Model-Agnostic Interpretability: Partial Dependence Plots

Marcos Gómez Soler, Joaquín Carrión Gil, and Gonzalo Hurtado Sanhermelando

Universitat Politècnica de València, Valencia, Spain
`mgomsol@etsinf.upv.es, jcargil@etsinf.upv.es, ghursan@etsinf.upv.es`

Abstract. This report explores model-agnostic interpretability methods, focusing on Partial Dependence Plots (PDP). We apply both one-dimensional and two-dimensional PDP to a random forest model predicting bike rentals. Additionally, we extend the methodology to a housing price prediction task using the `kc_house_data.csv` dataset.

1 Introduction

Model-agnostic interpretability techniques allow us to analyze the influence of features on predictions regardless of the underlying algorithm. In this report, we apply Partial Dependence Plots (PDPs) to understand how specific features affect a random forest model's predictions for bike rentals and housing prices.

2 One-Dimensional PDP: Bike Rental Prediction

2.1 Model and Features

In order to predict bike rentals (`cnt`) a random forest regression model was trained. For this machine learning model, features related to the weather and the date were used.

Once the model was trained, some features were analyzed making use of the PDP graphic. Those variables were **instant** (days since 2011), **temp**, **hum** and **windspeed**.

2.2 Partial Dependence Plots

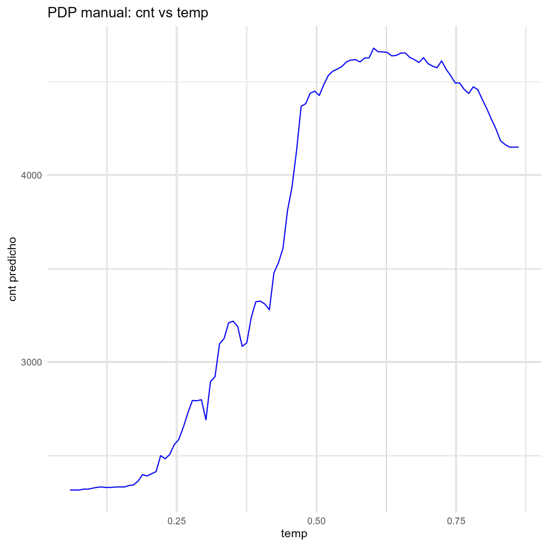

Fig. 1. PDP: Temperature

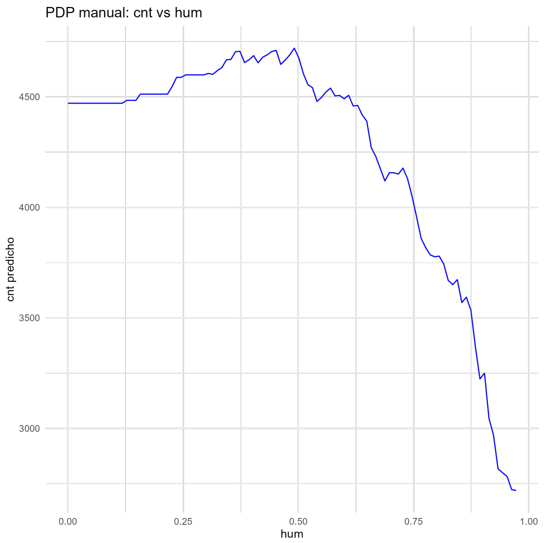

Fig. 2. PDP: Humidity

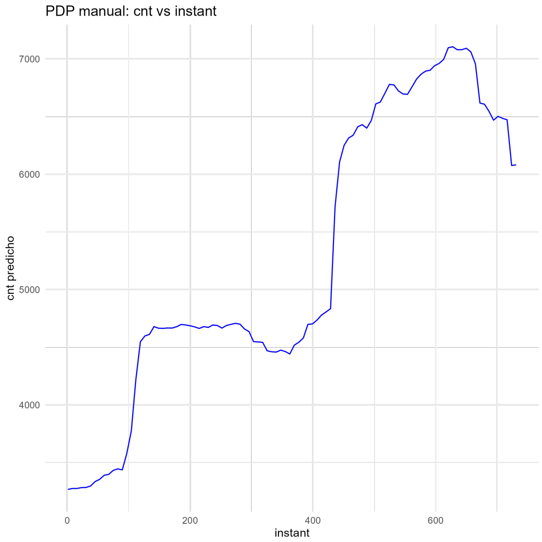

Fig. 3. PDP: Days since 2011

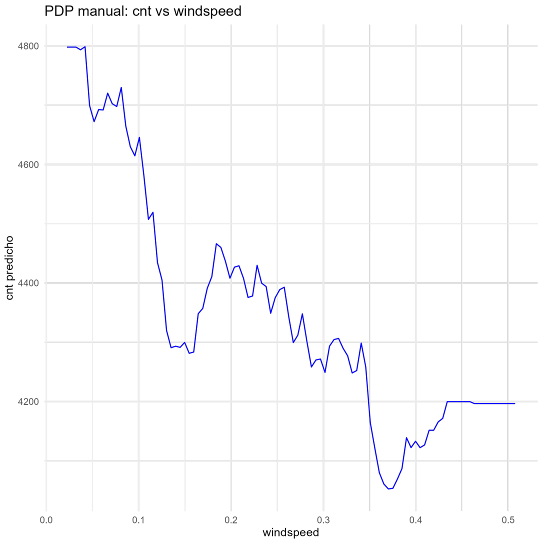

Fig. 4. PDP: Wind Speed

Interpretation

- **Temperature**: As temperature increases, the number of bike rentals also rises, reaching a peak around 0.65–0.70 (normalized scale). Beyond that point, rentals decrease, suggesting that excessively high temperatures may be uncomfortable and discourage bike use.
- **Humidity**: Rentals remain stable under low or moderate humidity. However, after a certain threshold (0.6), the number of rentals drops sharply, indicating that high humidity negatively impacts bike usage.
- **Instant (Days since 2011)**: There is a general upward trend over time, which may reflect growing adoption of the bike rental service or improvements in infrastructure. Toward the end of the period, there's a slight decline, possibly due to seasonality or external factors.
- **Wind Speed**: Overall, higher wind speeds are associated with fewer rentals. A noticeable decline occurs beyond approximately 0.1, suggesting that windy conditions reduce bike use, likely due to discomfort or safety concerns.

3 Two-Dimensional PDP: Temperature and Humidity

3.1 Method

A 2D PDP was generated using a subsample of the dataset to reduce computational cost. The features selected were: **temp** and **hum**.

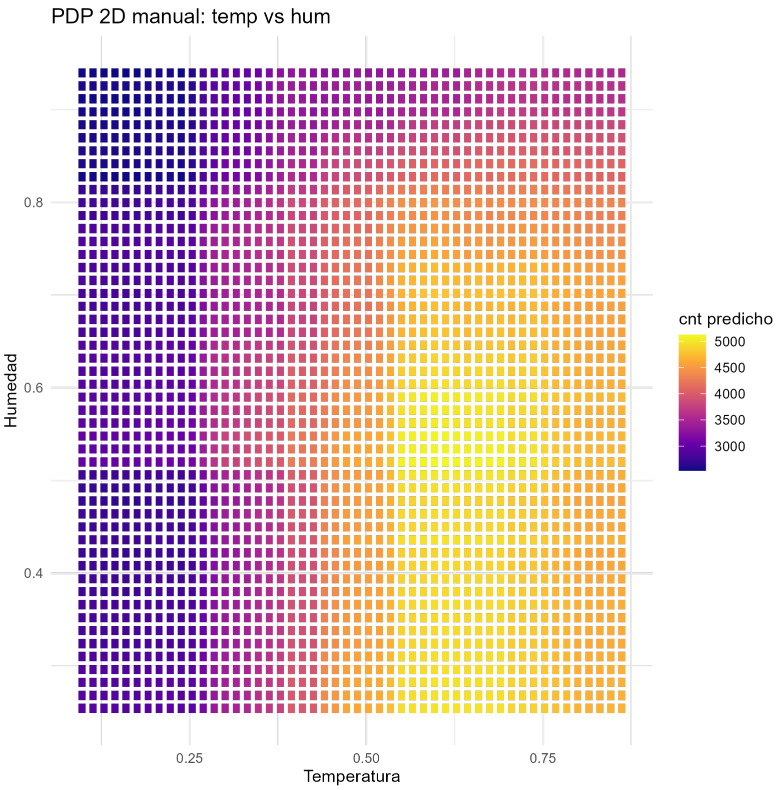

Fig. 5. 2D PDP showing interaction between temperature and humidity.

Interpretation

The 2D Partial Dependence Plot illustrates how the interaction between temperature and humidity affects the predicted number of bike rentals. The color gradient represents predicted values, with brighter colors (yellow) indicating higher predictions and darker tones (purple) indicating lower ones.

We can observe that the highest predicted rental values occur when temperatures are relatively high (above 0.5 on the normalized scale) and humidity is moderate (around 0.4 to 0.6). In contrast, low temperatures combined with high humidity result in the lowest predicted values, suggesting unfavorable conditions for bike use. This visualization highlights a synergistic relationship between temperature and humidity: while moderate heat encourages bike rentals, this positive effect is diminished when humidity is too high. The plot reveals not only the individual effects of these features but also how their interaction shapes model behavior.

4 PDP on Housing Prices

4.1 Model and Features

A random forest regression model was trained using the following features: **bedrooms**, **bathrooms**, **sqft_living**, **sqft_lot**, **floors** and **yr_built**.

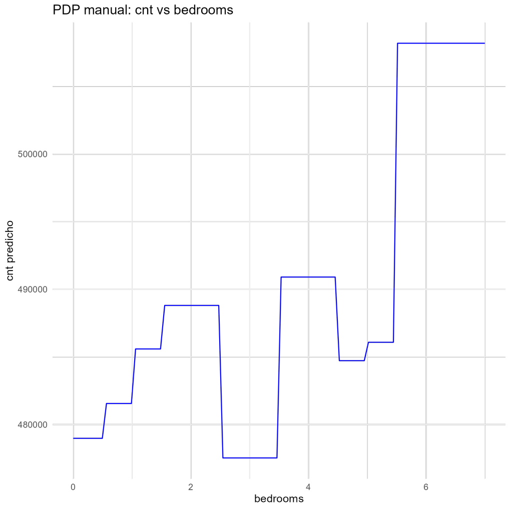

Fig. 6. PDP for bedrooms.

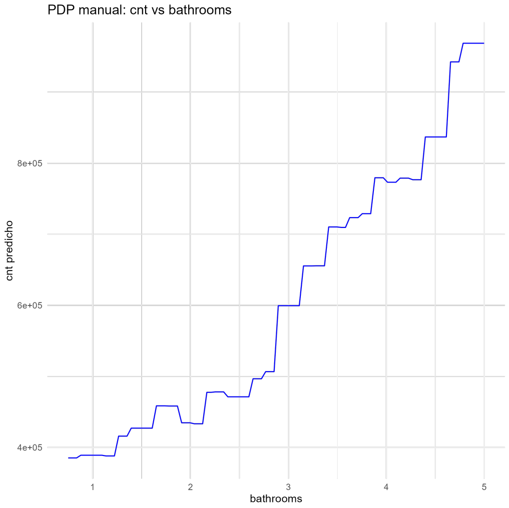

Fig. 7. PDP for bathrooms.

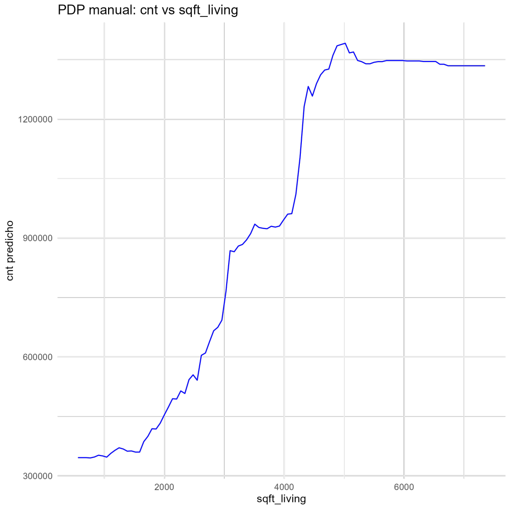

Fig. 8. PDP for square footage (living).

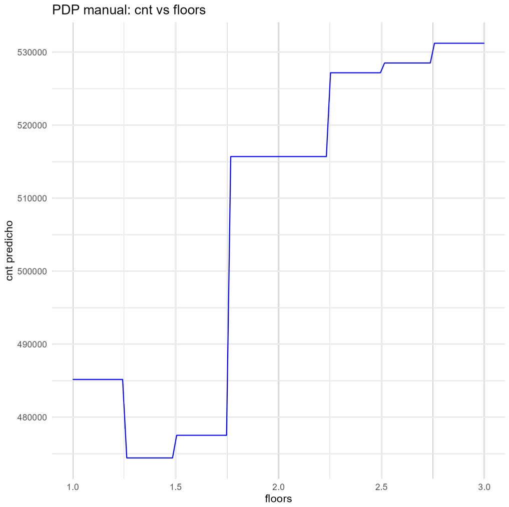

Fig. 9. PDP for number of floors.

Interpretation

- **Bedrooms**: The effect of the number of bedrooms on the predicted price is not linear. Prices fluctuate slightly as bedroom count increases, with a small peak at 3 bedrooms. Beyond this, the relationship appears unstable until a sharp increase is seen at 6 or more bedrooms, suggesting that only very large homes significantly increase value through bedroom count alone.
- **Bathrooms**: There is a clear positive relationship between the number of bathrooms and house price. As the number of bathrooms increases, the predicted price also rises steadily, reflecting the consistent added value that additional bathrooms contribute to property valuation.
- **Square Footage (Living Area)**: The most influential variable appears to be the living area. Prices increase sharply between 1000 and 4000 square feet, peaking around 4000–5000 sqft. Beyond that, the price levels off or slightly declines, indicating a saturation point where extra space yields diminishing returns in valuation.
- **Floors**: The number of floors has a step-wise effect on the predicted price. One-story homes are valued lower, while two-story and especially three-story homes have higher predicted prices. However, the effect is not continuous, suggesting thresholds in buyer or market preferences.

5 Conclusion

This report has demonstrated the effectiveness of Partial Dependence Plots (PDP) as a means to interpret complex machine learning models in a transparent and structured way. By applying PDPs to two distinct regression tasks—bike rental prediction and housing price estimation—we have been able to uncover meaningful patterns in how specific input features influence the model's output.

In the case of bike rentals, one-dimensional PDPs revealed clear non-linear relationships between the number of rentals and weather-related variables such as temperature, humidity, and wind speed. These insights align well with real-world expectations—for instance, that extremely high temperatures or high humidity levels tend to discourage bike usage, while moderate weather conditions promote it. The temporal variable also suggested an upward trend in usage, possibly reflecting growth in the system's popularity or external developments such as infrastructure improvements.

The two-dimensional PDP for temperature and humidity provided further depth by visualizing interactions between these variables. This allowed us to observe how combined conditions affect model predictions in a more realistic way, beyond the additive effects captured by single-variable plots.

In the housing price prediction task, PDPs enabled us to assess the relative influence of structural features such as the number of bedrooms and bathrooms, square footage, and the number of floors. Some features showed strong and consistent effects, while others exhibited threshold behaviors, indicating that property value may increase sharply after a certain point rather than linearly.

While PDPs (Partial Dependence Plots) are useful for visualizing the relationships between independent variables and the target variable, they can be misleading if presented in isolation. These plots assume independence between variables, which can lead to incorrect interpretations in the presence of complex interactions or combined effects among features. Additionally, PDPs average effects, potentially obscuring non-linear patterns or specific segmentations within the data. Therefore, it is essential to complement PDPs with other analyses, such as interaction plots or ICE (Individual Conditional Expectation) plots, to gain a more accurate and comprehensive understanding of the factors influencing the model.

Overall, PDPs serve as an accessible and visually intuitive method for understanding the behavior of complex models. Although they rely on certain assumptions, such as partial feature independence, their ability to highlight both individual and interaction effects makes them a valuable addition to the interpretability toolkit. Incorporating such techniques into the modeling process contributes to better model transparency, facilitates stakeholder communication, and supports informed decision-making based on machine learning outputs.