XAI: Unveiling Drivers of Bike Rentals and House Prices

María Dolores Bonet López, Claudia Piqueras Almario, and Elena Moya Domínguez
Universitat Politècnica de València
{mdbonlop, cpiqalm, emoydom}@etsinf.upv.es

May 15, 2025

Abstract

This study applies interpretable machine learning techniques to two real-world problems: forecasting daily bike rentals using weather and temporal features, and predicting residential house prices in King County based on property attributes. We employ Random Forest regression and Partial Dependence Plots (PDPs) to quantify the marginal effects of key predictors and visualize model behavior. Insights on feature influence and actionable recommendations are provided to support data-driven decision-making.

1 Introduction

Machine learning models have become an integral part of decision-making in numerous industries, from urban mobility to real estate. While highly flexible algorithms such as Random Forests often deliver state-of-the-art predictive performance, their complexity can obscure the reasoning behind individual predictions. For stakeholders—whether city planners seeking to optimize bike-sharing operations or property developers evaluating investment strategies—interpretable insights are as valuable as numerical accuracy.

Partial Dependence Plots (PDPs) offer a model-agnostic approach to interpretability by illustrating the marginal effect of one or more features on the model's prediction. By varying a feature of interest while holding other inputs constant (via averaging), PDPs reveal actionable relationships—soft constraints on feature ranges that increase desired outcomes or avoid adverse scenarios.

In this report, we demonstrate how PDPs can illuminate key drivers in two practical applications:

- Forecasting daily bike rental counts using temporal and weather variables from the `day.csv` dataset.
- Predicting residential house prices in King County based on property characteristics from the `kc_house_data.csv` dataset.

We fit Random Forest regression models to each dataset, compute both one-dimensional and two-dimensional PDPs, and integrate marginal rug plots to visualize the underlying data distribution. The insights derived guide operational tactics for bike rentals and inform strategic investment in real estate development.

2 Data and Preprocessing

Detailed data preparation steps were performed to ensure model quality and interpretability for each case study.

2.1 Bike Rental Data

We utilized the Capital Bikeshare daily dataset (2011–2012) provided in `day.csv`. Key preprocessing steps included:

1. **Date conversion**: Parsing the original `dteday` column as `Date` type and computing a cumulative time variable.

2. *days_since_2011*: Created as the difference (in days) between each date and January 1, 2011, to capture seasonal and annual trends.

3. **Feature inspection and normalization**: Although `temp`, `hum`, and `windspeed` were already normalized to [0,1], we verified their distributions via summary statistics and histograms.

4. **Missing values**: Conducted a completeness check; confirmed no missing entries in predictors or the target (`cnt`).

Descriptive statistics:

- *days_since_2011*: range 0–730 days (two years).
- Median normalized temperature 0.45 (IQR 0.30–0.60).
- Humidity and windspeed exhibited moderate spread without extreme outliers.

2.2 House Price Data

The King County house sales dataset (over 21,000 records) was preprocessed as follows:

1. Exploratory analysis: computed means, medians, and ranges for `price` and all candidate predictors.

2. Outlier removal: excluded fewer than 0.1% of records with implausible values (e.g., zero bedrooms).

3. **Subsampling**: Randomly sampled 2,000 observations (fixed seed) to facilitate efficient PDP computation.

Key attributes after sampling:

- Bedrooms: 1–8 (mode 3).
- Bathrooms: 1.0–7.5 (majority 1.0–2.5).
- Living area (`sqft_living`): median 1,500 sqft (max 10,000).
- Floors: 1–3 (predominantly one- or two-story).
- Year built: 1900–2015, capturing vintage effects.

3 Modeling Approach

We adopted a uniform modeling pipeline for both datasets:

1. **Model selection**: Random Forest regression for nonlinearity and interaction modeling, with built-in OOB error estimation.

2. **Training setup**: `importance=TRUE`, `ntree=500`, default `mtry`; fixed seeds ensured reproducibility.

3. **Performance monitoring**: Confirmed OOB mean squared error stabilized by 300 trees for both models.

4. **PDP computation**: Utilized `pdp::partial()` to generate 1D PDPs (grid size 50 or unique levels) for each feature of interest.

5. **2D PDP (bike data)**: Computed joint partial dependence on temperature and humidity over a 50×50 grid using a 2,000-row subsample.

4 Results

4.1 Bike Rental Forecast

Figure 1 presents the four one-dimensional PDP curves in a single composite graphic (top-left: days since 2011; top-right: temperature; bottom-left: humidity; bottom-right: windspeed), with marginal rug ticks marking the empirical data.

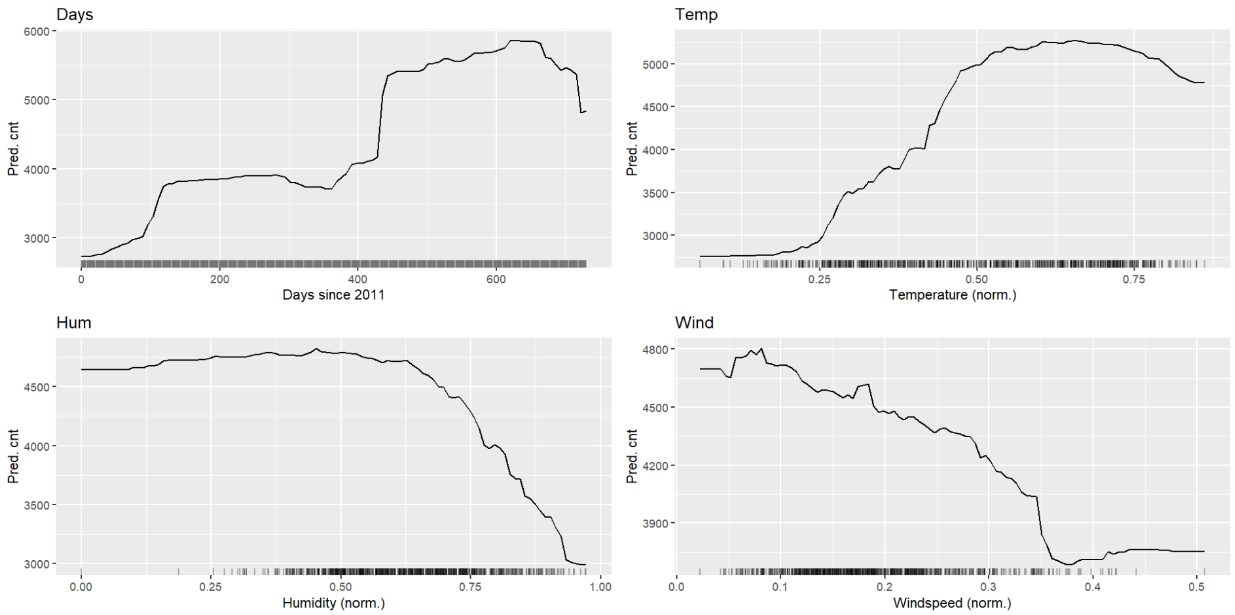

Figure 1: One-dimensional partial dependence plots for days since 2011, temperature, humidity and windspeed (with marginal rugs).

Days since 2011 Predicted rentals climb from about 2,700 in early 2011 to nearly 5,800 by late 2012, reflecting both seasonal peaks and growth of the system.

Temperature Below normalized 0.2, demand stays flat (2,700–2,800). From 0.25 to 0.60 it rises sharply above 5,000, then plateaus or dips slightly past 0.75—indicating extreme heat yields diminishing returns.

Humidity Usage remains high for humidity ¡0.6, then falls from 4,700 down to 3,000 at the highest values, showing that very humid days deter riders.

Windspeed A gentle negative trend shows rentals dropping from 4,700 to 3,900 as windspeed increases to 0.35, then flattening around 3,850—very windy days suppress demand but with a floor effect.

4.2 Bidimensional Partial Dependence Plot (Temperature vs. Humidity)

Figure 2 shows the joint partial dependence of temperature and humidity on predicted daily bike rentals. The heatmap colors indicate the model's predicted counts, and the rug ticks along the bottom and left margins mark the actual data distribution in each dimension.

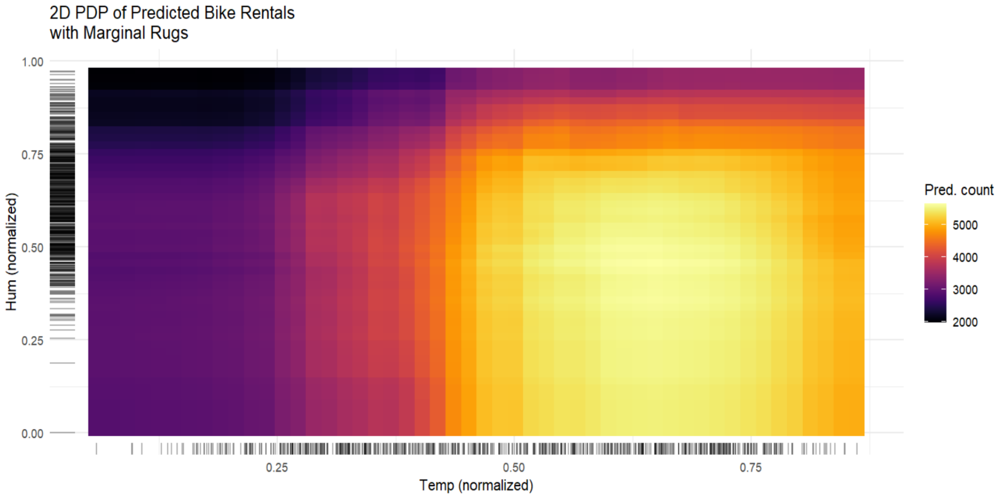

Figure 2: 2D Partial Dependence of predicted bike rentals with marginal rugs (temperature vs. humidity).

The key insights are:

- **High rentals under warm and dry conditions:** The brightest region in the heatmap occurs around normalized temperatures of 0.5 to 0.8 combined with humidity below 0.6, indicating that mild-to-warm, low-humidity days yield the highest rental counts.
- **Diminishing returns at extremes:** Very high humidity (¿0.7) or very low temperatures (¡0.2) are both associated with substantially lower predicted counts (dark purple), showing that both cool, damp and extremely humid days deter riders.
- **Data coverage:** The density of rug ticks confirms most observations fall within temp [0.2–0.8] and hum [0.2–0.7], so the model's predictions are most reliable in that central band.

Overall, the 2D PDP confirms that the Random Forest learned an intuitive interaction: rentals peak when the weather is neither too cold nor too humid, and drop off sharply in the opposite extremes.

4.3 One-Dimensional Partial Dependence for House Prices

Figure 3 presents the model's marginal effects for bedrooms, bathrooms, square-foot living area and number of floors in a single composite plot, with rug ticks indicating the data distribution beneath each curve.

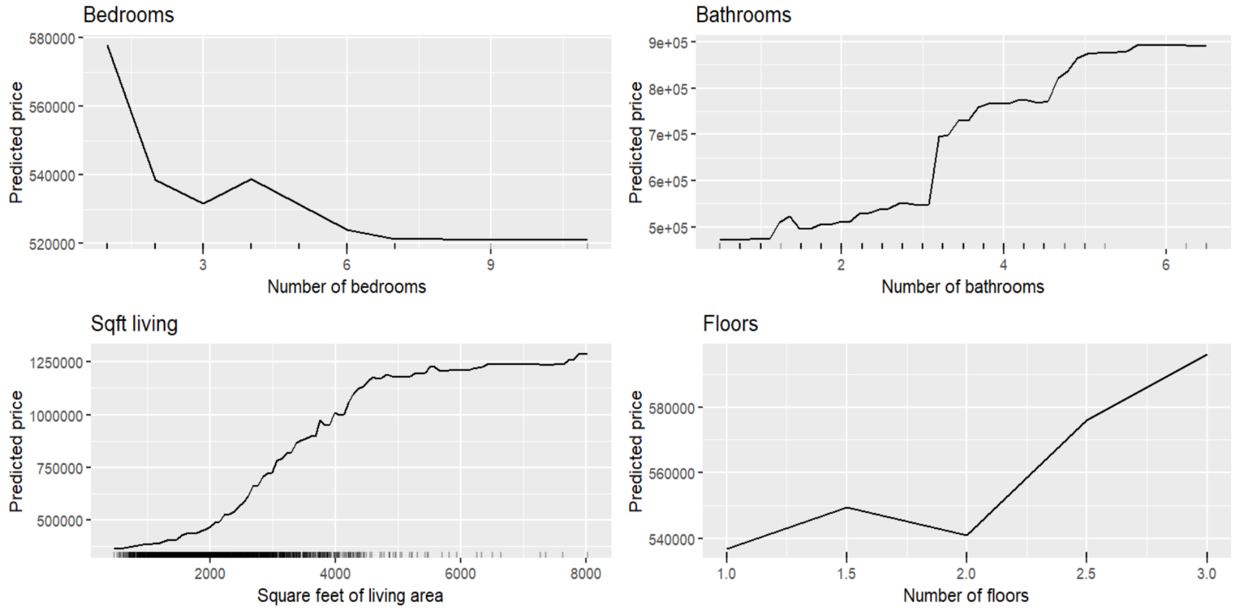

Figure 3: One-dimensional PDPs for number of bedrooms, number of bathrooms, living-area size and number of floors (with marginal rugs).

Bedrooms Predicted price decreases slightly as bedroom count increases from 1 to 3—reflecting that small homes with too few rooms tend to be less expensive—then levels off from 3 onward, indicating diminishing returns for additional bedrooms beyond the modal value.

Bathrooms Each additional bathroom shows an almost linear uplift in predicted price, with a noticeable jump between 2 and 3 bathrooms. This suggests that adding a third bathroom confers particularly high incremental value.

Living Area Size Predicted price rises monotonically with square footage, accelerating between 1,500 and 4,000 sqft before plateauing around 5,000–6,000 sqft. Luxury-sized homes (¿7,000 sqft) command the highest prices but occur less frequently.

Floors Single-story homes have the lowest predicted prices; adding a second floor yields a modest increase, while three-story homes show a further appreciable uplift, indicating vertical expansion also drives value.

These PDPs confirm that our Random Forest model has learned intuitive relationships: property value grows with living space and number of bathrooms, shows diminishing gains from extra bedrooms beyond three, and benefits from additional floors.

5 Conclusions and Recommendations

This report has demonstrated the value of model-agnostic interpretability through Partial Dependence Plots applied to two distinct regression tasks: forecasting daily bike rentals and predicting residential house prices. By fitting Random Forest models and visualizing both one-dimensional and two-dimensional PDPs with marginal rugs, we extracted clear, actionable insights:

- **Bike Rental Forecast:**
  - *Temporal growth & seasonality:* Rental counts rose from approximately 2,700 in early 2011 to nearly 5,800 by late 2012, with pronounced summer peaks.
  - *Temperature effect:* Demand remains flat at low temperatures (normalized ¡0.2), increases sharply between 0.25–0.60, and plateaus beyond 0.75, indicating diminishing returns at extreme heat.
  - *Humidity effect:* Rentals are highest under low-to-moderate humidity (¡0.6) and decline sharply beyond 0.7, showing high moisture deters riders.
  - *Windspeed effect:* A gentle negative relationship reveals fewer rentals as windspeed increases, with a floor effect around normalized 0.35.
  - *Interaction (2D PDP):* Optimal conditions occur at moderate-high temperatures (0.5–0.8) coupled with low-to-moderate humidity (0.2–0.6).
  - *Recommendation:* Leverage these insights for dynamic pricing or targeted promotions on days forecasted to fall within the optimal weather band to maximize ridership and revenue.
- **House Price Prediction:**
  - *Bedrooms:* Price increases up to three bedrooms, then shows diminishing returns, suggesting over-splitting living space yields limited value.
  - *Bathrooms:* Each additional bathroom contributes near-linear price uplift, with a marked jump when adding a third bathroom.
  - *Living-area size:* Predicted price grows monotonically with square footage, accelerating between 1,500–4,000 sqft and tapering off beyond 6,000 sqft.
  - *Floors:* Additional floors confer incremental value—two- and three-story homes command higher prices compared to single-story equivalents.
  - *Recommendation:* Real estate developers should prioritize bathroom additions and quality increases in living space, while recognizing limited ROI from bedrooms beyond three and considering vertical expansions where feasible.
- **Methodological Insights:** Random Forests combined with PDPs provide a robust framework for uncovering nonlinear and interaction effects in complex models. The inclusion of marginal rug plots ensures that interpretations focus on regions with sufficient data density, thereby maintaining reliability.

- **Next Steps:** Future work could extend this analysis by exploring higher-order interactions (e.g., 3D PDPs), investigating Shapley additive explanations (SHAP), or integrating real-time weather forecasts to operationalize bike-share pricing strategies.