Universidad Politécnica de Valencia

Escuela Técnica Superior de Ingeniería Informática

XA3: Model-agnostic methods

Evaluación y Despliegue de Modelos

Autors:
Marc Hurtado Beneyto
Alejandro Hervás Castillo
Jorge Guerrero Herrero

May 2025

First of all, we include the link to the GitHub repository here, as we were unable to add it to the assignment: Link to the GitHub repository

Introduction

Predictive modeling has become an essential component in strategic decision-making across industries. In this report, we explore the application of machine learning models—specifically, Random Forests—for understanding the impact of key features on model predictions. To interpret these models, we employ Partial Dependence Plots (PDPs), a powerful tool for visualizing the marginal effect of individual or pairs of input variables on a predicted outcome.

This document presents two practical case studies:

- Forecasting urban bike rental demand based on temporal and meteorological conditions.
- Estimating real estate value from structural housing attributes.

Through unidimensional and bidimensional PDPs, we identify influential factors, quantify their effects, and highlight interactions between variables that can inform both operational and investment strategies.

Forecasting Bike Rental Demand from Environmental Variables

In the context of urban mobility planning, anticipating the volume of bike rentals based on exogenous variables can significantly improve operational efficiency. We developed a predictive model using Random Forests trained on daily bike rental data. To understand how specific features influence the model's predictions, we generated Partial Dependence Plots (PDPs) and a 2D surface plot to analyze both individual and joint feature effects.

Individual Feature Effects

Days Since Launch (`instant`):

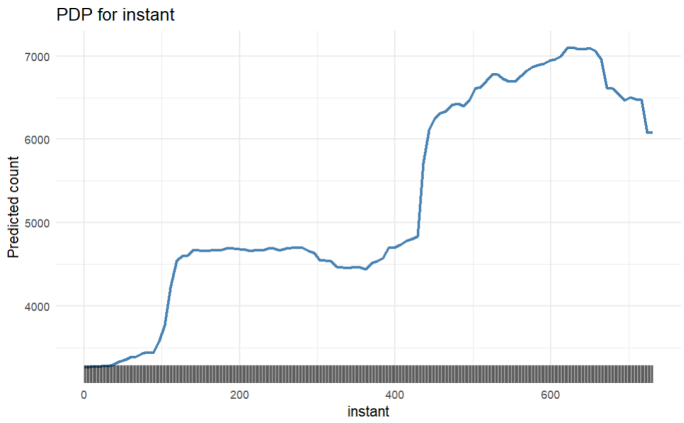

Figure 1: Predicted bike rentals over time since system launch

The plot reveals an increasing trend in predicted rentals over time, particularly after day 400. This suggests strong adoption growth and possibly seasonality patterns or operational improvements.

Temperature:

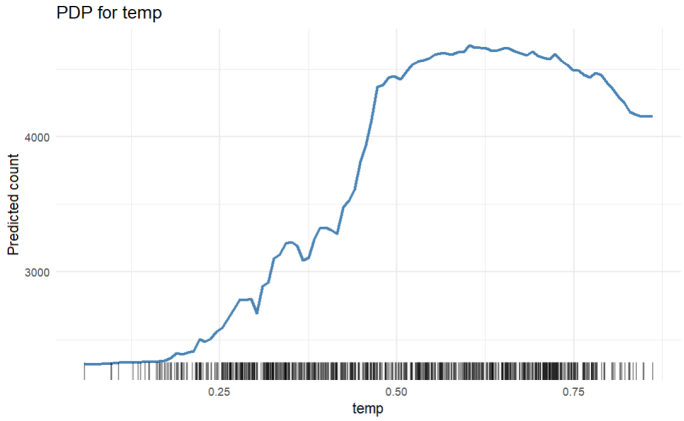

Figure 2: Predicted rentals vs. normalized temperature

Temperature shows a strong positive correlation with predicted rental volume, particularly up to 0.6 (normalized scale). Beyond that point, the effect begins to saturate or decline slightly, which may be due to excessively hot days deterring users.

Humidity:

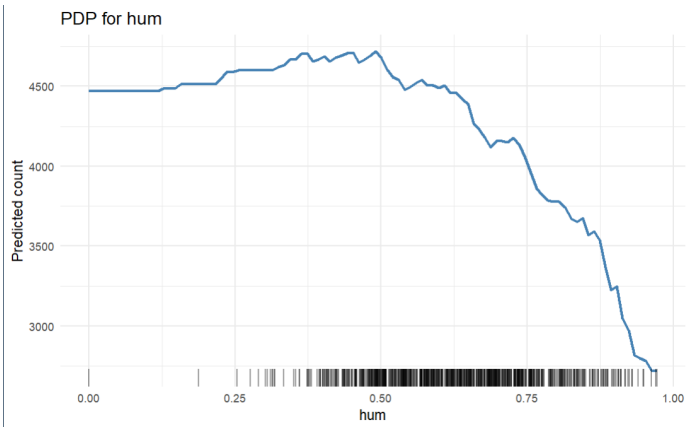

Figure 3: Predicted rentals vs. humidity

Humidity has a clear non-linear effect. While low to mid humidity shows relatively stable predictions, the model forecasts a sharp drop in rental volume beyond a humidity level of 0.7. This suggests discomfort in high-humidity conditions significantly reduces demand.

Wind Speed:

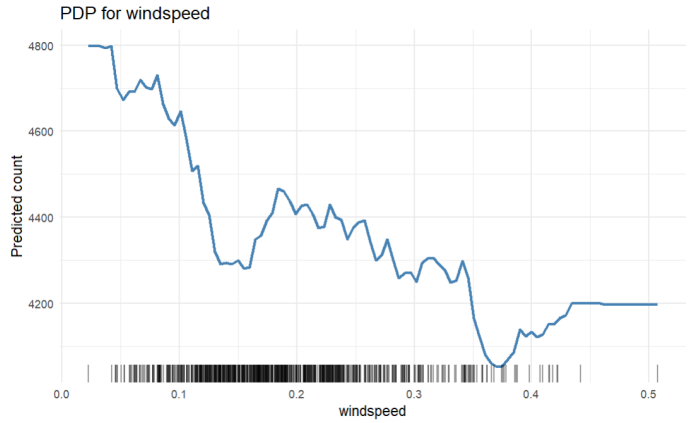

Figure 4: Predicted rentals vs. wind speed

Higher wind speeds are associated with decreased rental predictions. The downward trend is quite steady, especially for wind speeds above 0.15 (normalized). High wind likely reduces perceived safety and comfort.

Interaction Between Temperature and Humidity

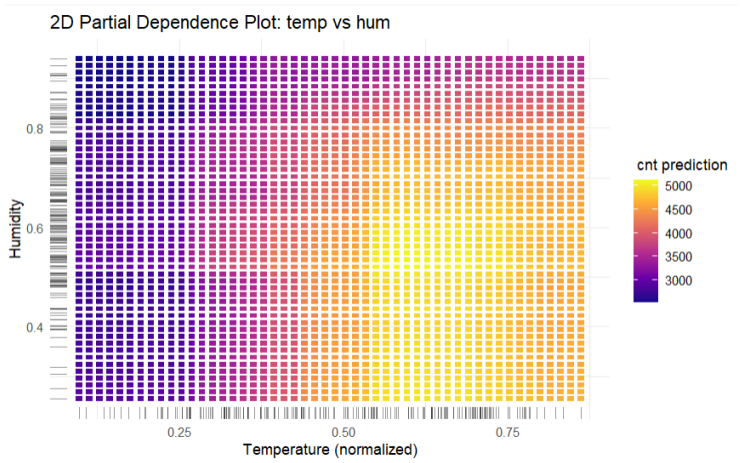

Figure 5: 2D Partial Dependence Plot: temperature vs. humidity

This heatmap provides a joint analysis of how temperature and humidity affect the predicted number of bike rentals. In addition to showing the model's predictions, rug plots have been added to the axes to display the actual data density for both features.

- The most favorable conditions for high demand (yellow zone) occur at high temperatures (around 0.6–0.75, normalized) and mid-level humidity (around 0.5). These regions also show a high density of training data, reinforcing the reliability of the observed patterns.
- High humidity, regardless of temperature, consistently reduces the model's predictions. However, there are fewer real data points in this region (as indicated by the rug plots), so the model's output in this area should be interpreted with caution.
- Conversely, the combination of low temperature and high humidity corresponds to the lowest predicted values (dark purple zone). This region also shows low data density, limiting the confidence we can place in these predictions.

These insights are valuable not only for understanding the model's behavior, but also for identifying regions in the input space with low data support, where prediction uncertainty is higher. This type of analysis is useful for anticipating drops in demand and optimizing resource allocation based on weather forecasts.

Estimating Real Estate Value from Property Attributes

To support better property valuation models, we trained a Random Forest regressor on a subset of the `kc_house_data.csv` dataset. Our model used the features `bedrooms`, `bathrooms`, `sqft_living`, `sqft_lot`, `floors`, and `yr_built`. We then applied Partial Dependence Plots (PDPs) to evaluate the individual effect of each variable on the predicted house price.

Feature Influence on Predicted Price

Number of Bedrooms:

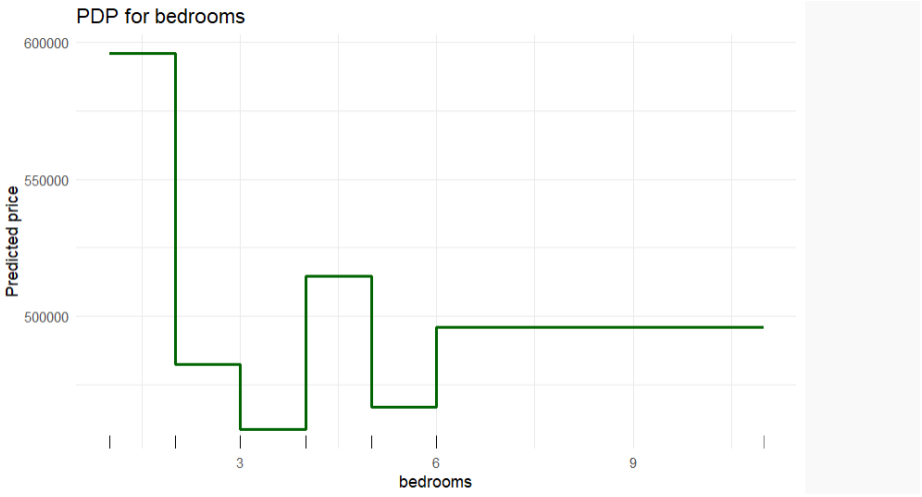

Figure 6: Predicted price vs. number of bedrooms

Contrary to expectations, more bedrooms do not always imply a higher price.Prices decrease sharply from 1 to 3 bedrooms and fluctuate slightly for higher values. The flat trend beyond 6 bedrooms suggests a limited or saturated influence. Rug marks reveal that extreme bedroom counts are rare, which may reduce prediction reliability in those regions.

Number of Bathrooms:

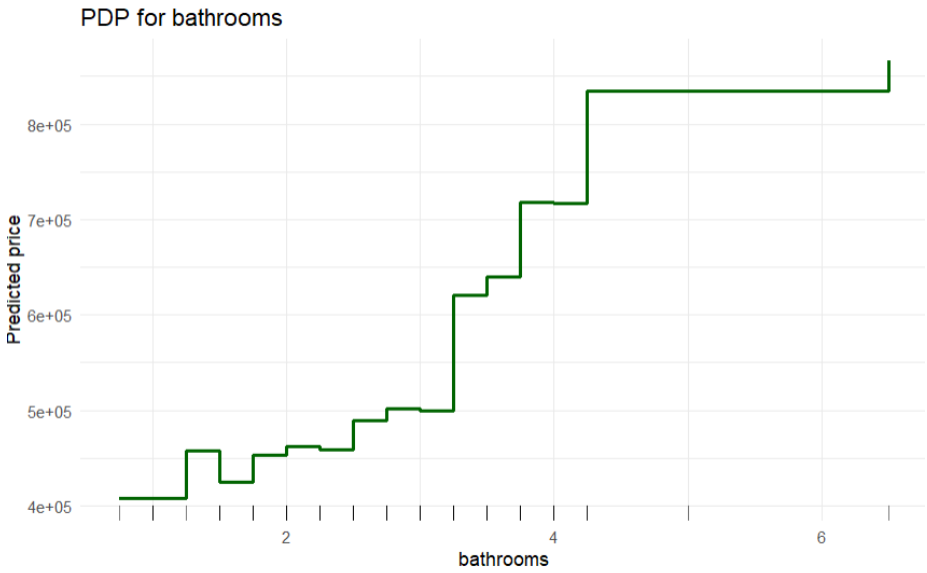

Figure 7: Predicted price vs. number of bathrooms

The plot reveals a clear upward trend, with price predictions increasing as the number of bathrooms grows. The step-like shape reflects the discrete nature of the variable and how the Random Forest model captures non-linear jumps in value. Rug marks along the x-axis show the distribution of real data points, indicating that the predictions are well supported across most bathroom levels.

Living Area (sqft):

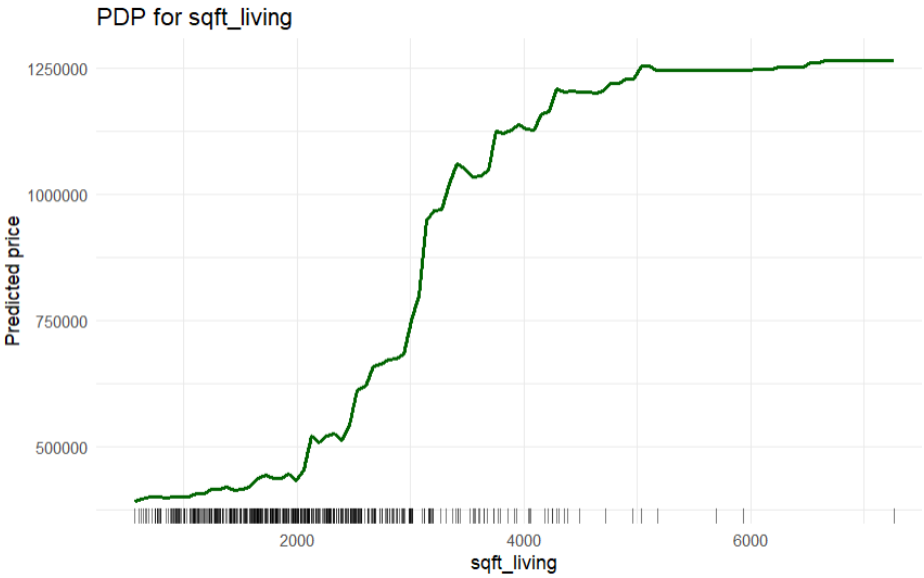

Figure 8: Predicted price vs. living space

This PDP shows a strong positive relationship between living area size and predicted house price. The curve increases steadily, especially between 1500 and 4000 square feet, indicating that larger homes are associated with higher prices. The growth flattens beyond 5000 sqft, suggesting diminishing returns. Rug marks indicate that most training data is concentrated below 4000 sqft, making predictions beyond that range less supported.

Number of Floors:

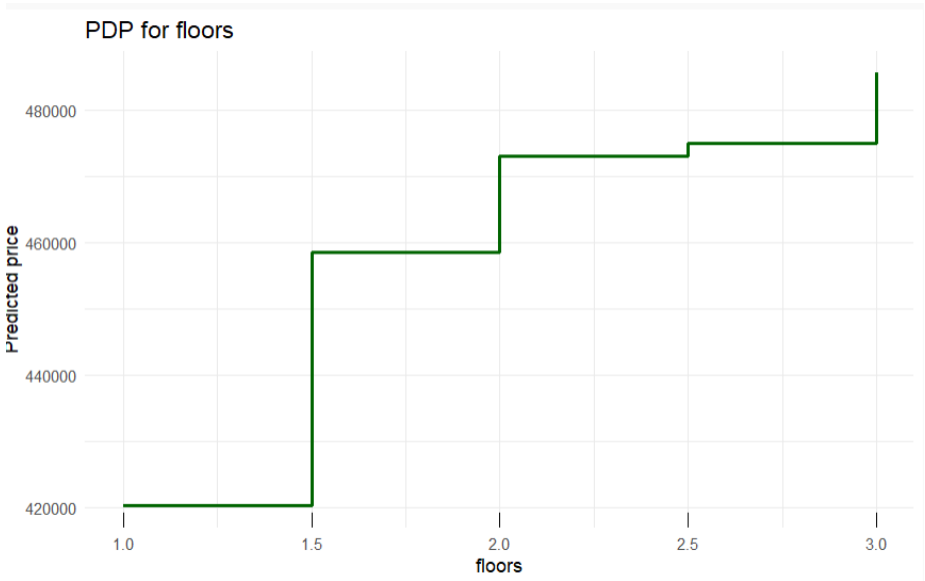

Figure 9: Predicted price vs. number of floors

The relationship is stepwise and slightly positive: properties with more floors tend to have higher predicted values, especially from 1.5 to 2 floors. The overall effect is moderate, suggesting that floor count contributes to price, but with less influence than other features like square footage.

Conclusions

This report demonstrates the value of Partial Dependence Plots as a means to interpret complex machine learning models applied to real-world data. Across two distinct domains—urban mobility and real estate—we identified key features that influence model predictions and derived insights relevant to decision-making.

In the case of bike rental demand, we found that temperature and time were strong positive drivers of usage, while high humidity and wind speed had a dampening effect. These results can guide dynamic resource allocation, infrastructure deployment, and seasonal strategies for shared mobility systems.

For housing price prediction, we observed that living area (sqft) and number of bathrooms were major contributors to price variation. Meanwhile, the number of bedrooms had a weaker and non-linear impact, and the number of floors contributed in discrete steps. These insights inform both property valuation and investment prioritization.

Overall, the PDP methodology bridges the gap between black-box model performance and human interpretability, enhancing trust and transparency in predictive systems.