XAI Exercise 5: Model-Agnostic Interpretability Using Partial Dependency Plots

Jaime Ballester Solá, Juan Carlos Trujillo Rodríguez, Marc Romeu Ferrás

May 16, 2025

Introduction to Model-Agnostic Methods

Model-agnostic interpretability methods are versatile techniques designed to explain the predictions of almost any machine learning model, regardless of its internal complexity. This capability is particularly valuable when dealing with "black-box" models, such as the Random Forests employed in this exercise or neural networks, where understanding model behavior through direct inspection of parameters is often challenging or impossible.

These methods offer significant flexibility: they can be applied across different model types without requiring changes to the explanation technique itself, and they provide various forms of explanations—be it graphical, statistical, or focused on individual predictions. The core idea is to place an interpretability layer between the model's predictions and the human user, making complex behaviors more transparent. Partial Dependency Plots (PDPs), which have roots in statistical analysis, are a fundamental model-agnostic technique explored in this report. PDPs show average feature effects; interpretations are most reliable where data is dense (often indicated by rug marks or marginal distributions) and less certain in sparsely populated regions. It's also worth noting they might obscure heterogeneous individual effects, a limitation that techniques like Individual Conditional Expectation (ICE) plots address.

1 One-Dimensional Partial Dependence Plots for Bike Sharing

For this exercise, one-dimensional Partial Dependence Plots (PDPs) were computed for key variables influencing bike rentals, using a Random Forest model trained on the Capital Bikeshare dataset. The rug plot (small ticks along the x-axis) in each PDP indicates the distribution density of the data points for that feature.

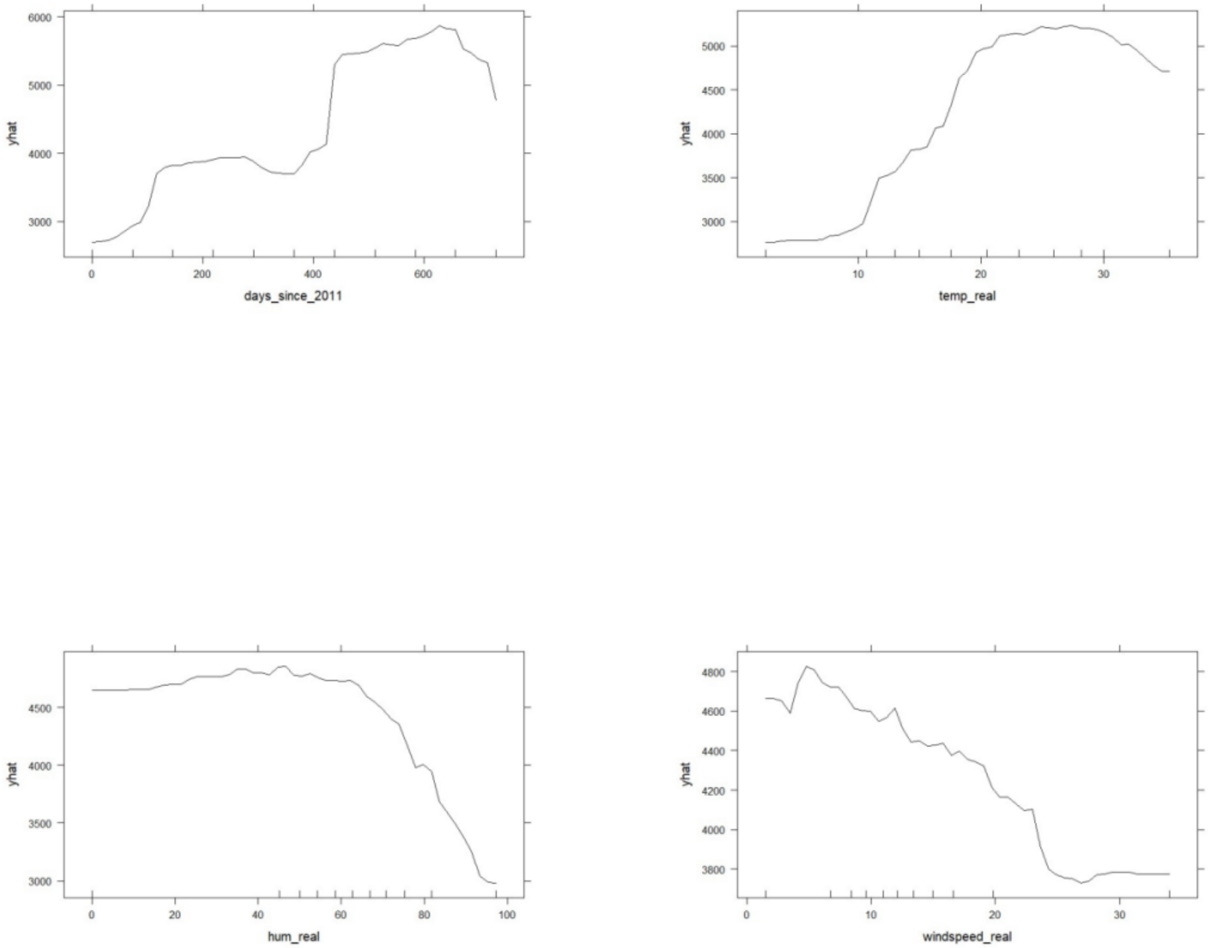

Figure 1: PDPs for features influencing bike rentals. Rug marks indicate data density.

Interpretation of 1D PDPs for Bike Sharing

The Partial Dependence Plots (PDPs) for key variables influencing bike rentals (Figure 1) provide valuable insights, keeping in mind that conclusions are most robust where the data (indicated by rug marks) is concentrated.

The PDP for **days_since_2011** reveals a generally increasing trend in the predicted number of rentals over time, showing two outstanding increases around 80 days and around 420 days. The rug marks are fairly evenly distributed, suggesting this trend is well-supported across the observed period. This indicates a growth in bike usage, likely reflecting increased adoption in two dates or seasonal patterns embedded within the overall upward trend.

For **temp_real** (temperature), the PDP exhibits a clear, nonlinear relationship. Predicted rentals rise significantly as temperature increases from low values 10°C up to an optimal range of approximately 25-30°C. Beyond this, rental activity slightly declines. The rug marks show a high concentration of data points in the moderate temperature ranges(10°C to 30°C), making the observed peak and decline in these areas reliable. Interpretations at extreme cold or hot temperatures, being the data sparse there, should not be made.

Humidity, represented by **hum_real**, shows a distinctly negative effect. Rentals are highest at low to moderate humidity levels and decrease sharply once humidity exceeds 65%. The data appears well-distributed across various this levels (from 45% ro 80%), supporting the reliability of this observed negative trend. High humidity likely deters bike usage due to discomfort. As told before the levels it is not posible to do accurate interpretations.

Similarly, wind speed (**windspeed_real**) displays a consistent negative correlation with rentals. Predicted rentals are high when wind is light (between 8km/h and 10 km/h) and drop steadily with increasing wind speed. The rug plot indicates most data points fall within lower to moderate wind speeds(8km/h to 20km/h), so the sharp decline observed at higher speeds (>20 km/h) should be interpreted with consideration of potentially fewer data points supporting that part of the cure or ignored.

Taken together, these plots illustrate how temporal progression and weather factors shape bike rental behavior. Understanding these dependencies, and where the model's understanding is best supported by data, enables more accurate forecasting and better conclusions.

2 Two-Dimensional Partial Dependency Plot (Temperature × Humidity)

The 2D Partial Dependence Plot (PDP) in Figure 2 illustrates the combined effect of temperature (°C) and humidity (%) on predicted bike rentals. To properly interpret such a plot, one should ideally also consider the marginal distributions (or density of data points) for both temperature and humidity, as predictions are more reliable in regions of the 2D space where data is abundant.

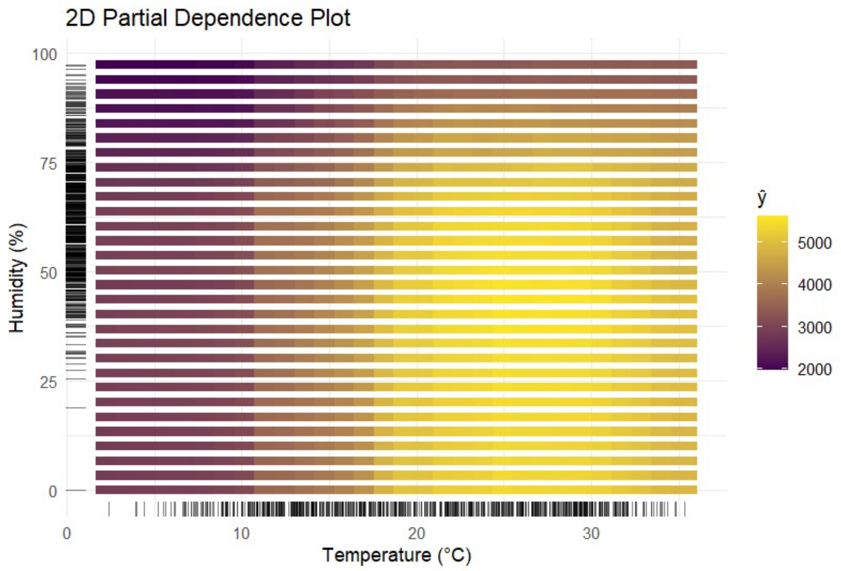

Figure 2: Humidity against Temperature (2D PDP for Bike Rentals). Interpretations are most reliable in data-dense regions of this 2D space.

The plot primarily shows a horizontal color gradient, indicating that temperature (x-axis) is the slightly dominant factor influencing predicted rentals in this model's view: lighter shades (higher rentals) appear as temperature increases. The vertical gradient, representing humidity (y-axis), shows less variation in color, suggesting a weaker isolated effect of humidity as learned by this model, or that its effect is more nuanced when interacting with temperature.

The reliability of these interpretations depends on the data density across the temperature-humidity plane. If, for example, there are very few data points for "high temperature AND high humidity" or "low temperature AND low humidity", the model's predictions in those corners are less certain. Generally, for bike rentals, one would expect:

- Peak rentals at moderate temperatures and moderate humidity, where data is often concentrated.
- Low rentals at temperature extremes, significantly lower for les than 15ºC and slightly lower form more than 30.
- Suppressed demand at humidity extremes above 75%, and has an stable effect for less.

The specific plot in Figure 2 suggests temperature is the primary driver captured by the model, these reasonings are more accurate around 10 to 30 ºC and between 30 to 95 % of humidity as there are more samples.

3 PDPs Applied to House Price Prediction

A Random Forest model was fitted on a subsample of the King County house-sales data. One-dimensional PDPs were generated for selected features. The rug plot on each x-axis in Figure 4 helps assess where data is concentrated.

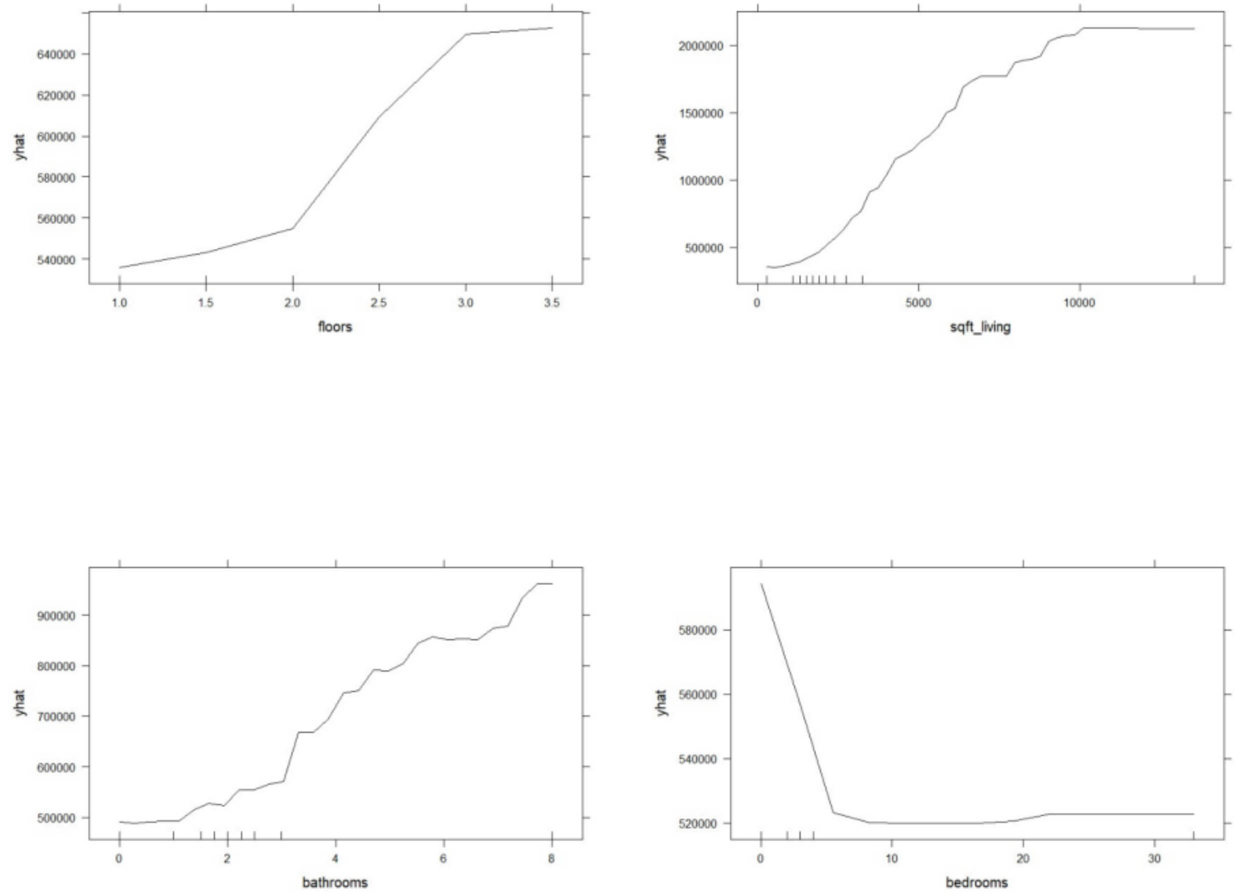

Figure 3: PDPs for house features. Rug marks indicate data density.

Interpretation of PDPs for House Price Prediction

The PDPs for selected house features (Figure 4) offer these insights, which are most reliable where rug marks indicate sufficient data:

The PDP for bedrooms shows a linear pattern. Predicted price appears to decrease from one to around six bedrooms, with a notable spike only at seven bedrooms. The rug marks indicate that most houses in the dataset have between 2 to 5 bedrooms. The behavior at 1 bedroom or 7+ bedrooms is based on fewer instances, making those interpretations less certain. The initial decrease might be due to confounding with ‘sqft_living‘ (smaller houses with many rooms).

For bathrooms, there’s a clear positive relationship with price. Prices rise steadily with each additional bathroom, especially beyond 3. The data is concentrated around 1 to 3 bathrooms, so the strong upward trend for 4+ bathrooms, while clear in the PDP, is based on a smaller portion of the data and a consequence less certain.

The PDP for sqft_living (living area) demonstrates a strong, positive, and mostly linear effect on predicted prices, especially up to around 500 to 4000 sqft where most of the data lies. Beyond this, while the trend continues upwards, however it’s supported by fewer or no data points.

The PDP for floors indicates that two story homes generally have higher predicted prices than single-story homes. The trend for homes with more than two floors (2.5 to 3.5 in the plot) is less clear or shows minimal additional gain, and is based on fewer instances. Most houses in the dataset appear to be 1 or 2 stories.

Additionally, we tried to explain the two remaining variables in the random forest in order to complete our conclusions.

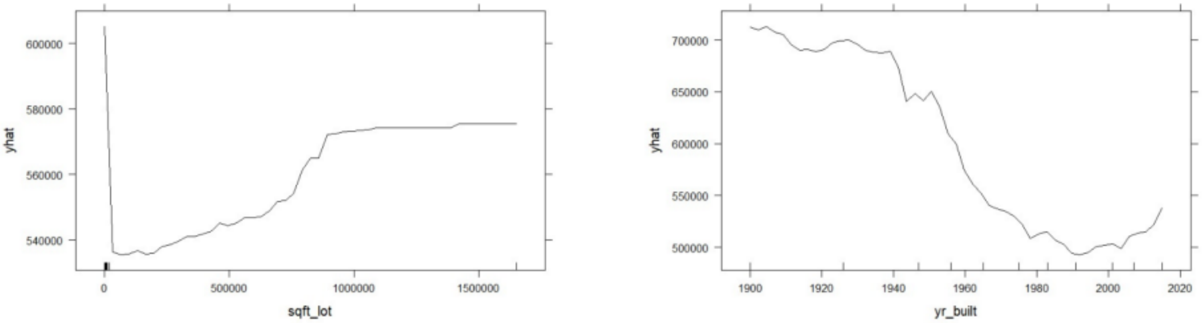

Figure 4: PDPs for sqft_lot and yr_build. Rug marks indicate data density.

For sqft_lot. lot size has a first negative impact untill around 1000 where it apears a positive impact for beyond. Most data points are concentrated in smaller lot sizes, less than 2000, so the behavior for very large lots is less certain.

Lastly, yr_built generally shows newer homes having higher prices, those built for 2000 and beyond, for those built before the price increases as we go back until 1940. The rug marks would confirm the distribution of houses by construction year. Any unusual patterns for very old should not be taken into acount because of data density(earlier than 1940) is lower. For instance, the minimum is located around 2000 is homes followed by a rise for modern constructions is plausible, also there is a rise for older.

Conclusion

This exercise demonstrated the practical application PDPs, we were able to extract meaningful insights into how different features influence predictions in two distinct contexts: bike share demand forecasting and house price estimation.

In the bike-sharing case study, PDPs effectively visualized the impact of temporal trends and key weather variables on rental volumes. The analysis highlighted non-linear relationships and interaction effects, such as the optimal temperature and humidity ranges for maximizing rentals, which carry direct implications for operational planning. Consideration of data density, as indicated by rug marks or marginal distributions, proved crucial for assessing the reliability of interpretations across different feature values.

Similarly, in the house price prediction exercise, PDPs illuminated how structural features like the number of bedrooms and bathrooms, living area, lot size, number of floors, and year built contribute to the model's price estimations. These plots revealed both expected trends and some counterintuitive patterns, underscoring the importance of looking beyond simple correlations and understanding the model's learned behavior, again with due attention to the underlying data distribution for each feature.

Overall, Partial Dependency Plots offer a valuable and accessible approach to enhance the transparency of "black-box" models. They provide a visual means to understand average marginal effects, helping to validate model behavior, generate hypotheses, and communicate complex model logic. While aware of their limitations, such as potentially obscuring heterogeneous effects (better addressed by techniques like ICE plots) and their reliance on data density for robust interpretation.