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

Jonas Bachmann, Juan Alfonso García Real, Xinyi Zhuo

May 16, 2025

Introduction to Partial Dependence Plots

As machine learning models grow increasingly complex, they often become black boxes, making their behavior difficult to interpret. This lack of transparency is particularly concerning when such models are used in contexts that impact human lives, such as healthcare, finance, or criminal justice. To address this, we need robust methods that can help explain the decisions made by these models. One powerful family of such techniques is known as model-agnostic methods, which can be applied to any type of model regardless of its internal structure. In this exercise, we will explore one of such model-agnostic approaches: the *Partial Dependency Plot* (PDP), which helps us understand how individual features influence the model's predictions.
The method works by fixing the value of the chosen features in the dataset and averaging the predictions over all the modified samples in the dataset. This corresponds to the value

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{N} \hat{f}(x_s, x_{C_i}),$$

where $\hat{f}$ is the predictor, x_S are the fixed features, and x_C are the other features. One important assumption is that there is no feature interaction between x_S and x_C, meaning they are uncorrelated. If this is not the case, the average might include datapoints that are very unlikely or impossible. However, if independence is given, these plots perfectly show the average relationship between the features and the target.

1 One-dimensional Partial Dependency Plot

We consider a bike rental dataset and the regression problem of predicting the bike rental count given a variety of different features. We fit a Random Forest model and build PDPs for single features to analyze their influence on the predictions of the model. For each feature, we measure the partial dependence for 30 equally spaced points in its value range.

1.1 Number of Days since 01.01.2011

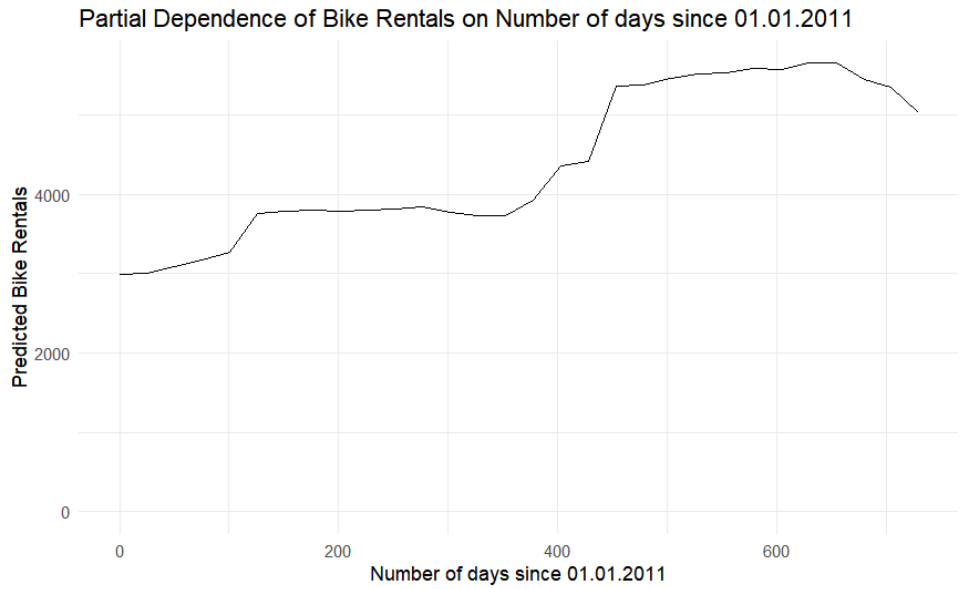

Figure 1: Partial Dependence Plot for the number of days since 01.01.2011

We can see in Figure 11 that the predicted count increases over time. However, for the duration of roughly half a year, the prediction plateaus. Given that the measurement starts on the 01.01., we see that bike rental counts increase until the summer, then stay roughly the same within that year and then increase significantly until the next summer.

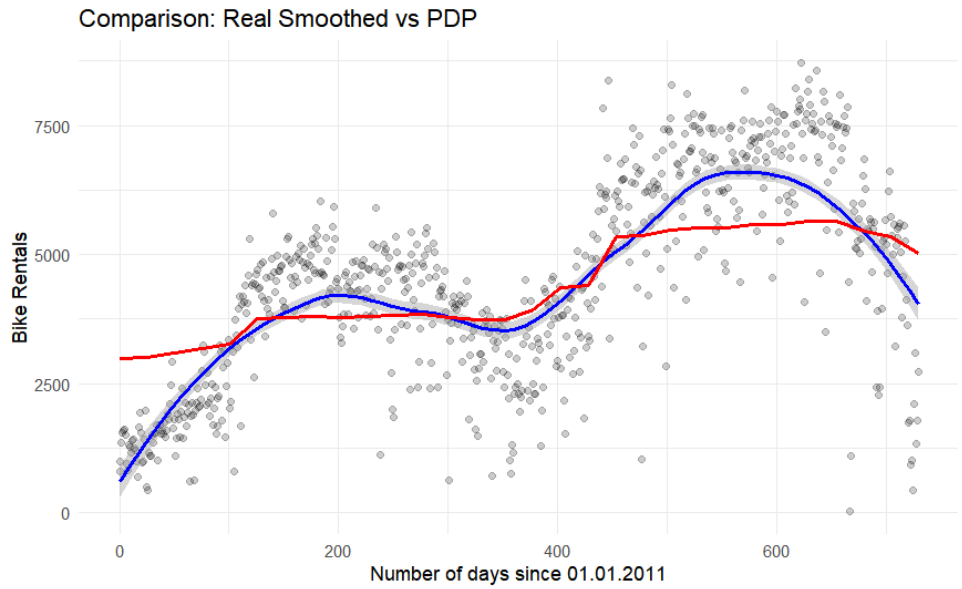

Figure 2: Real Smoothed Data Distribution (blue) vs. Partial Dependence Plot (red).

In Figure 2, it is evident that the real fluctuations during a year are larger than the PDP shows. This might be because the model has learned that while the rental count increases over the years, within one year other features (such as temperature) are better correlated and suited to predict the rental count.

1.2 Temperature

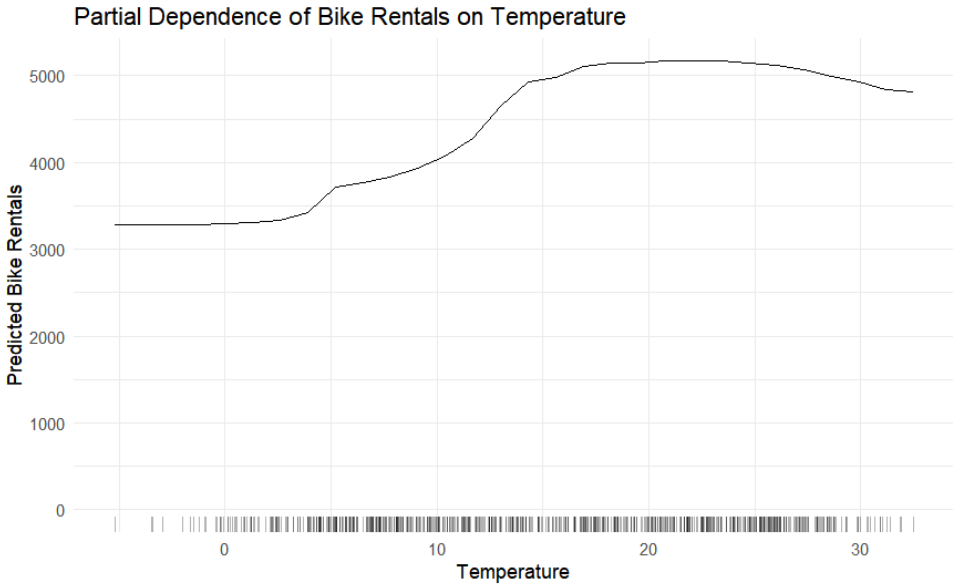

Figure 3: Partial Dependence Plot for Temperature. The tick marks over the x-axis indicate the density of the temperature feature.

The PDP in Figure 3 shows that temperature has a significant influence on the predictions. The temperature alone can make a difference of around 2000 rentals, which is significant given a maximal total average of rentals of 5000 per day. Note that rental counts tend to decrease once the temperature exceeds the 25 degrees mark. The sweet spot lies between 15 and 25 degrees, not too cold to freeze and not too hot to sweat.

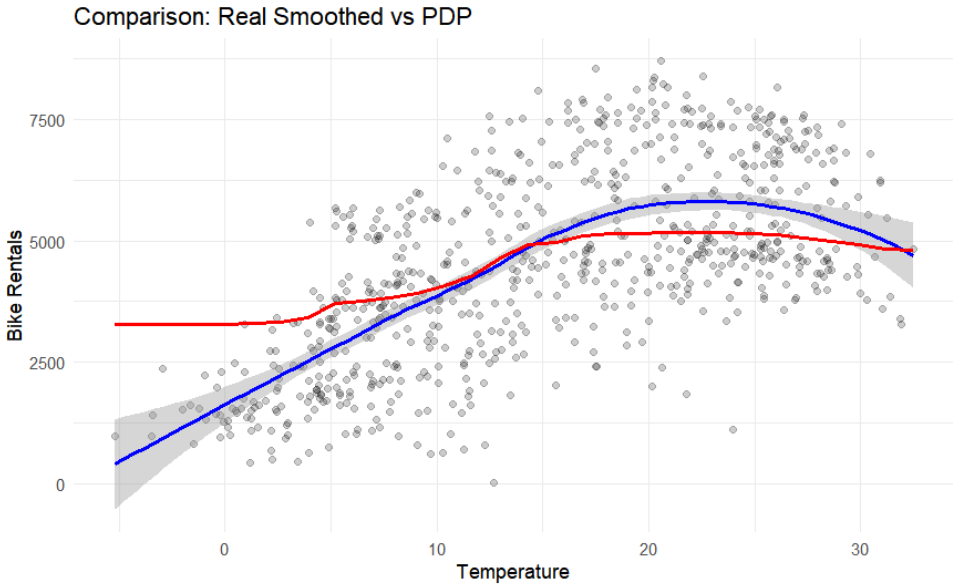

Figure 4: Real Smoothed Data Distribution (blue) vs. Partial Dependence Plot (red).

While the model follows the curve of the real data, we can see in Figure 4 that it underestimates its effect. This is certainly due to correlations with other features. For

example, we would probably never see a datapoint with 0 degrees of temperature in the summer. As a result we average over unlikely datapoints. Additionally, the model might explain the drop in rental numbers in winter not only by temperature but also by correlated features, such as the binary winter season feature.

1.3 Humidity

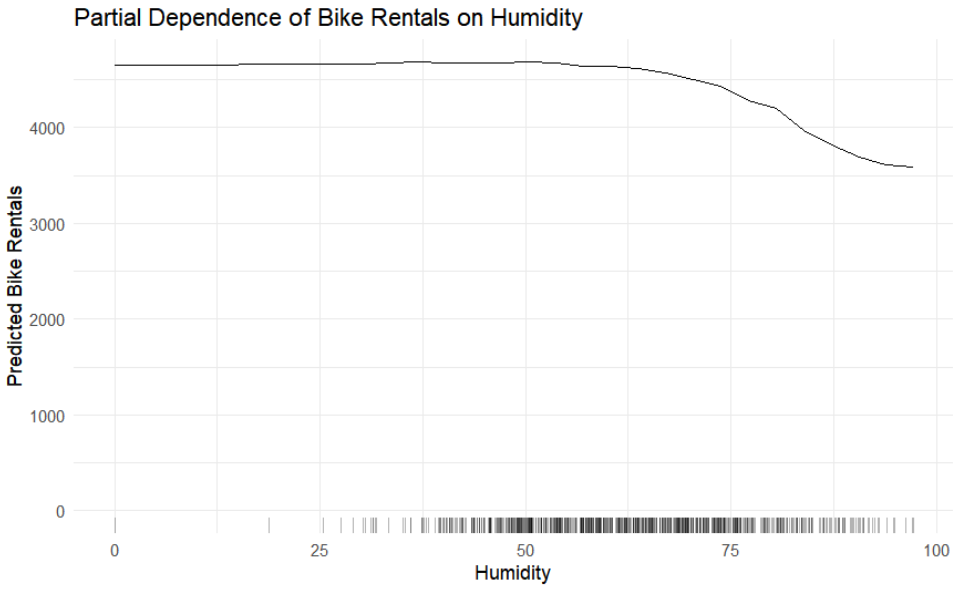

Figure 5: Partial Dependence Plot for humidity. The tick marks over the x-axis indicate the density of the humidity feature.

The partial dependence of the prediction on humidity in Figure 6 seem intuitive. Below 40% humidity we don't have many datapoints in the dataset, so the model could not learn to make reliable predictions in that area. We observe that once the humidity exceeds 70%, the prediction begins to decrease. This can be explained by the higher chance of rain. The correlation between having rain and humidity $> 80\%$ is approximately 0.4.

1.4 Windspeed

2 Bidimensional Partial Dependency Plot

To analyze the joint effect of temperature and humidity on bike rental predictions, a 2D Partial Dependence Plot was generated using a Random Forest model. Due to the size of the dataset, we used a random sample of 500 observations to speed up computations, as suggested in the assignment.

The following figure illustrates the predicted number of bike rentals depending on the temperature (in °C) and humidity (in %) using the `geom_tile()` function.

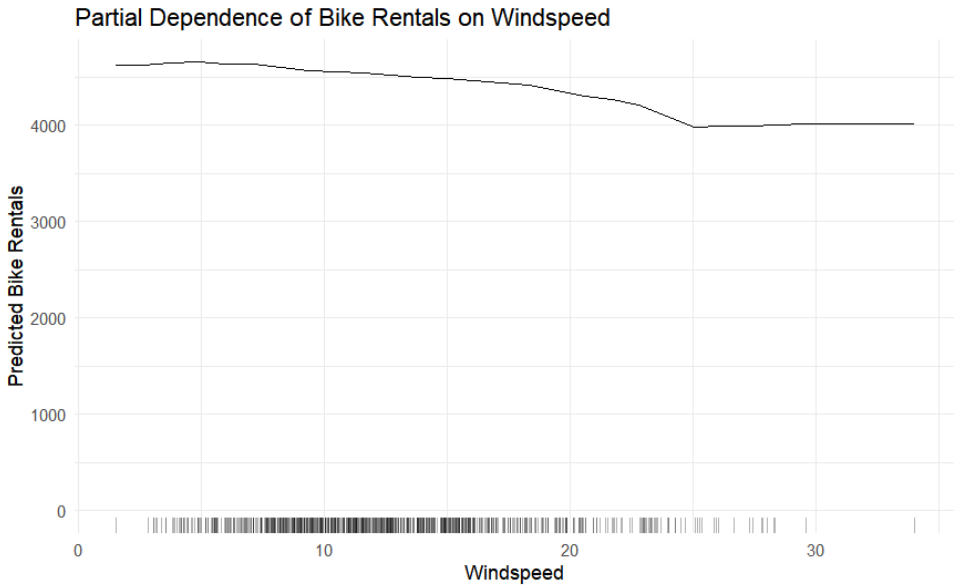

Figure 6: Partial Dependence Plot for windspeed. The tick marks over the x-axis indicate the density of the windspeed feature.

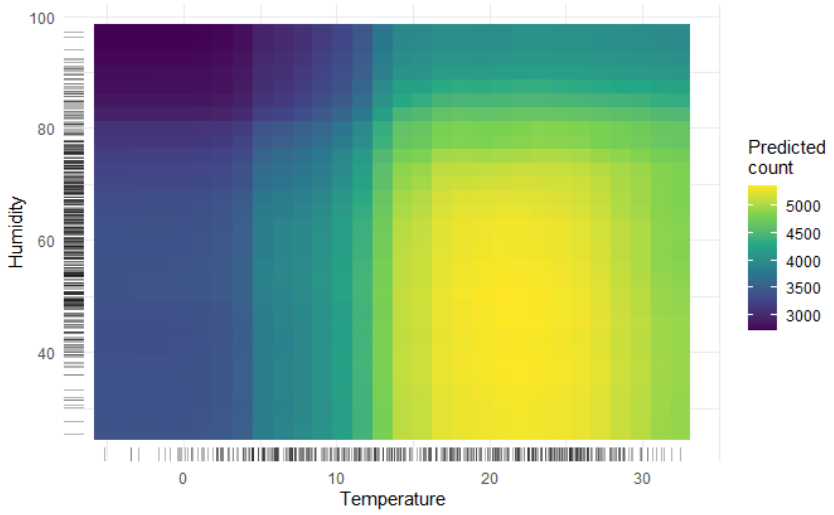

Figure 7: 2D Partial Dependence Plot for temperature (ºC) and humidity (%). Marginal density distributions are shown on the axes (rug plots).

Interpretation: The plot shows that the predicted bike rentals increase significantly with moderate temperatures, especially around 15–25°C. High humidity levels, however, tend to reduce the number of predicted rentals across all temperature values. The optimal condition for bike rentals, according to the model, occurs at moderate temperatures (approximately 20°C) and low to moderate humidity (below 60%). On the contrary, cold and highly humid days show lower rental predictions.

3 PDP for house price prediction

4 Conclusion

PDPs can give a good estimate about the global effect of one feature on the target variable. However, this only holds true as long as the correlations between features are small or non-existent. If correlations exist, we include improbable or impossible datapoints into our average. It may also happen, that the effect of one feature is completely or parcially explained by another feature. In that case, the PDP only serves for knowing how much the model relies on the value of the feature alone for its predictions. The real relationship between the target variable and the feature can not be seen due to the missing feature interactions. In that case, the individual conditional expectation plot (ICE) is probably the better option.