Escola Tècnica
Superior d'Enginyeria
Informàtica
etsinf

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

DEPLOYMENT: MODEL-AGNOSTIC METHODS

Evaluation, deployment and monitoring of models
Bachelor Degree in Data Science
Universitat Politècnica de València (UPV)

Mikel Baraza Vidal
Adriana Chust Vendrell
Belén Inglés Granero

Index

Introduction

Partial Dependence Plots (PDPs) are a powerful and widely used tool in machine learning for understanding the relationship between a target variable and a single model feature. These plots provide a graphical representation of how the target variable changes as the feature of interest vary while keeping all other features constant.

PDPs are particularly useful for identifying and visualizing non-linear relationships between the target variable and the feature of interest, which can be challenging to detect using other methods. They are often used in conjunction with other techniques, such as feature importance and permutation importance, to gain a comprehensive understanding of the factors that contribute to a model's performance.

In this report, we will explore the concept of PDPs in detail. We will also discuss the advantages and limitations of PDPs and provide examples of how they can be used with different models in order to gain insights into complex models and improve model performance.

One dimensional Partial Dependence Plot

A One-dimensional Partial Dependence Plot (PDP) is a graphical representation that shows the relationship between a target variable and a single feature of a machine learning model while holding all other features constant.

To create a one-dimensional PDP, the feature of interest varies across its range, and the model's output is recorded at each feature value. The resulting plot shows how the target variable changes as the feature of interest varies while keeping all other features constant.

This type of plot is particularly useful in identifying and visualizing non-linear relationships between the target variable and the feature of interest. It can be used to gain insights into how changing the value of the feature will impact the model's output.

In this section, we will perform a random forest regression to predict the number of bike rentals in one day based on several variables, such as *days since 2011, temperature, humidity, wind speed*…

The basic idea behind random forest regression is to create many decision trees, each trained on a different subset of the input data and a random subset of input features. Then, the model averages the predictions of all the decision trees to arrive at the final output. This ensemble approach helps reduce the model's variance and improve its overall accuracy.

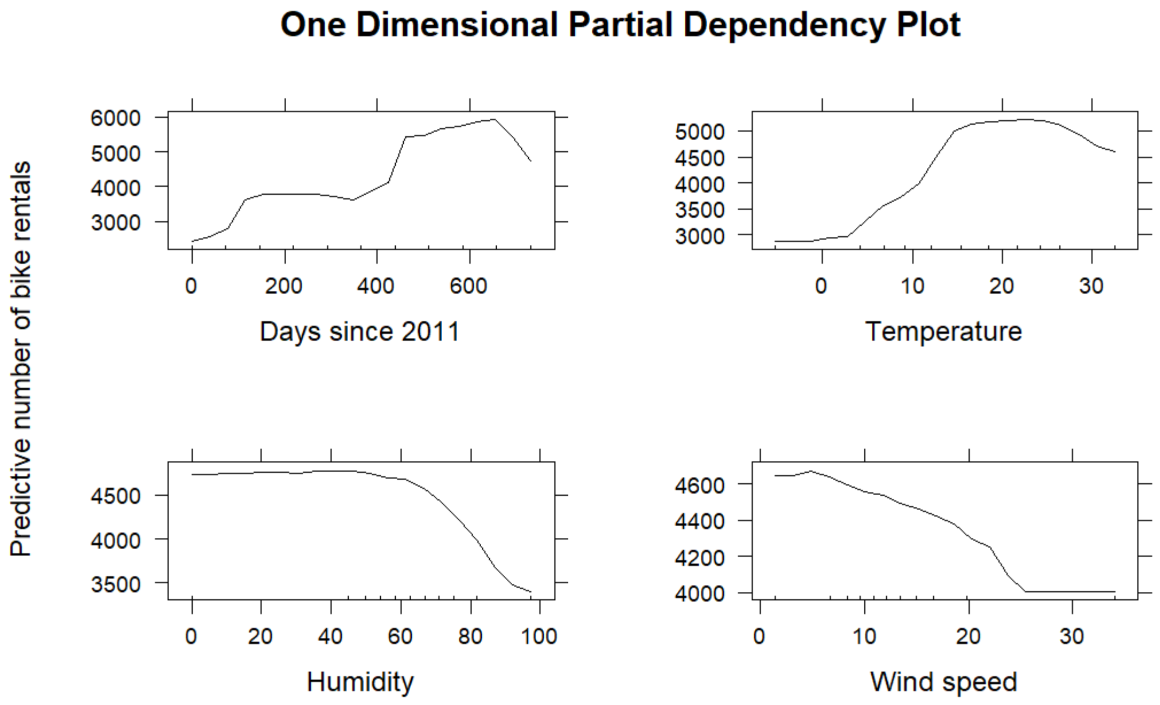

Figure 1. Unidimensional PDP

In a random forest regression model, the PDP shows the relationship between the feature of interest and the predicted target variable after averaging across all the decision trees in the forest. This helps to reduce the impact of individual decision trees on the overall model and provides a more robust estimate of the relationship between the feature and the target variable.

In order to understand this plot, we need to take into account the y-axis magnitude, the shape of the curve, and the x-axis density.

Regarding *days since 2011*, there are samples in the range from 0 to 700 days. The predicted number of bikes seems to increase the further it is from 2011. However, as the shape of the curve is neither linear nor monotonic, it seems that the relationship between the number of bikes predicted and this variable is more complex, or there might be an interaction with one or more other features in the model.

For warm but not too hot weather, the model predicts a high number of rented bicycles on average. For temperatures around 0ºC, there are not enough samples, and it happens the same around 30ºC. Temperatures from 15ºC to 25ºC show a high magnitude of the effect of temperature on the predicted target variable, around 5000 bikes rented.

For humidity above 40%, rental of bikes predicted is around 4500. However, from 70% onwards it starts to decrease, which makes sense, as it is more difficult to ride a bike when humidity increases. For humidity under 40% there are not enough samples to be able to make any conclusion.

The lower wind speed is, the more bikes are predicted to be rented. Its peak is below 10 km/h, being around 4600 bikes. As wind speed increases, the number decreases, being 20 km/h, a stop point where there are not any further samples. Wind speed from 2-5 km/h does not have enough training data, maybe because there are not many days with that feature.

Bidimensional Partial Dependence Plot

Partial Dependence Plot can also be visualized for two variables. It is commonly helpful in visualizing the relationship between two variables while holding all other variables constant. Bidimensional partial dependence plots are so popular because they provide a simple and intuitive way to see how a particular variable affects the model's prediction.

By plotting the model's predicted values against a range of values for the two variables, we can see how changes in each variable impact the prediction. This can be useful for identifying patterns and relationships that may be absent from just looking at the raw data.

Additionally, bidimensional partial dependence plots can be used to identify potential interactions between two variables. An interaction occurs when the effect of one variable on the prediction of the model depends on the value of another variable. By plotting the partial dependence of two variables, we can identify potential interactions and better understand how they may impact the model's prediction.

Using the previous section as a foundation (we used the same model), we plot the bidimensional partial dependence plot for variables hum and temp, respectively, Humidity and Temperature in each day of bike rentals.

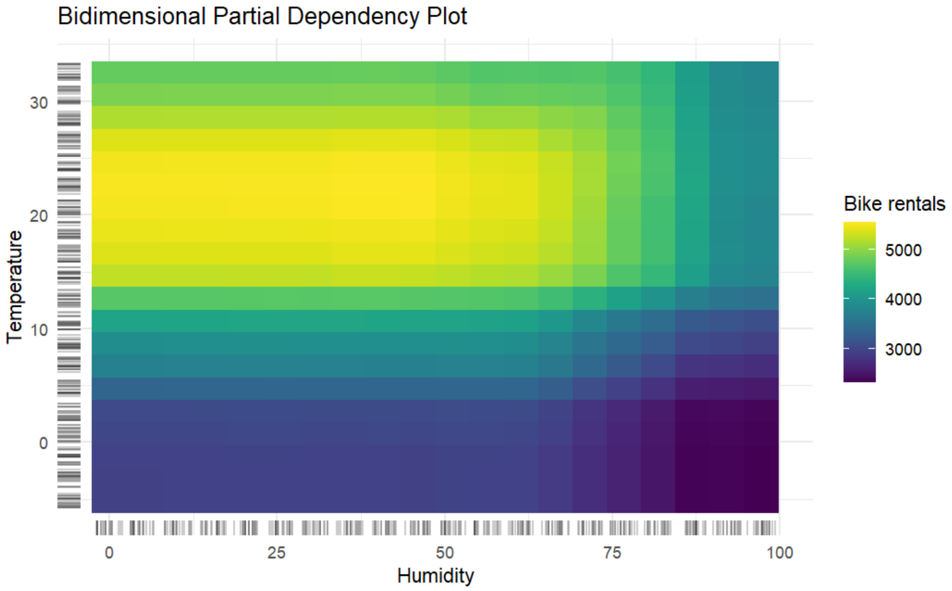

Figure 2. Bidimensional PDP

The plot shows the increase in bike rentals when humidity is lower than 60%, and temperature is between 14ºC and 26ºC. The worst bike rentals tend to be when the humidity is higher than 80%, and the temperature is lower than 5ºC, which are, obviously, the worst weather conditions. Regarding the distribution of the variables, both seem to be equally distributed and it can't be seen values with much higher density than others.

However, these conclusions can't be taken as causal effects since the correlation of both features does not necessarily imply causation. In a bidimensional partial dependence plot, we can only observe the relationship between two variables while holding all other variables constant. Therefore, any observed correlation between the two variables may be due to confounding variables or other factors not included in the model.

PDP to explain the price of a house

As we previously said, the PDP is a graph that displays the connection between a target variable and a solitary attribute of a machine learning model, maintaining all other features constant.

In the following dataset, different houses prices are studied along with their information on characteristics such as square footage of rooms, number of bathrooms, or floors, for instance. Initially, we will perform a random forest in order to predict the resulting price according to different variables. This allows us to attempt to study the correlation they have with price.

We first divided the data into a training set and a testing set. And we tried to get the predictions using the following variables: bedrooms, bathrooms, sqft_living, sqft_lot, floors, and yr_buil.

To visualize the importance of variables graphically, we use the **varImpPlot** function. Although we have represented all the features in the PDP, only bedrooms, bathrooms, sqft_living, and floors will be interpreted.

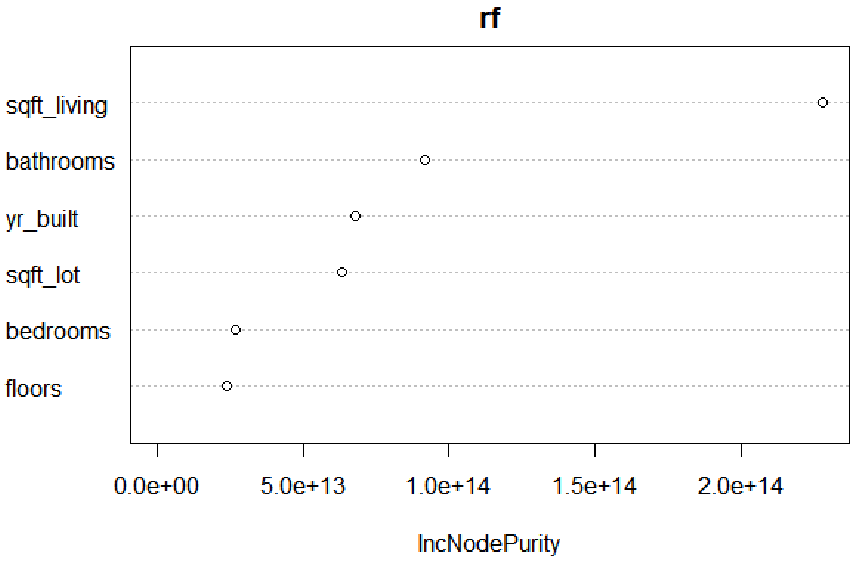

Figure 3. Feature importance PDP

To discuss the results obtained in the graph, we must first understand that the IncNodePurity measures the importance of a variable in a Random Forest model. It indicates how much the impurity (i.e., lack of homogeneity) in the tree nodes would be reduced if that variable were used to split the data. A high value indicates that the variable is important for prediction.

After having explained this, we can easily observe that the variable referring to the size of the floor in square feet (value = 2.0e+14) is the most important for making the prediction and, therefore, crucial when determining the house price.

In addition, it can be seen how the number of bathrooms, bedrooms, and floors follows it. The lower importance of these variables is so interesting and may be due to the fact that they are implicitly included in the square footage of the dwelling. In other words, larger houses have more bedrooms, bathrooms, and floors, so these variables are indirectly represented through the square footage.

We will now show the one-dimensional partial dependence plot, that as we have said, shows the influence of one variable on the target variable being predicted, while keeping all others constant.

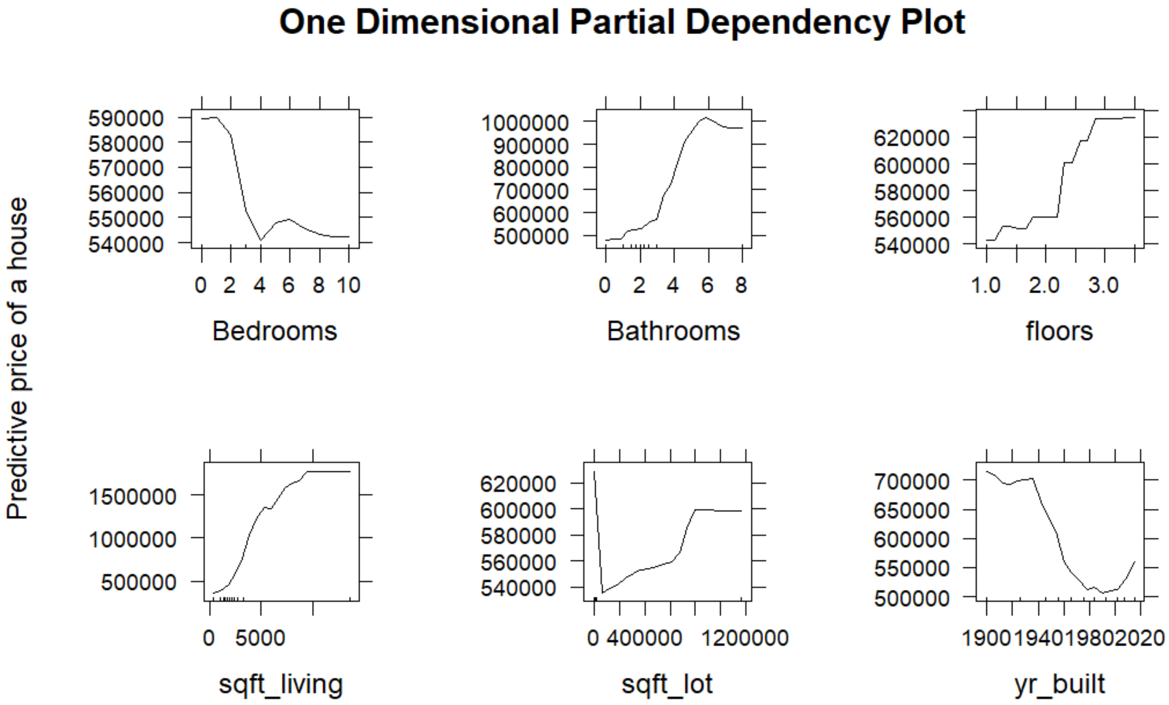

Figure 4. Housing dataset PDP

Regarding the number of bedrooms, we can see how it moves up to a maximum value close to 10. We can see that the curve does follow a monotonically decreasing direction as the number of bedrooms increases. We can also see how the predicted values of higher prices are found in those houses with values of approximately 0 to 4 bedrooms. This could be due to multiple reasons. As we have mentioned, the house size implicitly includes the number of bedrooms or bathrooms, and as we have seen, it is much more related than the rest to the price variable. This can make variables like this not be predicted correctly.

On the other hand, it could also be due to a misinterpretation of the database. Perhaps the available beds are counted as the number of bedrooms, making rooms with ten beds, such as bunk beds, have a lower price. This is not likely, but we cannot deny it as we are not sure how this variable is being taken.

If we now focus on the bathrooms, we can see how we enter values between 0 and 8. Here we see a positive relationship, as the predicted price increases as the number of bathrooms increases. As we have said, the apartment size implicitly includes the number of bathrooms. As bathrooms are smaller rooms, they may not be affected in the same way as the number of bedrooms seems to have been affected. These are significant changes since the price varies by approximately 100000 euros for every two bathrooms.

Moving onto the number of floors, we also see a positive relationship with the price; as the number of floors increases, so does the cost of the house. There are options from 1 to 3 floors, and approximately for each additional floor, the price increases by 520000 euros. This may seem like a small change since we are talking about floors, but it is because the size of the house also includes the number of floors, which may cause these variables not to relate correctly.

As for the apartment's square footage, we see a clear positive relationship, with values approximately between 0 and almost 150000 square feet. We can see how the price varies by about 50000 euros for every 5000 square feet. This is where we are allowed to see the highest prices, as the size, including many other implicit variables, will be the best related to the price, as we can observe in the variable importance plot.

When it comes to the square lot of the apartment, we observed how the price increases when sqft_lot is higher than 400000. However, as we can see in the distribution, the majority of houses are near zero sqft_lot. Therefore the predictions tend to be lower for this variable.

Finally, the year of the build also does not follow a normal distribution, since most houses are around 1960 and 1990. The price predictions are higher when the year of construction is older, being the newest homes cheaper.