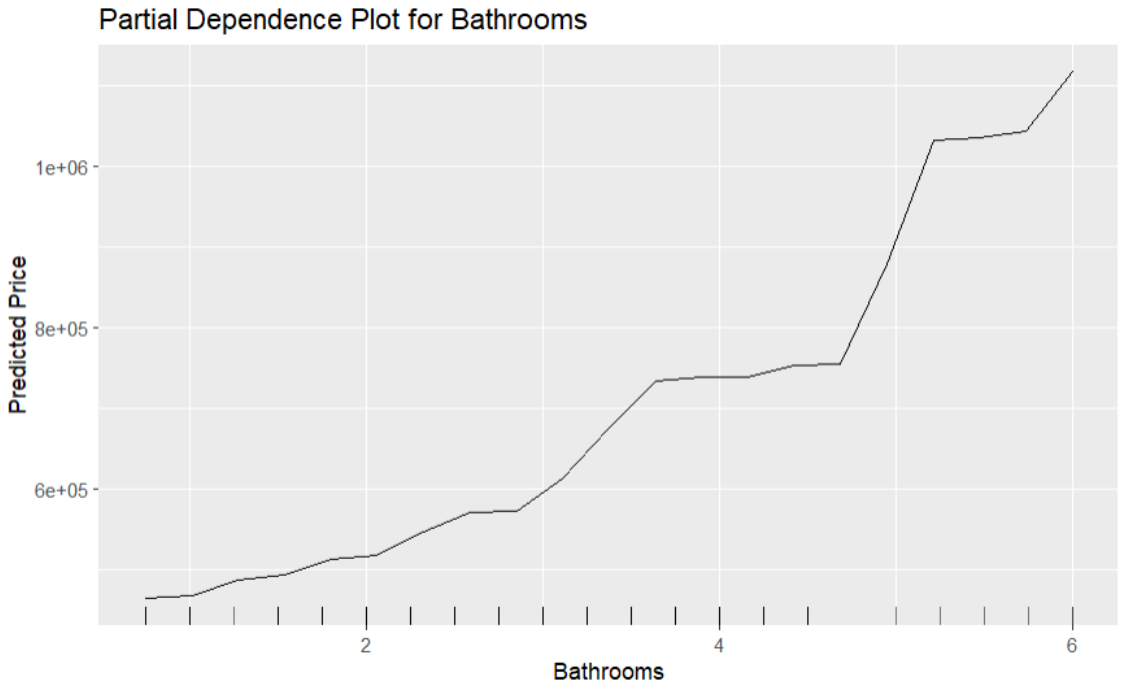
Looking at this graph, we see that the combination that gives us the highest number of bike rentals is when the temperature is between 20 and 25 degrees Celsius and the humidity is around 50. Beyond this optimal point, any deviation from these values decreases the number of bike rentals. For example, even if the temperature is optimal, if the humidity is above 80, the number of bike rentals would decrease from about 5000 to about 4000. But it's when, in addition to the humidity rising, the temperature drops that we see a significant decrease. In fact, when the humidity is above 80 and the temperature is around 10 degrees Celsius, around 3500 bikes are rented, and if it drops below 10, the number drops to 3000. It also seems that temperature has more weight than humidity because with a temperature yielding poor results, it reaches lower values than with humidity at a poor level if the temperature is good.

PDP to explain the price of a house

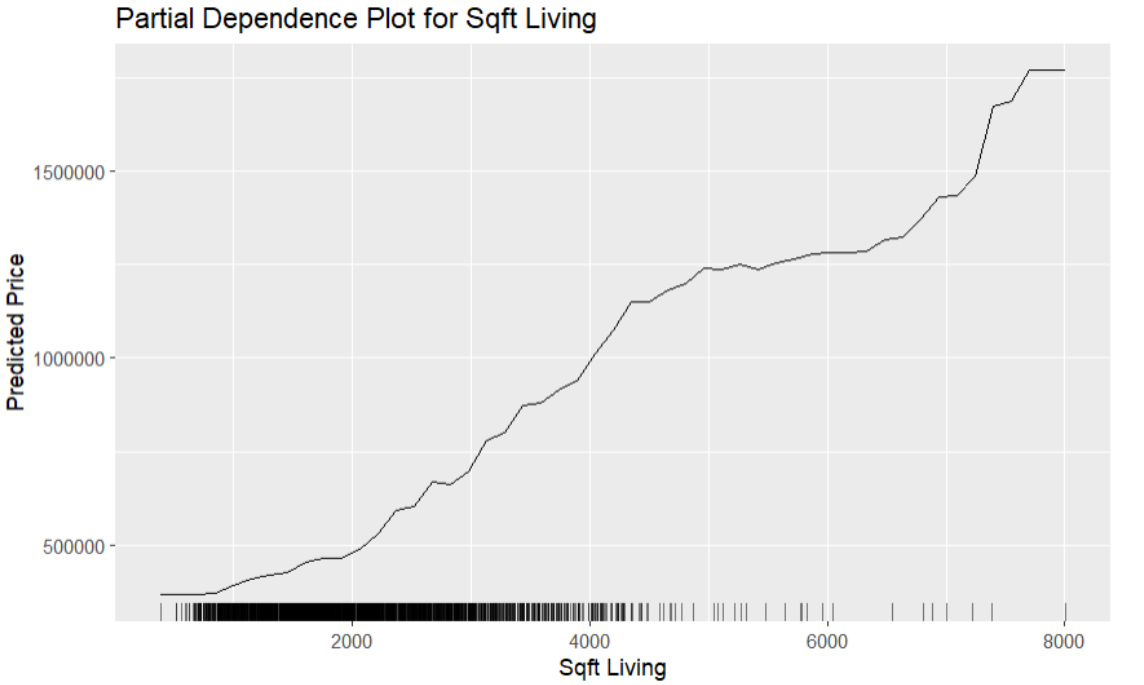
Now we will move on to doing individual PDPs for some variables in the new housing dataset.

Now we are going to assess the importance of the 'bedrooms' variable, which indicates the number of bedrooms in the house. Taking into account the distributions, it seems that apart from 0, houses with 6, 7, and 8 bedrooms should not be considered much as they have few records. Interestingly, houses with 2 bedrooms tend to be the most expensive, while when they have 3, their price decreases, and the lowest-priced ones are those with 4 bedrooms. From here, we could speculate that houses with this number of bedrooms are

less in demand. It's curious that the price increases again with 5 bedrooms.

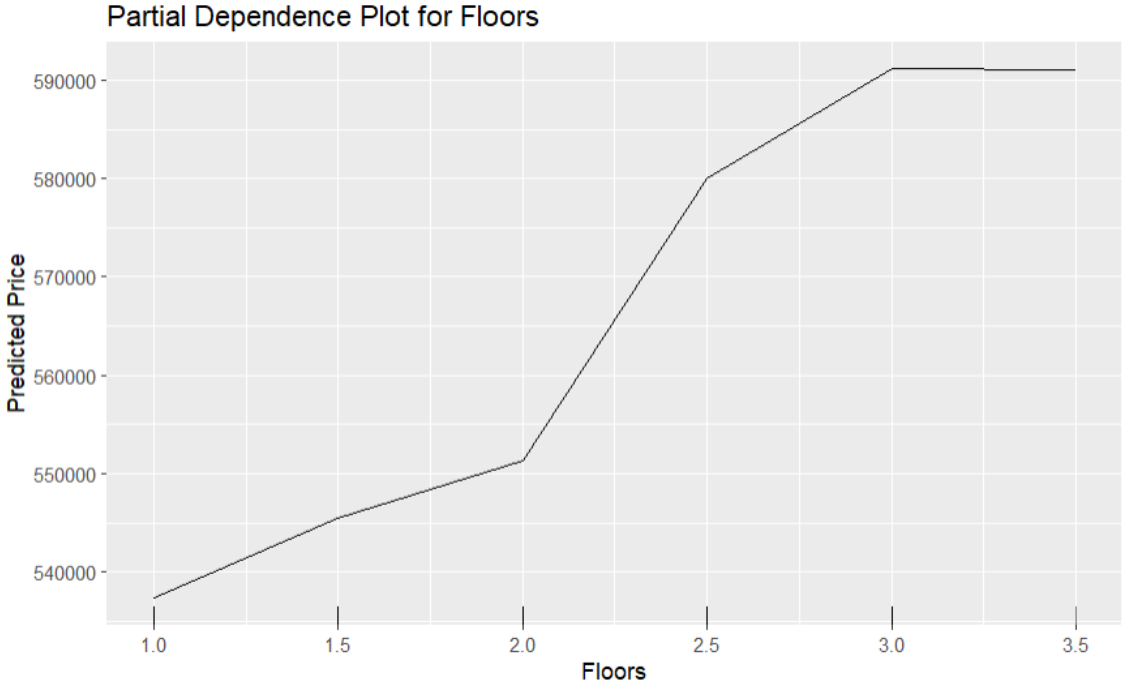


When evaluating the importance of the number of bathrooms, we see that there are few houses with between 4 and 6 bathrooms, so we will not consider these results. However, with 6 bathrooms, there do seem to be enough. We see that as the number of bathrooms increases, the price of the house also rises, and indeed the value will be higher when there are 6 bathrooms.



In the variable of square footage of the house, there are very few values beyond 4000 square feet, so these interpretations are not significant. Nevertheless, we see that as the

size of the house increases, the price also goes up, as one would logically expect.



Finally, let's evaluate the number of floors in the house. Houses with 3.5 floors are infrequent, so we will not consider them for conclusions. Here we can see that the more floors the house has, the higher the predicted price. Furthermore, we can observe that the price change from 2 floors to 3 is greater than from 1 floor to 2.