Universitat Politècnica de València

Escuela Técnica Superior de Ingeniería Informática

XAI: Model-agnostic methods

Partial Dependency Plot (PDP)

Subject: Evaluation, Deployment and Monitoring of models

Authors:

Víctor Ferrando Chelvi
Marcos Ranchal García
Marcos Valero Navarro

May 16, 2025

Contents

1 Introduction

As machine learning models continue to grow in complexity and are increasingly deployed across diverse domains—from healthcare and finance to transportation and marketing—the need for transparent and interpretable predictions becomes more critical. Complex models, such as ensemble methods and deep neural networks, often act as "black boxes," providing highly accurate results but little insight into how those results are generated. This opacity can hinder trust, raise ethical concerns, and limit the ability of practitioners and stakeholders to make informed decisions based on the model's outputs.

To address these challenges, a variety of interpretability techniques have been developed, aiming to shed light on the inner workings of machine learning models. Among these, the Partial Dependence Plot (PDP) stands out as a widely used and intuitive tool for understanding model behavior. PDPs offer a visual representation of the marginal effect that one or two input features have on the predicted outcome, while averaging out the influence of all other variables in the dataset. This ability to isolate the effect of specific features makes PDPs invaluable for diagnosing model behavior, detecting potential biases, and validating model assumptions.

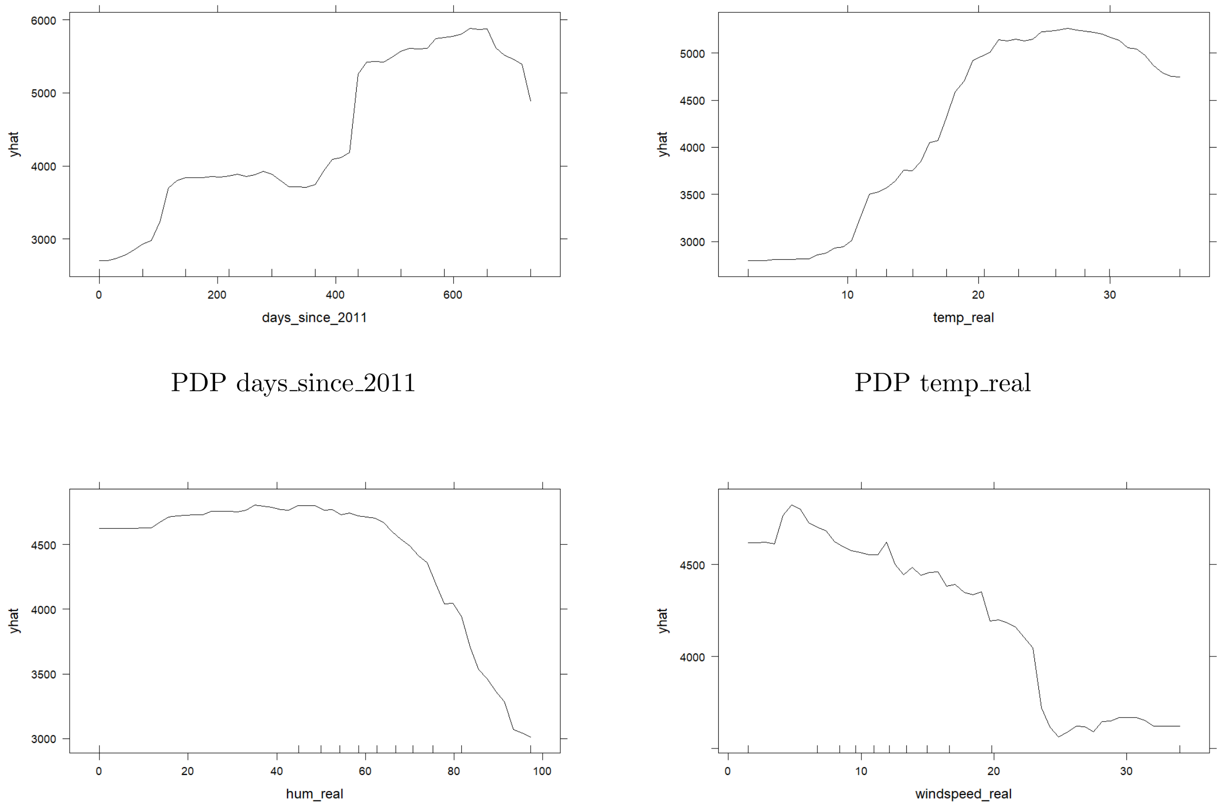
The core principle behind a PDP involves systematically varying the value of the feature(s) of interest across their range and computing the average prediction for each setting, holding other features constant. This process generates a smooth curve or surface that reveals how changes in the selected feature(s) influence the prediction. Crucially, this method allows analysts to explore relationships in models that are highly nonlinear or involve complex interactions—relationships that are difficult to capture with simple linear approximations.

Beyond just visualization, PDPs help uncover important insights about feature influence, including linear trends, nonlinear patterns, threshold effects, and saturation points. For example, PDPs can reveal whether a feature positively or negatively impacts predictions, whether this effect is consistent across its range, or whether there are critical points where the relationship changes direction. Such insights are vital for domain experts who need to ensure that the model's learned patterns align with real-world knowledge and expectations.

Furthermore, PDPs play a crucial role in the growing field of explainable artificial intelligence (XAI), which strives to make AI systems more accountable and accessible. By providing a bridge between complex model outputs and human-understandable explanations, PDPs contribute to enhancing user trust, facilitating regulatory compliance, and enabling effective communication between data scientists, decision-makers, and end-users.

In summary, Partial Dependence Plots represent a fundamental technique for interpreting machine learning models. They empower users to dissect and comprehend the nuanced influences of individual features on predictions, making models not only powerful but also transparent and trustworthy.

2 One dimensional Partial Dependence Plot



PDP days_since_2011

PDP temp_real

PDP hum_real

PDP windspeed_real

Figure 1: PDP's for features

The Partial Dependence Plots (PDPs) for key variables influencing bike rentals provide valuable insights into user behavior under different temporal and weather-related conditions.

To begin with, the PDP for `days_since_2011` reveals a generally increasing trend in the predicted number of rentals over time. This suggests that as the days progress from the beginning of 2011, bike usage tends to rise, likely reflecting factors such as increased adoption of the bike-sharing system, expansion of the service, and seasonal or habitual usage patterns. Although the trend is not perfectly linear—showing occasional plateaus and periods of slower growth—the overall direction remains positive, indicating a growing user base or system demand.

Weather-related variables also show strong and distinct effects. For example, the `temp_real` (temperature) variable exhibits a clear nonlinear relationship with rentals. As temperature increases from low values up to approximately 25–30°C, the predicted rental count rises significantly, indicating that users are more inclined to rent bikes under mild to warm conditions. However, beyond this optimal range, rental activity slightly declines, likely due to discomfort or health risks associated with excessive heat—an observation consistent with general outdoor activity patterns.

Humidity, represented by `hum_real`, has a distinctly negative effect. Rentals remain relatively stable at low to moderate humidity levels but start to decrease sharply once humidity exceeds 60%, with the decline becoming more pronounced at higher values. This suggests that high humidity—often associated with discomfort, perspiration, or the likelihood of rain—deters bike usage considerably.

Similarly, wind speed, measured by `windspeed_real`, shows a consistent negative correlation with the number of rentals. When wind is light (under 10 km/h), the predicted number of rentals is high. However, as wind speed increases, expected rentals drop steadily, likely due to the physical effort and safety concerns linked to cycling in windy conditions. At speeds above 20 km/h, the deterrent effect becomes particularly strong, with rental predictions significantly lower.

Taken together, these plots illustrate how both temporal progression and weather factors such as temperature, humidity, and wind speed shape bike rental behavior. Understanding these dependencies enables more accurate forecasting and more responsive management of bike-sharing systems.

3 Bidimensional Partial Dependency Plot

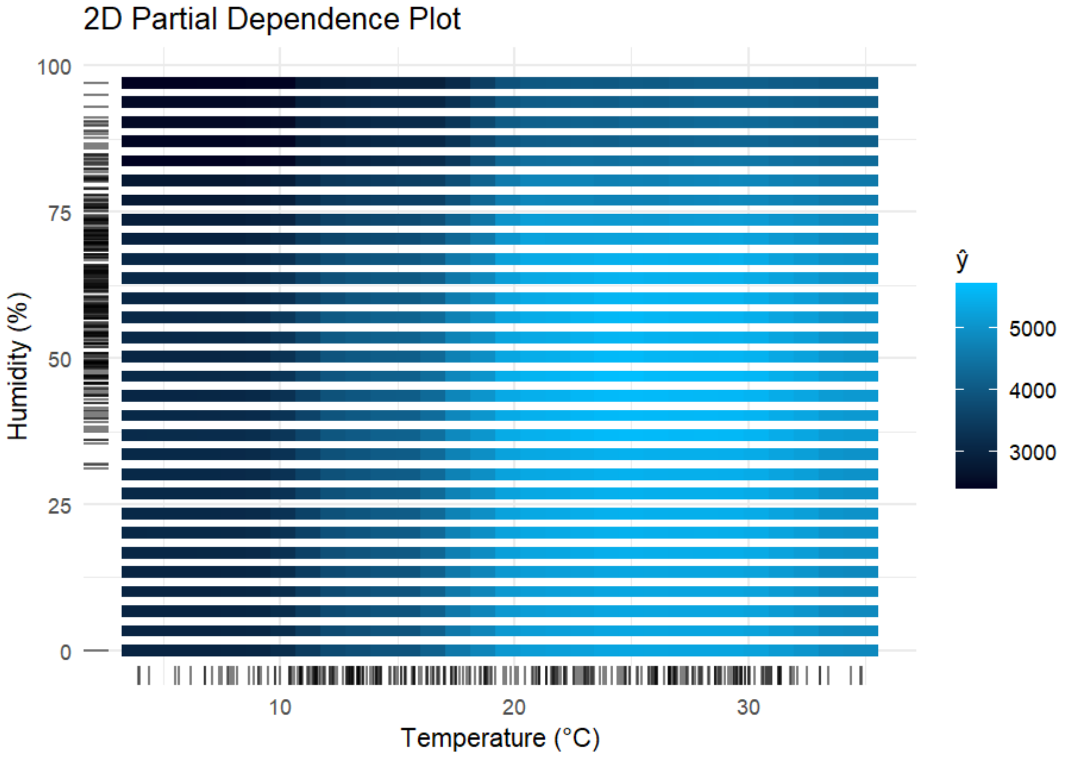


Figure 2: Humidity against Temperature

The 2D Partial Dependence Plot (PDP) displayed above illustrates the combined effect of temperature (°C) and humidity (%) on the predicted number of bike rentals (ŷ), as estimated by a machine learning model. This type of plot is useful for visualizing how a model's predictions respond to changes in two input variables, while marginalizing over the influence of all other variables in the dataset.

As observed, temperature plays a prominent role in shaping the prediction output. The horizontal color gradient, shifting from dark shades (representing lower predicted rentals) to lighter shades (indicating higher predicted rentals) as temperature increases, clearly shows that higher temperatures are associated with increased bike rental predictions. This trend is consistent with real-world behavior, where people are generally more likely to rent bikes in warmer weather.

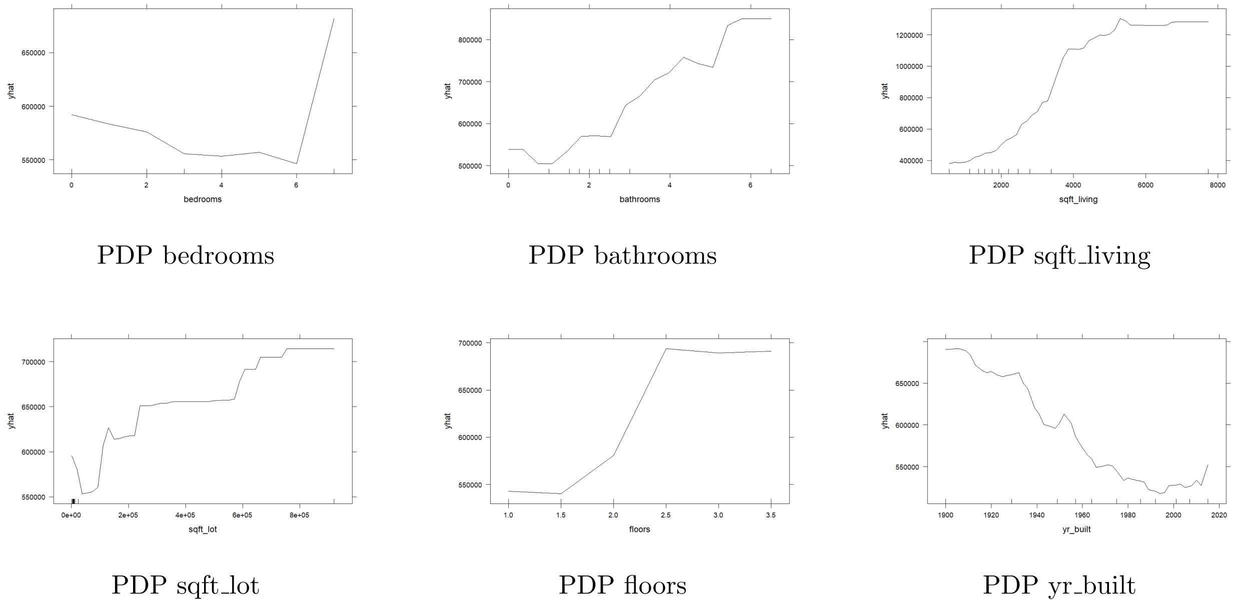
In contrast, humidity appears to exert a relatively minor effect on the model's predictions. The vertical axis, representing humidity, shows minimal color variation from top to bottom, suggesting that changes in humidity alone do not significantly alter the predicted rental count. This may be because humidity either has a limited effect on bike rental decisions or its influence is not strongly captured by the model, potentially due to a non-linear or weak relationship.

The interaction between temperature and humidity also seems limited. While the plot shows distinct horizontal gradients (driven by temperature), there are no corresponding sharp vertical gradients (which would indicate a strong effect from humidity). This indicates that temperature is the dominant variable, and that the combined or interaction effect between temperature and humidity is either weak or negligible within the model.

The rug plots—black tick marks along both axes—represent the distribution of actual data used to train the model. These ticks indicate that the training data is most densely concentrated around 20–30°C and 50–80% humidity. This suggests that the model's predictions are most reliable within this range, and that caution should be used when interpreting predictions in areas with fewer data points (e.g., temperatures below 10°C or humidity above 90%), as the model may be extrapolating in those regions.

In conclusion, the Random Forest model used for predicting bike rentals has effectively captured a strong and intuitive relationship with temperature, showing that bike rentals increase as temperatures rise. In contrast, humidity has a much weaker or nearly neutral influence, and its interaction with temperature appears minimal. These insights can help inform decision-making related to demand forecasting and planning in bike-sharing systems under different weather conditions.

4 PDP to explain the price of a house



PDP bedrooms

PDP bathrooms

PDP sqft_living

PDP sqft_lot

PDP floors

PDP yr_built

Figure 3: PDPs for house features

The Partial Dependence Plot for "bedrooms" reveals a somewhat counterintuitive pattern: as the number of bedrooms increases from one to around six, the predicted house price tends to remain flat or even slightly decrease, with a sharp spike occurring only at seven bedrooms. This suggests that simply adding bedrooms does not necessarily enhance a home's value unless it belongs to a luxury or high-end property segment, which could explain the sudden rise at seven bedrooms. Typically, more bedrooms correlate with larger homes, but this feature alone may not strongly influence price without considering complementary factors such as square footage or location.

In contrast, the plot for "bathrooms" demonstrates a clear and positive relationship with house price. As the number of bathrooms increases from one to six, predicted prices rise steadily, especially beyond 2.5 bathrooms. This likely reflects buyers' appreciation for additional bathrooms due to the convenience they provide, particularly in larger or multi-family households. Homes boasting more than four bathrooms tend to be associated with significantly higher price predictions, indicating a potential shift toward luxury market characteristics in this range.

Similarly, the PDP for “sqft_living” (living area) shows a strong, mostly linear, and positive effect on predicted prices. As living space expands from about 500 to 4000 square feet, prices increase sharply, reflecting the general expectation that larger homes command higher value. Beyond 4000 square feet, however, the curve flattens, suggesting diminishing returns where additional space adds comparatively less to the price.

When examining “sqft_lot” (lot size), the trend becomes less pronounced. While larger lots correspond to somewhat higher predicted prices, the relationship is irregular with several plateaus. This indicates that although lot size can impact value, its effect is less direct and more variable than that of living space, likely due to other moderating factors such as zoning laws, location desirability, or land usability.

The PDP for “floors” indicates that increasing the number of floors from one to approximately 2.5 is associated with a noticeable rise in predicted house price. This may be because multiple floors suggest increased living space or modern architectural design, traits buyers often favor. Yet beyond 2.5 floors, the effect plateaus or slightly declines, implying that additional stories may not always contribute positively to price and might even detract from desirability in certain contexts.

Lastly, the plot for “yr_built” (year built) exhibits a somewhat unexpected downward trend in predicted prices as homes become newer, especially from 1900 through the 2000s. While one might assume newer homes to be more valuable, this trend likely reflects dataset-specific factors such as older homes being located in premium or historic neighborhoods, or possessing architectural value that commands a price premium over newer constructions in less desirable areas.

5 Conclusion

This analysis highlights the value of Partial Dependence Plots (PDPs) as a powerful interpretability tool for understanding complex machine learning models. By isolating the effect of individual features while averaging out others, PDPs provide clear, intuitive insights into how different variables influence model predictions across distinct domains.

In the context of bike rentals, temporal and weather-related variables exhibit meaningful and expected patterns. The increasing trend in rentals over time likely reflects growing adoption and usage, while weather factors such as temperature, humidity, and wind speed strongly shape rental demand. Specifically, moderate temperatures encourage higher bike usage, whereas high humidity and strong winds serve as deterrents. The 2D PDP further confirms temperature as the dominant driver with minimal interaction effects from humidity, underscoring the importance of weather in operational planning for bike-sharing systems.

Regarding housing prices, the PDPs reveal nuanced relationships that underscore the complexity of real estate valuation. While the number of bedrooms alone shows limited influence except at luxury levels, bathrooms and living area exhibit strong positive effects on price. Lot size and number of floors contribute less consistently, suggesting that additional factors like location, zoning, and home design interplay with these features. Interestingly, the decreasing price trend for newer homes likely reflects contextual biases, such as the premium placed on older, historic properties in desirable neighborhoods.

Overall, these findings demonstrate how PDPs can enhance model transparency by aligning learned patterns with domain knowledge and highlighting areas for further investigation. Incorporating such interpretability techniques is crucial for building trust, improving model refinement, and supporting informed decision-making across diverse applications.

References

[1] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 29(5), 1189–1232.

[2] Molnar, C. (2019). *Interpretable Machine Learning.* https://christophm.github.io/interpretable-ml-book/

[3] Friedman, J. H. (2003). *Partial Dependence Plots.* Stanford University Technical Report.

[4] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[5] Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems (NeurIPS), 4765–4774.

[6] Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. Journal of Computational and Graphical Statistics, 24(1), 44–65.