EXPLAINABLE AI: MODEL-AGNOSTIC METHODS

LUCÍA LIN MARTÍNEZ, ALEJANDRO CHAFER RIBOT,
KIRIL IVAYLOV TZENKOV

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

1 INTRODUCTION

This work focuses on practicing Explainable AI (XAI) by utilizing Partial Dependence Plots (PDP) to interpret the predictions of machine learning models. Specifically, we apply this method to understand how different features, such as temperature, humidity, and the number of rooms, influence the predictions made by Random Forest models for bike rentals and house prices. The goal is to extract

meaningful insights and conclusions about the model's behavior, making it easier to interpret and trust its decision-making process. By leveraging PDPs, we aim to provide a clearer understanding of how these models work, fostering transparency in machine learning applications.

2 ONE DIMENSIONAL PARTIAL DEPENDENCE PLOT.

PDP helps us interpret the Random Forest regression model for bike rentals by revealing whether each feature has a positive, negative, or non-linear influence on the predicted daily rental count. Here we examine PDPs for four key features from the `day.csv` dataset: *Days Since 2011*,*Temperature*, *Humidity*, and *Windspeed*. Each subsection below includes the PDP for the feature (plotted as the feature value vs. predicted bike count) and an interpretation of the relationship it captures.

2.1 Days Since 2011

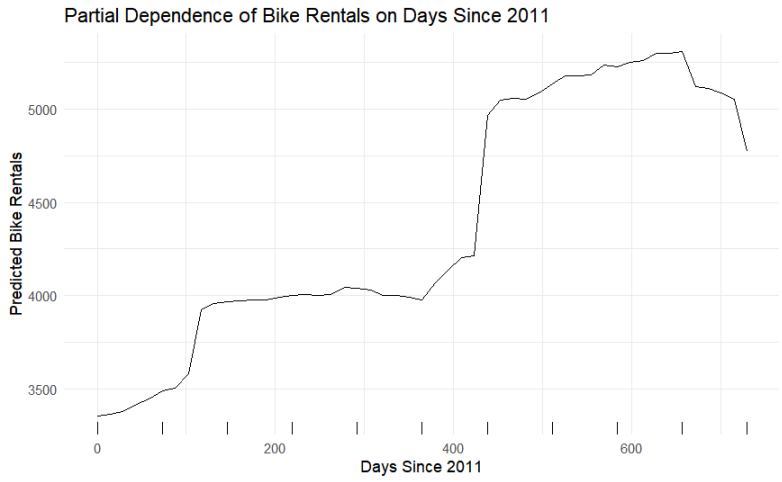

Figura 1: Partial Dependence of Bike Rentals on *Days Since 2011*

Figure 1 illustrates the partial dependence of predicted bike rentals on the *Days Since 2011* feature. The PDP reveals a clear upward trend over time: predicted rental counts are lower in early 2011 and higher in late 2012. In fact, the curve shows a step-like increase, with a lower plateau for dates in 2011 (around 3,500–4,000 predicted rentals) and a higher plateau for dates in 2012 (around 5,000 or more). This indicates that, after accounting for weather and other factors, the model has learned a positive time-based trend – i.e., bike rentals increased as time progressed. The jump between the two plateaus suggests that the year 2012 had significantly higher baseline users than 2011. Toward the end of the timeframe the PDP dips slightly from its peak, which likely corresponds to winter months of 2012 when ridership declined seasonally. Overall, the *Days Since 2011* feature has a strong influence: later dates have higher predicted rentals, implying a temporal growth effect in the model's predictions.

2.2 Temperature

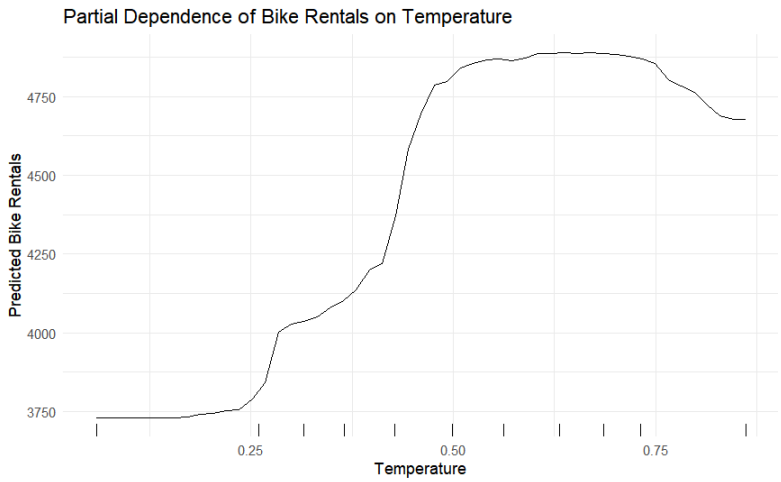

Figura 2: Partial Dependence of Bike Rentals on *Temperature*

The partial dependence plot for *Temperature* (Figure 2) shows a distinctly non-linear relationship with bike rental counts. At the coldest temperatures (left), the predicted number of rentals is very low (around 3,750) and remains relatively flat – indicating that when it is extremely cold, increases in temperature from frigid levels do not yet lead to big gains in usage. However, as the temperature rises into the moderate/comfortable range, the PDP curve climbs steeply. This sharp increase (starting roughly from the 10°C–15°C range onward) suggests that warmer weather strongly boosts bike usage – each additional degree in this range leads to a substantial rise in predicted rentals. The curve then levels off at the higher end of the temperature range: once the weather is warm (around 20°C and above), the predicted rental count reaches a plateau indicating diminishing returns – further warming yields little to no additional increase in usage. Notably, at the hottest temperatures the PDP actually trends slightly downward. This slight decline suggests that when it becomes extremely hot, bike rentals are predicted to drop off. In summary, the model has learned that temperature has a positive effect on ridership up to an optimal point– pleasant warm days see the highest bike rental predictions – but both extreme cold and extreme heat lead to lower predicted rentals.

2.3 Humidity

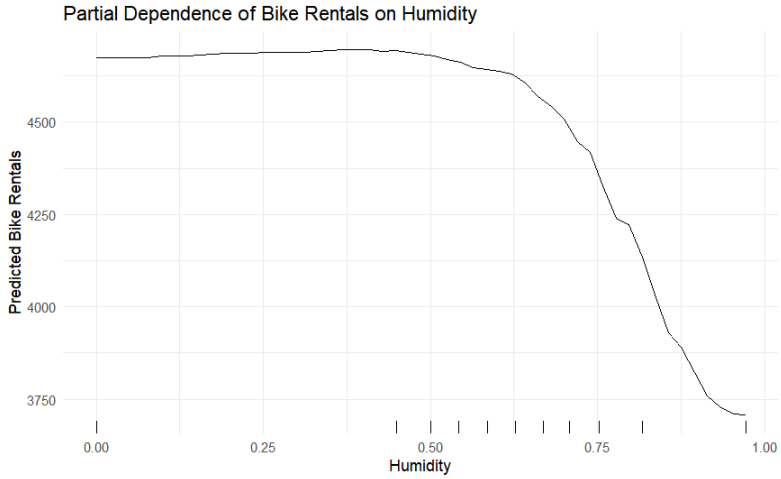

Figura 3: Partial Dependence of Bike Rentals on *Humidity*

Figure 3 shows the partial dependence of predicted rentals on *Humidity*. The overall relationship is negative: as humidity increases, the model's predicted bike rental count generally decreases. In the PDP curve, the left portion is relatively high and flat – for humidity values from very dry conditions up to about 50–60 %, the predicted rentals stay around 4,500–4,700 bikes, and even show a slight, subtle rise in the mid-range. This indicates that under comfortable or only mildly humid conditions, humidity itself has little adverse effect on ridership (the model does not significantly change its prediction until humidity becomes high). Beyond roughly the 60 % humidity mark, however, the curve drops. At the highest humidity levels (approaching 80–100 % relative humidity), the predicted rental count falls to the lowest values. This steep downward slope for high humidity suggests that the model has learned that very humid days greatly reduce bike rentals. The PDP confirms that humidity has an inhibitory influence on bike usage.

2.4 Windspeed

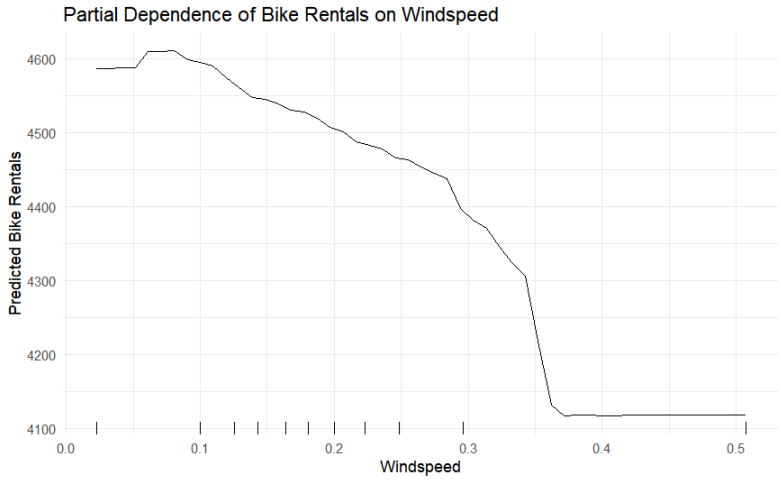

Figura 4: Partial Dependence of Bike Rentals on *Windspeed*

The partial dependence plot for *Windspeed* (Figure 4) demonstrates a clear **negative correlation between wind speed and predicted bike rentals. On the far left of

the plot (near zero wind), the predicted rental count is highest. There is even a tiny uptick at very low wind speeds, suggesting that perfectly calm conditions are not significantly different from a light breeze in their effect – if anything, a small breeze might be neutral or marginally beneficial. However, once wind speed increases beyond this minimal range, the PDP line trends steadily downward. Moderate winds lead to noticeably fewer predicted rentals and as wind speed reaches the higher end, the predicted rental count falls sharply. The curve bottoms out around 4,100 rentals and then flattens, indicating that beyond a certain high wind threshold, additional wind has little further effect because the predicted rentals are already at a minimum. In short, the model infers that windy conditions discourage cycling. Calm or light wind days are favorable for bike rentals, whereas strong winds significantly reduce the expected number of rentals.

3 2D PARTIAL DEPENDENCE PLOT FOR TEMPERATURE AND HUMIDITY

A partial dependence plot (PDP) illustrates how predicted bike rentals vary with `temp` (temperature) and `hum` (humidity), marginalizing over the effects of all other features. This two-dimensional PDP is visualized as a heatmap, which helps reveal any interaction between temperature and humidity in their influence on the model's bike rental predictions.

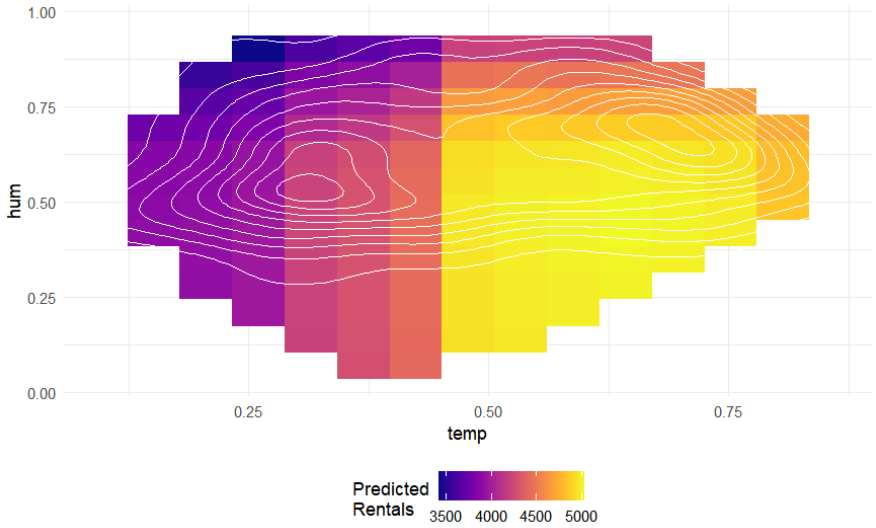

Figura 5: Partial dependence heatmap for bike rental

Figure 5 shows that the highest predicted rentals occur under warm and dry conditions (represented by the warmest colors in the heatmap), indicating that the model expects more bikes to be rented on hot days with low humidity. Conversely, the lowest predicted rentals are found in cool and very humid conditions (shown by the coolest colors), suggesting that rain decreases bike usage. We also observe a clear interaction between temperature and humidity: when temperatures are high (above roughly 20°C), humidity becomes a critical factor—high humidity on a hot day significantly reduces the predicted rentals compared to a hot day with low humidity. At lower temperatures, both factors contribute to keeping rental predictions low; even a dry but cool day yields relatively few rentals, and adding high humidity to a cool day further suppresses the predicted demand.

4 PARTIAL DEPENDENCE PLOT FOR HOUSE PRICES

To explore how individual housing characteristics influence the model's predicted sale prices, we employed Partial Dependence Plots (PDPs) as an interpretability tool. Focusing on four influential features—*Bedrooms*, *Bathrooms*, *Sqft_living*, and *Floors*—each plot isolates the marginal effect of a single variable by averaging out the influence of all others. This approach allows us to visualize the model's behavior in response to changes in specific attributes, highlighting trends such as linear growth, saturation points, or non-linear patterns. In the following sections, we analyze the relationships captured by each PDP to better understand how these features contribute to the predicted property values.

4.1 Bedrooms

Figura 6: Partial Dependence of House Price on *Bedrooms*

The PDP for *Bedrooms* (Figure 6) suggests a non-linear and somewhat plateauing relationship with house price. From one to five bedrooms, predicted prices increase gradually, reflecting the added utility and flexibility that additional rooms provide for family members, guests, or dedicated spaces like home offices. However, beyond five bedrooms, the upward trend diminishes and even slightly declines, implying that further increases may no longer be viewed as added value by typical buyers. This may be due to practical considerations such as maintenance costs or inefficient layouts, or because such properties cater to niche markets with different pricing dynamics. The model thus captures a limited but meaningful positive impact of bedroom count on price, with diminishing returns beyond a practical threshold.

4.2 Bathrooms

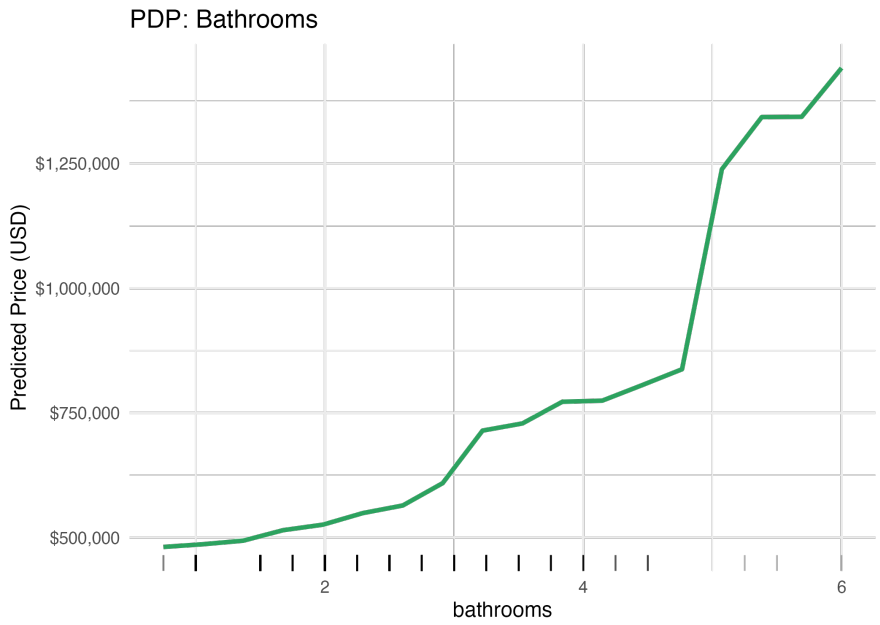

Figura 7: Partial Dependence of House Price on *Bathrooms*

The plot in Figure 7 reveals a clear and sustained positive effect of the number of bathrooms on predicted house prices, particularly between 1 and 3.5 bathrooms. During this range, the model captures a steady increase in value—reflecting buyer preferences for greater comfort, privacy, and convenience. Interestingly, beyond four bathrooms, the curve begins to flatten and even shows a slight decline, suggesting that additional bathrooms beyond this point offer diminishing returns. This plateau may indicate the transition into luxury or atypical properties, where price is shaped more by a combination of high-end features than by bathrooms alone. Overall, the trend illustrates a classic case of diminishing marginal utility: while more bathrooms do raise predicted prices, the added value per unit decreases after a practical threshold is reached.

4.3 Sqft_living

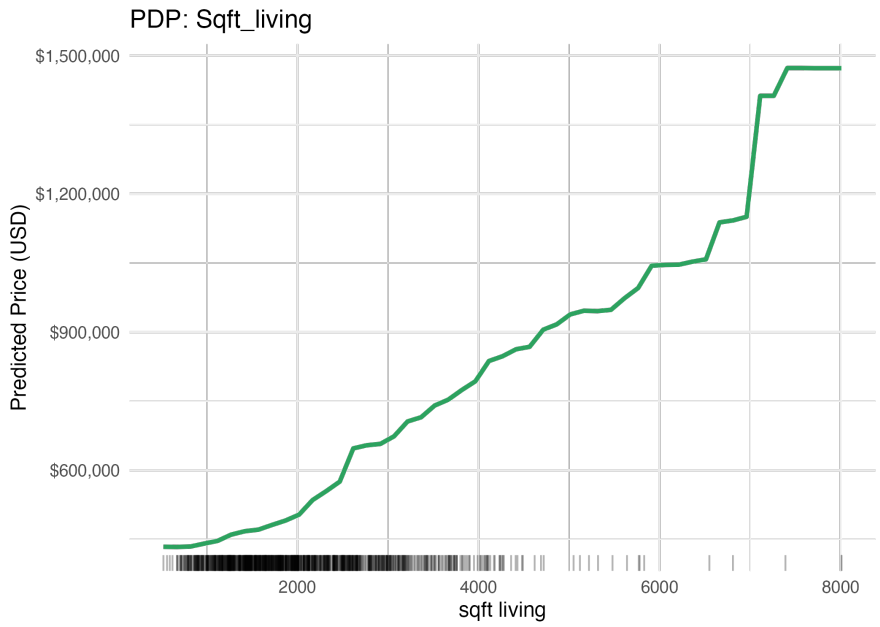

Figura 8: Partial Dependence of House Price on *Sqft_living*

Figure 8 shows a strong, nearly linear relationship between the *Sqft_living* variable and the house price, making it one of the most influential features in the model. As the interior living area increases, predicted prices rise consistently across the observed range. This trend aligns with expectations in real estate: larger homes typically command higher market values. The absence of clear saturation or decline suggests that, within the sample, more living space is uniformly perceived as an asset. The simplicity and regularity of this effect make it especially interpretable and confirm that *Sqft_living* serves as a robust proxy for overall property size and comfort.

4.4 Floors

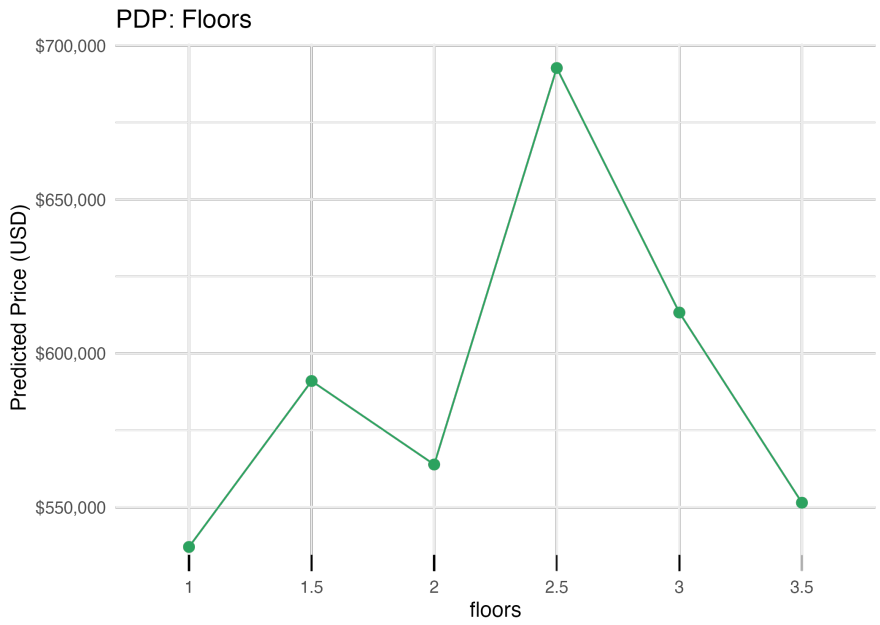

Figura 9: Partial Dependence of House Price on *Floors*

In Figure 9, we can observe that the behavior of the *Floors* variable is characterized by distinct stepwise changes, consistent with its discrete nature. The model predicts a modest increase in price when moving from single-story homes to those with two or two and a half floors, which may reflect improved space utilization without expanding the property's footprint. This added vertical space can enhance living capacity without requiring a larger lot. However, beyond this point, the effect plateaus and even shows a slight decline for properties with more than three floors. Such homes may be less desirable due to reduced accessibility, unconventional layouts, or maintenance concerns. The overall pattern suggests that while adding a second floor may offer a moderate value boost, further increases in height yield limited or even negative returns in the context of typical residential preferences.