# TASK XAI3: Model-agnostic methods

# REPORT

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

Juan Martínez Uradci
Sergio Tortosa Borrell
Ángel Ros Badules

# INDEX



## FIGURE INDEX

# 1. Introduction

In the context of explainable artificial intelligence, understanding how machine learning models make decisions is essential for building trust, ensuring accountability, and enabling informed action. While many modern algorithms, such as random forests or gradient boosting machines, offer strong predictive performance, their internal mechanics are often opaque to end users. Model-agnostic techniques provide a way to interpret these so-called “black-box” models without needing to inspect their inner workings. One such technique is the Partial Dependence Plot (PDP), which estimates the marginal effect of one or more features on the predicted outcome.

This task focuses on applying PDPs to analyze and interpret the output of random forest regressors in two distinct real-world scenarios. The first involves predicting the number of daily bike rentals in a city, using temporal and environmental factors such as temperature, humidity, wind speed, and calendar information. The second addresses the prediction of house prices based on structural features including number of bedrooms and bathrooms, square footage, and number of floors.

The report is structured into three analytical sections. The first section investigates univariate PDPs to evaluate the isolated effect of each predictor on bike rental predictions. The second section builds a bivariate PDP combining temperature and humidity, allowing us to explore more complex interactions between features. Finally, the third section shifts to the real estate domain, applying PDPs to assess how house characteristics influence property value as predicted by the model.

By applying PDPs across these cases, we aim not only to interpret model behavior but also to uncover actionable insights in mobility and housing. This approach highlights how interpretable machine learning methods can bridge the gap between accuracy and transparency, enhancing decision-making in applied domains such as transportation planning and real estate analytics.

# 2. One dimensional Partial Dependence Plot.

In this section, we examine how four specific features (days since 2011, temperature, humidity, and wind speed) influence the predicted number of bike rentals in a Random Forest regression model. To do so, we compute individual Partial Dependence Plots (PDPs) for each of these variables. These plots allow us to isolate the marginal effect of each feature on the model's output, providing an interpretable, model-agnostic view of how user behavior responds to different temporal and meteorological conditions. The analysis aims to identify patterns or thresholds that significantly affect rental volume, and to assess the reliability of the model across the full range of each variable.

## 2.1.1 Days since 2011

The PDP for this variable reveals a clear upward trend in bike rentals over time. In the first 100 days, the expected rentals are below 3500. Between days 100 and 400, predictions remain relatively stable near 3900. A marked increase occurs between days 400 and 600, where predicted rentals exceed 5000. This peak suggests either improved infrastructure or growing user adoption during that period. After day 600, a slight decline is observed, although still above earlier levels. These results indicate a long-term trend of increasing bike usage, potentially influenced by seasonality, policy changes, or public engagement.

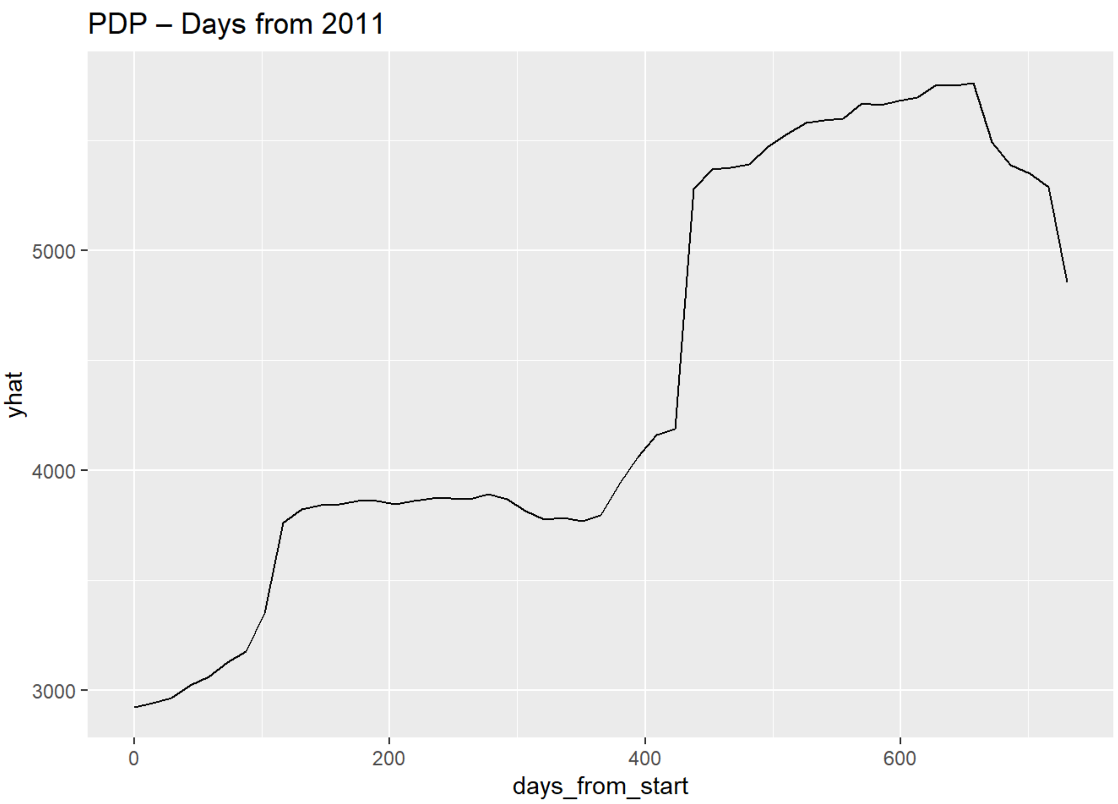


**Figure 1. Partial Dependence Plot – Days from 2011**

### 2.1.2 Temperature

The temperature plot shows a strong positive correlation with the number of rented bikes. Starting from around 5°C, predictions gradually increase, reaching a peak of over 5200 daily rentals between 15°C and 25°C. This range appears to represent optimal weather conditions for cycling. Beyond 25°C, predicted rentals decline slightly, falling to around 4800 at 30°C. This suggests that while warmer weather generally encourages cycling, excessive heat may reduce user comfort and therefore demand. Very low temperature values are not considered reliable due to sparse data.

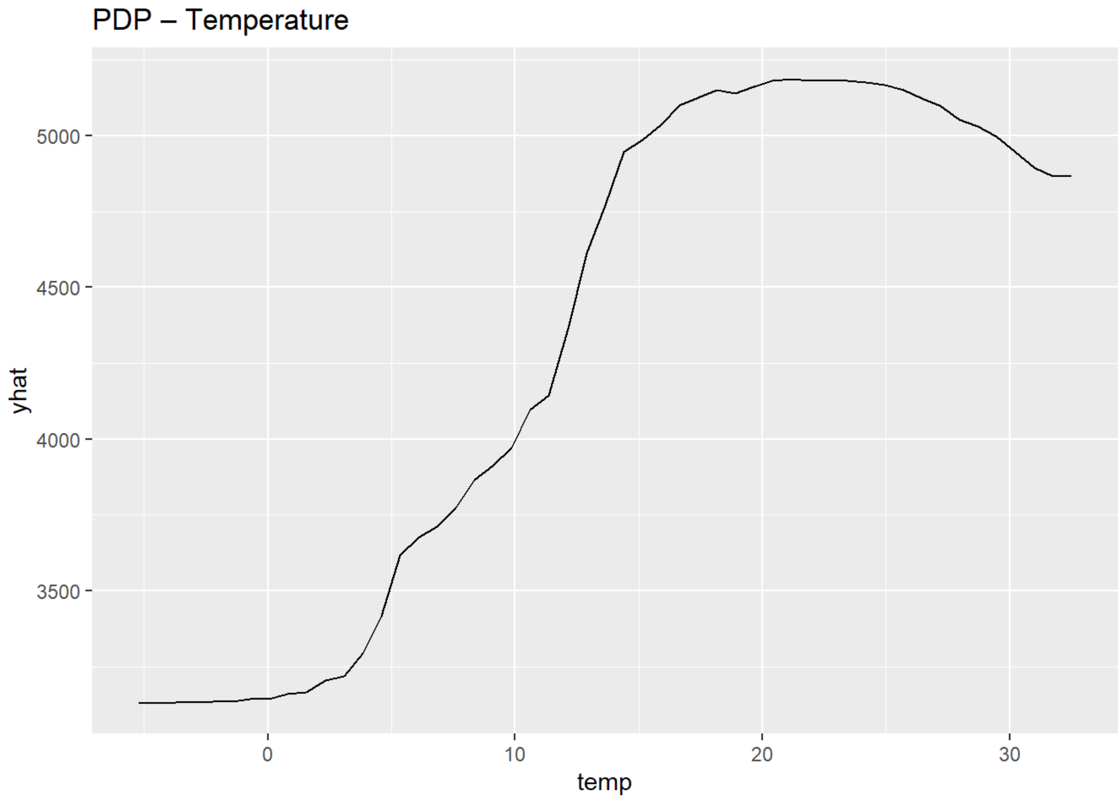


**Figure 2. Partial Dependence Plot – Temperature**

### 2.1.3 Humidity

Humidity presents an inverse relationship with bike rentals. When humidity levels are below 50%, predicted rentals remain high, averaging around 4700. From 50% to 75%, a gradual decrease occurs, followed by a sharper drop beyond 75%. At 100% humidity, the expected number of rentals falls below 3700. This trend highlights how discomfort caused by high humidity negatively impacts outdoor activity. The model's reliability decreases for humidity below 20%, where few samples are available, making that range less conclusive.

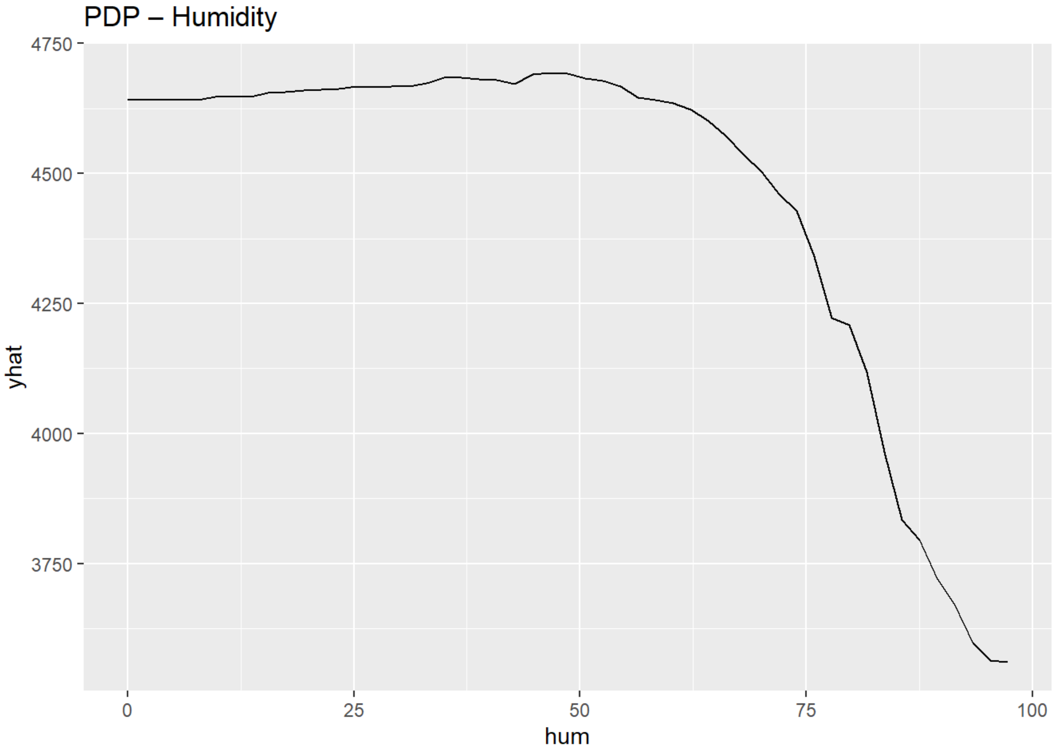


**Figure 3. Partial Dependence Plot – Humidity**

## 2.1.4 Wind Speed

The impact of wind speed on bike usage is also negative. Rentals start around 4600 at wind speeds near zero and decline steadily with increasing wind intensity. By 25 km/h, expected rentals have dropped below 4100. After this point, the curve stabilizes, indicating that extreme wind conditions significantly deter users, but beyond a threshold, additional increases have diminishing marginal effects. Again, data sparsity at the highest wind values should caution against overinterpreting those regions.

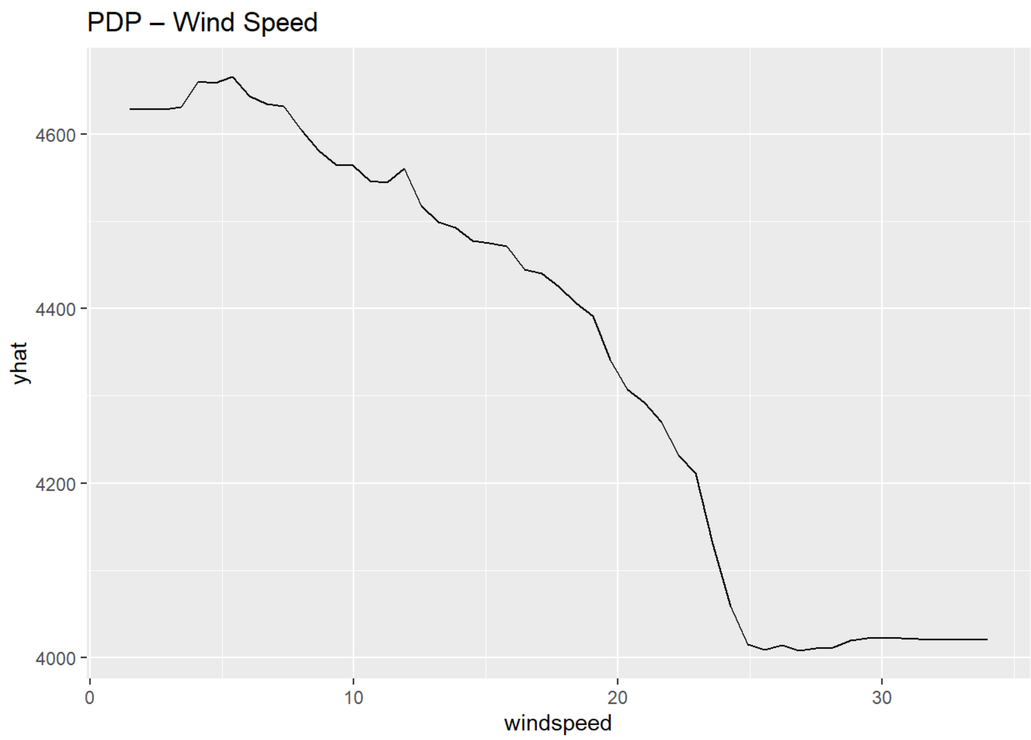


**Figure 4. Partial Dependence Plot – Wind Speed**

This one-dimensional PDP analysis confirms that bike rentals are positively influenced by time progression and temperature, while humidity and wind speed act as deterrents. These insights can support strategic planning for urban mobility services, especially in adapting operations based on seasonal and daily forecasts.

# 3. Bidimensional Partial Dependency Plot

In this section, a two-dimensional Partial Dependence Plot (2D PDP) is generated to analyze how the combined variation of humidity and temperature affects the predicted number of rented bikes. This analysis allows us to explore interaction effects between these two meteorological variables using a model-agnostic approach. To ensure computational efficiency and avoid overfitting to noise, a random subsample of 500 observations from the dataset was used to train the model.

A Random Forest regressor was fitted using only humidity and temperature as predictors. The PDP was computed on a 50x50 grid, and the resulting surface was visualized using geom_tile() with color encoding for predicted counts. Additionally, a contour-based density plot was overlaid to assess the data distribution in the input space.

## 3.1 Visual Interpretation of the PDP Surface

The 2D PDP surface clearly illustrates that lower temperatures, regardless of humidity, consistently correspond to low predicted rental counts. As temperatures increase, especially in the range 15 °C to 25 °C, the model predicts a significantly higher number of rentals. Within this optimal temperature band, the most favorable humidity levels are observed between 50% and 75%, where predicted rentals peak.

On the other hand, high humidity levels (above 80%) lead to a reduction in predicted rentals, even when temperatures remain within the otherwise optimal range. This suggests that high humidity introduces discomfort that discourages outdoor cycling activity.

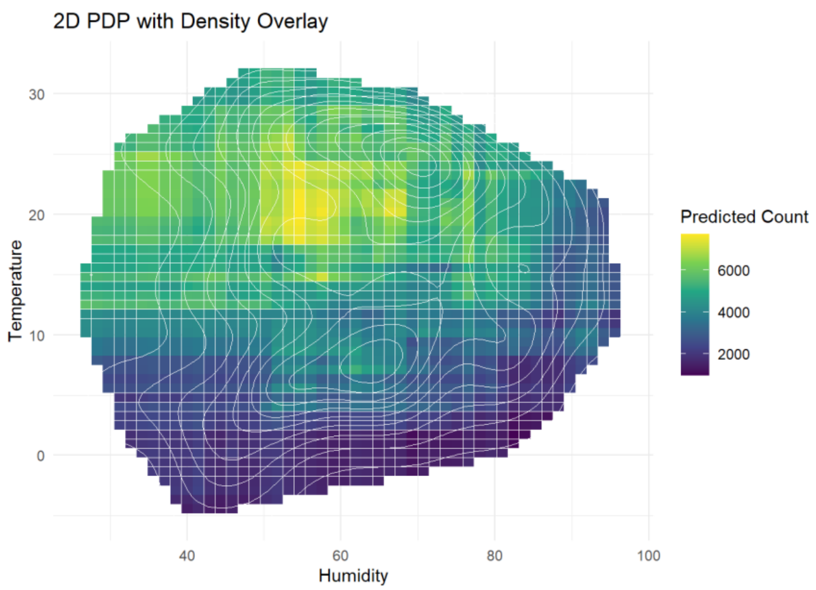


**Figure 5. 2D Partial Dependence Plot with Density Overlay (Humidity vs Temperature)**

## 3.2 Contribution of Temperature and Humidity

From the PDP, it is evident that temperature has a stronger direct impact on rentals than humidity, especially as values increase beyond 15 °C. However, the joint effect of both features produces a non-linear relationship: only under specific combinations (moderate humidity and warm temperatures) does the model output its highest predictions.

At high humidity and low temperature, the lowest values for predicted bike rentals are reached. These combinations represent clearly unfavorable environmental conditions for outdoor activity.

## 3.3 Data Density and Model Confidence

To evaluate the reliability of the PDP predictions, the analysis incorporates a density overlay that reflects the distribution of training data across the humidity–temperature space. Most data points are concentrated between 10 °C and 20 °C, and between 40% and 70% humidity. Predictions within this central zone are therefore supported by a higher number of real observations, making them more reliable.

Conversely, the regions corresponding to extreme humidity and high temperatures, although associated with high predicted rentals, are based on sparse data. Caution should be taken when interpreting predictions in these areas, as the model may be extrapolating from limited information.

## 3.4 Conclusion

This bidimensional analysis provides a deeper understanding of how temperature and humidity jointly influence predicted bike rental demand. It confirms that favorable cycling conditions occur within a specific climate window, and highlights the need to account for data density when interpreting model behavior. The use of 2D PDPs proves valuable in capturing interactions between features that cannot be observed in univariate analysis alone.

# 4. PDP to Explain the Price of a House

This section explores how specific physical characteristics of a house influence its predicted price using a Random Forest model and univariate Partial Dependence Plots (PDPs). The goal is to assess the marginal contribution of each variable to the model's output, isolating the effect of individual features while holding all others constant. The selected features for analysis are: bedrooms, bathrooms, sqft_living (square footage of interior living space), and floors.

A random subset of 1000 instances was used to train the model due to the large size of the original dataset.

## 4.1 Bathrooms

The PDP for the number of bathrooms shows a clear positive trend, with predicted price increasing as the number of bathrooms rises. This trend is especially steep between 3 and 5 bathrooms, where the expected price jumps significantly. However, the curve becomes unstable beyond 5 bathrooms, where a sharp rise is observed. This behavior should be interpreted cautiously due to low sample density in that range, houses with more than 5 bathrooms are rare in the dataset, and their effect may be exaggerated by the model.

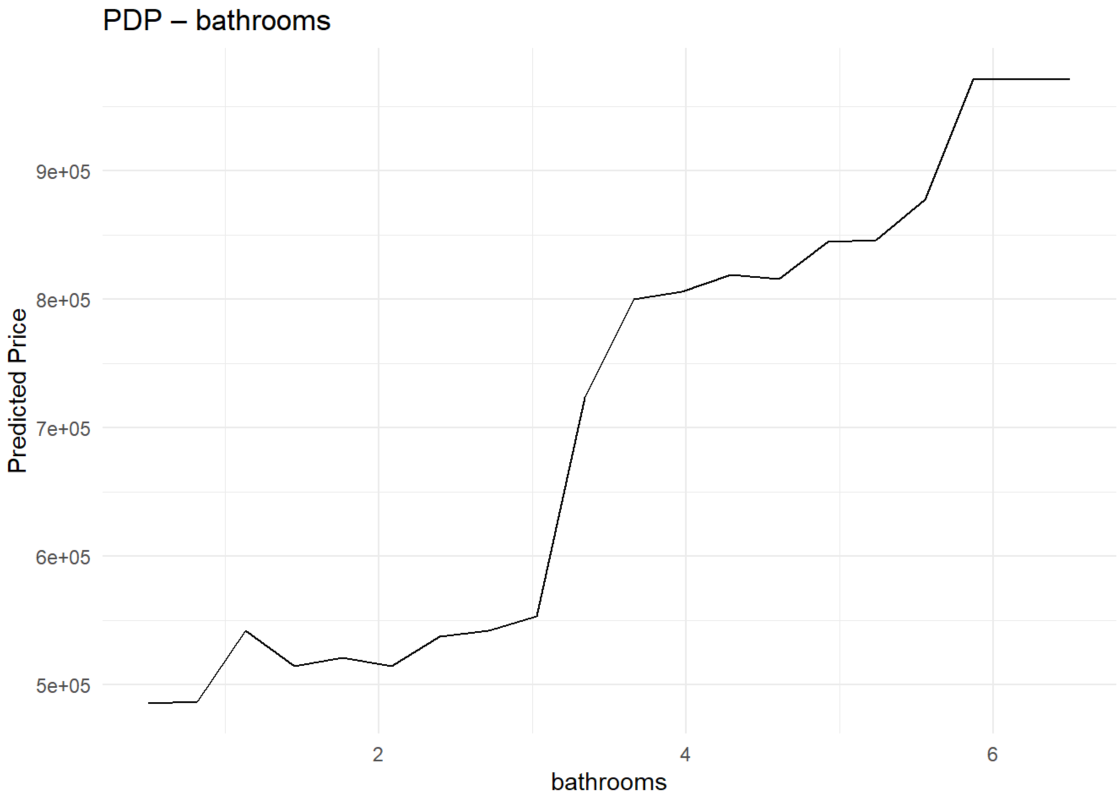


**Figure 6. Partial Dependence Plot – Bathrooms**

## 4.2 Bedrooms

The relationship between number of bedrooms and predicted price appears non-linear. Initially, from 1 to 3 bedrooms, the predicted price declines slightly. Between 3 and 5 bedrooms, the price remains fairly constant, and then slowly increases again from 6 onwards. Surprisingly, more bedrooms do not always equate to a higher price. This can be explained by the fact that the model evaluates bedrooms independently from other relevant variables like living space or bathrooms. For example, a large number of bedrooms in a small or old house may not increase value significantly. Additionally, PDPs assume independence, which may limit interpretability when strong correlations exist.

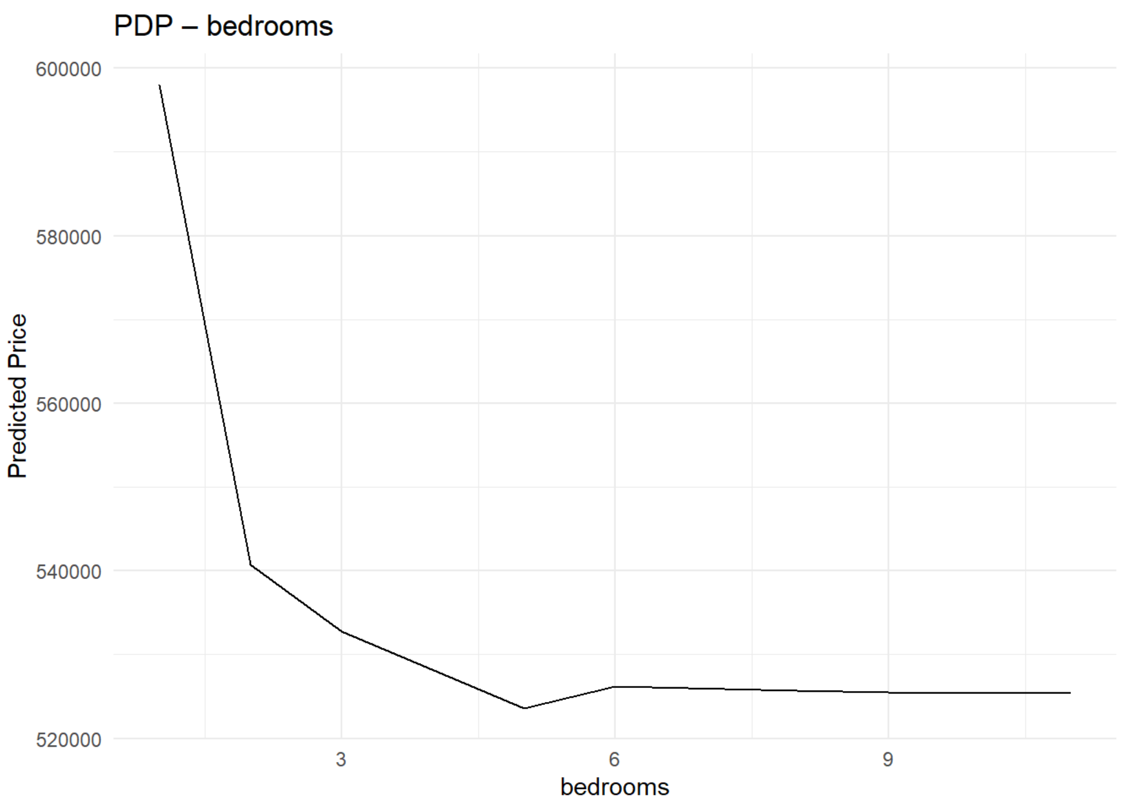


**Figure 7. Partial Dependence Plot – Bedrooms**

## 4.3 Square Footage

As expected, the square footage of interior living space displays a strong positive correlation with predicted price. The price increases steadily from 500 to 4000 sqft, with a particularly steep slope between 2000 and 3500 sqft, suggesting this is a highly sensitive range for market valuation. Beyond 5000 sqft, the curve flattens, indicating a saturation effect where additional space yields diminishing returns. It is important to note that the number of observations in the very large property range is limited, which may reduce reliability.

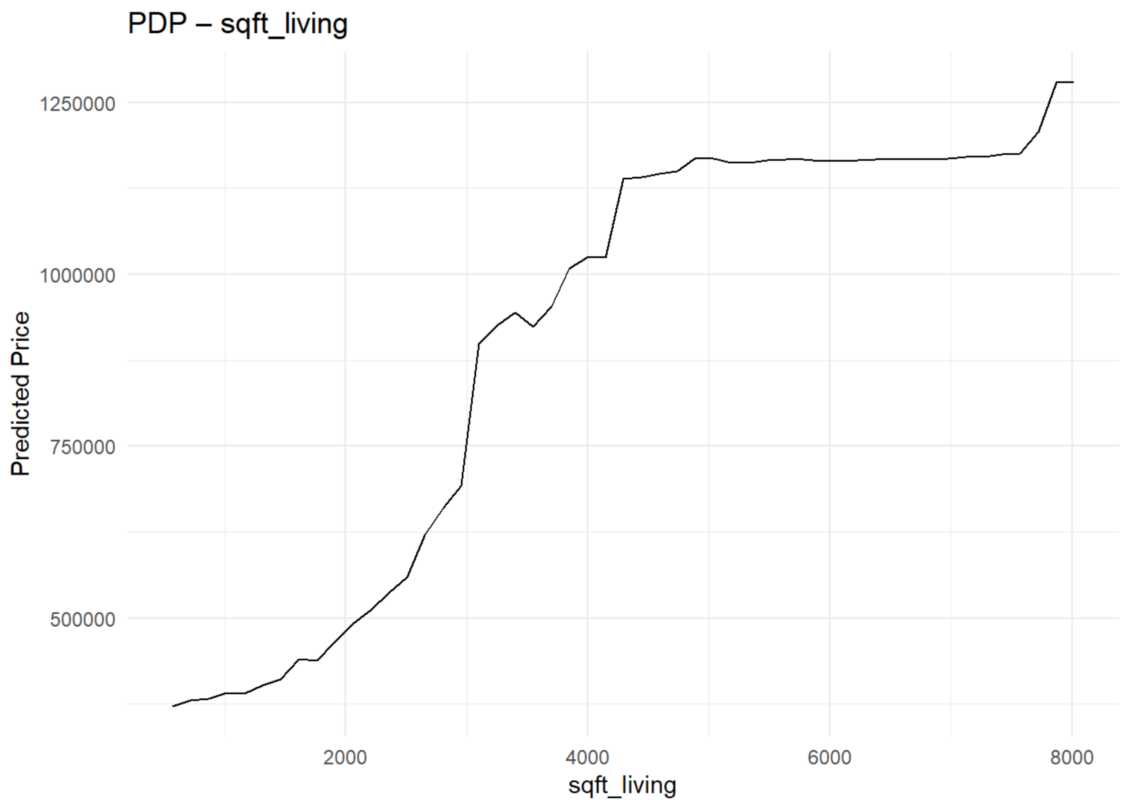


**Figure 8. Partial Dependence Plot – Living Area (sqft_living)**

## 4.4 Floors

The number of floors also impacts predicted price, but the relationship is less consistent. From 1 to 1.5 floors, prices remain stable. A slight dip is observed at 2 floors, followed by an increase beyond that point. The highest predicted prices occur at 3 floors, although the number of such cases is minimal. This irregularity may reflect market preferences or demographic factors, for example, buyers in certain areas may favor one-story homes due to accessibility needs, which influences valuation.

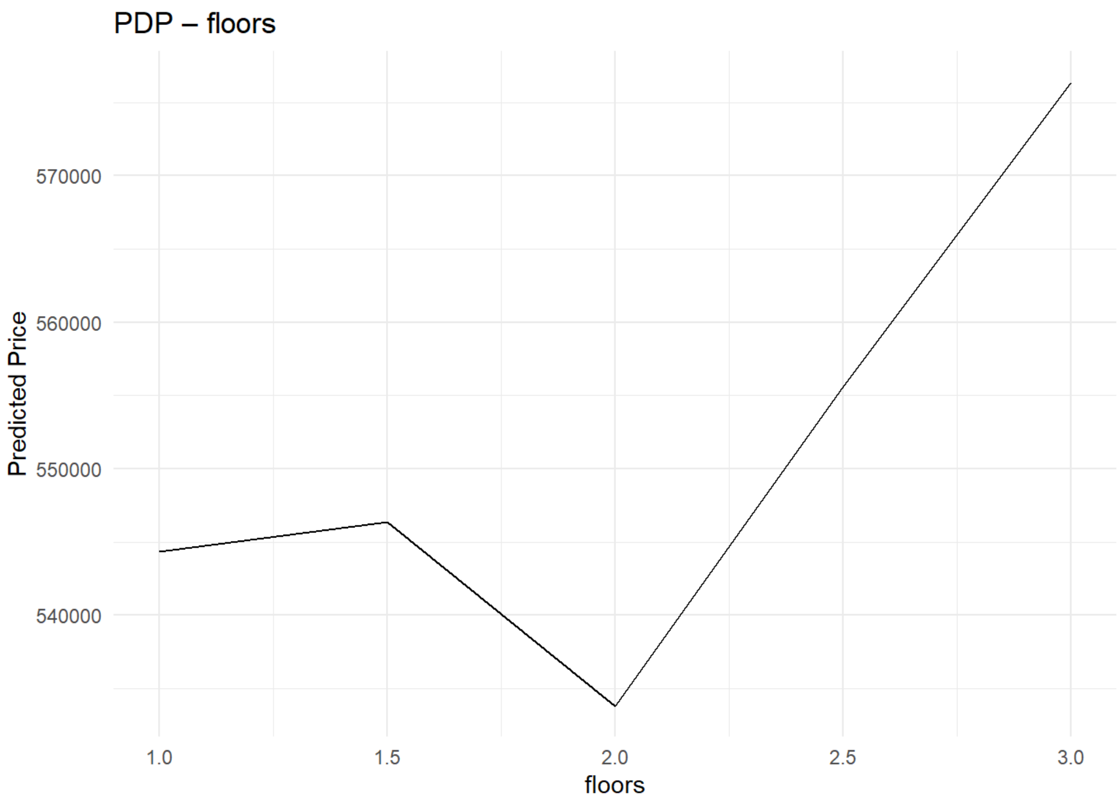


**Figure 9. Partial Dependence Plot – Number of Floors**

## 4.5 Conclusion

The univariate PDP analysis applied to house price prediction reveals clear patterns in how structural attributes affect model output. Among the variables studied, sqft_living demonstrates the strongest and most stable positive influence, indicating that larger interior living spaces are consistently associated with higher predicted prices.

The number of bathrooms also correlates positively with price, though the sharp increases observed beyond five bathrooms should be interpreted cautiously due to limited sample representation. The bedrooms variable exhibits a more complex pattern, with price stabilizing or even slightly decreasing across common bedroom counts, suggesting potential confounding effects with other features like size or location. Lastly, floors show a moderate influence, though the pattern is irregular and may reflect market preferences not captured directly by the model.

Overall, this analysis confirms the value of PDPs in real estate analytics, providing interpretable insights into how individual housing features contribute to model predictions. It also reinforces the importance of data distribution awareness, as predictions in underrepresented regions of the feature space may be less reliable.

# 5. Final Conclusions

This project has illustrated the value of model-agnostic interpretability methods, with a focus on Partial Dependence Plots (PDPs), to analyze the influence of individual and paired features on the predictions of complex machine learning models. Through three different case studies, predicting bike rentals based on weather and time, analyzing feature interactions, and estimating house prices based on structural attributes, PDPs have enabled a clear understanding of how specific inputs affect model outcomes.

In the first section, one-dimensional PDPs were applied to a Random Forest model to investigate how environmental and temporal features influence bike rental demand. The analysis confirmed intuitive trends, such as higher rentals with warmer temperatures and lower rentals with increased humidity or wind speed. Additionally, the steady rise in predicted rentals over time suggests an underlying pattern of increased usage or improved service availability. These results provide valuable insights for urban planners and mobility service providers who aim to align operations with environmental conditions.

The second section used a two-dimensional PDP to evaluate the combined effect of temperature and humidity. This analysis revealed that bike rentals tend to increase under warm and moderately humid conditions. However, it also showed that predictions become less reliable in regions where the training data is sparse. By incorporating a density overlay, the analysis helped identify where model outputs are well-supported and where they may be extrapolating, which is important when making decisions based on limited observations.

In the final section, PDPs were employed to examine how features such as number of bedrooms, number of bathrooms, living area size, and number of floors affect house prices. While the analysis supported common expectations that larger houses and more bathrooms tend to raise property values, it also highlighted non-linear effects and diminishing returns. For instance, adding a third floor or a sixth bedroom does not consistently increase price, especially when such configurations are rare in the dataset. These findings help clarify how different combinations of features influence valuation and can inform both sellers and buyers in the housing market.

Altogether, the use of PDPs in this task has demonstrated how interpretable methods can reveal important patterns in black-box models, support transparent decision-making, and build trust in predictive systems. This reinforces the importance of combining model performance with interpretability when deploying machine learning solutions in real-world scenarios.