Explainable AI: Model-Agnostic Methods Partial Dependence Plots for Model Interpretation

Daniel Reinón García
Carlos Pérez Faus
Miguel Abarca Casares

May 16, 2025

Contents

1 Introduction

1.1 Partial Dependence Plots

Partial Dependence Plots (PDPs) are a powerful model-agnostic technique for visualizing the relationship between a subset of features and the predicted outcome of a machine learning model. PDPs show the marginal effect of one or more features on the predicted outcome, averaging out the effects of all other features. This makes them particularly useful for understanding how specific features influence model predictions, regardless of the complexity of the underlying model.

The mathematical formulation for a partial dependence function for a feature x_S is:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_{C,i}) \tag{1}$$

where x_S represents the feature(s) of interest, x_C represents the other features, $\hat{f}$ is the prediction function, and n is the number of instances in the dataset.

1.2 Objectives

This report aims to:

- Apply one-dimensional PDPs to analyze the influence of various features on bike rental predictions
- Generate and interpret two-dimensional PDPs to understand feature interactions
- Apply PDPs to a house price prediction problem to demonstrate their versatility
- Provide comprehensive interpretations of the relationships learned by the models

2 Methodology

2.1 Datasets

For this study, we utilize two distinct datasets:

- **Bike Rental Dataset (day.csv)**: Contains daily counts of bike rentals along with various features such as temperature, humidity, wind speed, and days since 2011.
- **House Price Dataset (kc_house_data.csv)**: Contains house prices in King County, USA, along with features such as number of bedrooms, bathrooms, square footage, and year built.

2.2 Model Selection

For both datasets, we employ Random Forest models due to their:

- High predictive accuracy
- Ability to capture non-linear relationships

- Robustness to overfitting
- Capability to handle mixed data types

However, Random Forests are inherently complex and difficult to interpret directly, making model-agnostic interpretation methods like PDPs particularly valuable.

2.3 Implementation Tools

We implement our analysis using R with the following key packages:

- `randomForest`: For fitting Random Forest models
- `iml`: For generating Partial Dependence Plots
- `tidyverse`: For data manipulation and visualization
- `grid` and `gridExtra`: For arranging multiple plots

2.4 Methodological Process

Our methodological approach follows these fundamental steps:

1. **Data preparation**: We load the datasets and perform necessary transformations. For the house price dataset and for the two-dimensional analysis, we perform random sampling to reduce computational load.
2. **Model training**: We fit Random Forest models using selected features. For the bike dataset, we use days since 2011, temperature, humidity, and wind speed as predictors. For the house dataset, we use bedrooms, bathrooms, living area, lot size, floors, and year built.
3. **Model encapsulation**: We use the `Predictor` class from the `iml` package to encapsulate our trained models, facilitating the generation of PDPs.
4. **One-dimensional PDP generation**: For each feature of interest, we create a `FeatureEffect` object with the "pdp" method and generate visualizations showing how the average prediction changes as the feature varies.
5. **Two-dimensional PDP generation**: For the interaction analysis between temperature and humidity, we create a two-dimensional PDP showing how these two features interact to influence predictions.
6. **Visualization and organization**: We organize the resulting plots into grids for easier comparison and add appropriate labels to enhance interpretability.
7. **Analysis and interpretation**: We carefully examine each PDP to extract insights about the relationships learned by the model and their practical implications.

This methodological approach allows us to systematically explore how different features influence our models' predictions, providing a solid foundation for interpretation and decision-making.

3 One-Dimensional Partial Dependence Plots for Bike Rental Prediction

3.1 Implementation Process

To analyze how different features influence the prediction of bike rental counts, we followed a structured process:

1. **Data loading and preparation**: We imported the `day.csv` dataset containing information about daily bike rentals along with various meteorological and temporal variables.

2. **Feature selection**: We focused on four key features: days since 2011 (`instant`), normalized temperature (`temp`), normalized humidity (`hum`), and normalized wind speed (`windspeed`).

3. **Model training**: We fit a Random Forest model using these features to predict the bike count (`cnt`). We used 100 trees to balance accuracy and computational efficiency.

4. **Preparation for interpretation**: We encapsulated the trained model using the `Predictor` class from the `iml` package, which facilitates the generation of model-agnostic interpretation methods.

5. **Generation of individual PDPs**: For each feature, we created a `FeatureEffect` object with the "pdp" method and generated visualizations showing how the average prediction varies as the feature changes while all others are held constant.

6. **Visualization configuration**: We adjusted axis limits and labels to ensure consistency and facilitate comparison between plots.

7. **Results organization**: We combined the four PDPs into a 2x2 grid using `arrangeGrob` and `grid.arrange` to facilitate visual comparison.

3.2 Results and Interpretation

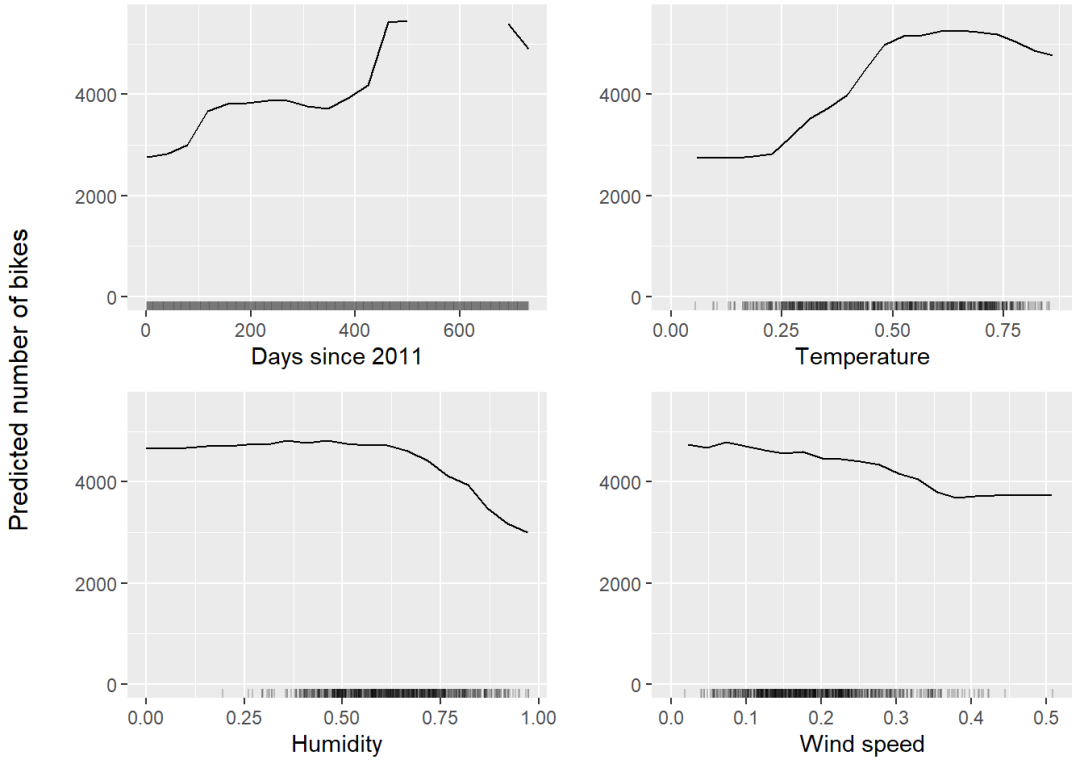

Figure 1: One-dimensional Partial Dependence Plots for bike rental prediction showing the influence of days since 2011, temperature, humidity, and wind speed on predicted bike counts.

Figure 1 shows the one-dimensional PDPs for the four selected features. These plots reveal how each feature individually influences the predicted bike rental counts.

3.2.1 Days Since 2011 (instant)

The PDP for the `instant` feature reveals a clear upward trend in predicted bike rentals over time. This trend can be attributed to several factors:

- **Growing popularity**: The bike-sharing service likely gained more users over time as awareness increased.
- **Infrastructure improvements**: Expansion of bike lanes, docking stations, or service areas may have occurred.
- **Cultural shifts**: Increasing environmental consciousness and preference for alternative transportation.
- **Seasonal accumulation effects**: The data may capture multiple seasonal cycles, with overall growth trends.

The steady increase suggests that temporal factors play a significant role in predicting bike rentals, independent of other features like weather conditions. The plot shows a

nearly linear increase from approximately 3,000 predicted rentals at the beginning of the period to around 5,000 predicted rentals toward the end, indicating a substantial growth in the service's usage over time.

3.2.2 Temperature (temp)

The temperature PDP exhibits a strong positive relationship with predicted bike rentals, characterized by:

- A sharp increase in predicted rentals as temperature rises from low to moderate levels
- A peak at moderate-to-high temperatures (approximately 0.6-0.8 on the normalized scale)
- A slight decline at extremely high temperatures

This pattern aligns with intuitive expectations: people are more likely to cycle in pleasant weather conditions and less likely to do so in extreme cold or heat. The slight decline at very high temperatures likely reflects discomfort from excessive heat. The effect size is substantial, with predicted rentals increasing from approximately 2,000 at the lowest temperatures to over 5,000 at optimal temperatures, making temperature one of the most influential features in the model.

3.2.3 Humidity (hum)

The humidity PDP shows a clear negative relationship with predicted bike rentals:

- Relatively stable predictions at low to moderate humidity levels (up to approximately 0.6)
- A sharp decline in predicted rentals as humidity increases beyond 0.6-0.7
- Lowest predictions at the highest humidity levels

This pattern suggests that while low to moderate humidity has minimal impact on cycling behavior, high humidity significantly deters bike usage, likely due to physical discomfort and perceived exertion. The effect is particularly pronounced at the highest humidity levels, where predicted rentals drop to around 3,000, compared to approximately 4,500 at moderate humidity levels.

3.2.4 Wind Speed (windspeed)

The wind speed PDP reveals a consistent negative relationship with predicted bike rentals:

- Highest predicted rentals at the lowest wind speeds
- A steady decline as wind speed increases
- The effect appears to be approximately linear across the observed range

This negative relationship can be attributed to several factors:

- **Increased physical effort**: Cycling against the wind requires more energy
- **Safety concerns**: Higher wind speeds may make cycling feel less safe
- **Comfort reduction**: Wind can make the cycling experience less pleasant

The effect size is moderate, with predicted rentals decreasing from approximately 4,500 at the lowest wind speeds to around 3,500 at the highest wind speeds.

3.3 Comparative Influence Analysis

Comparing the four PDPs, we can rank the features by their apparent influence on predicted bike rentals:

1. **Temperature**: Shows the largest range of effect on predictions (approximately 3,000 units)
2. **Humidity**: Particularly influential at high levels (approximately 1,500 units)
3. **Days since 2011**: Demonstrates a consistent positive trend (approximately 2,000 units)
4. **Wind speed**: Shows a moderate negative effect (approximately 1,000 units)

This ranking helps prioritize which features are most critical for accurate predictions and provides insights for stakeholders in bike-sharing services. Temperature emerges as the dominant factor, suggesting that weather conditions, particularly temperature, are the primary drivers of bike rental behavior.

4 Two-Dimensional Partial Dependence Plot for Bike Rental Prediction

4.1 Implementation Process

To understand the interaction between temperature and humidity in predicting bike rentals, we generated a two-dimensional PDP following these steps:

1. **Dataset sampling**: Due to computational constraints, we first performed random sampling of the original dataset, selecting 500 observations. We set a random seed (123) to ensure reproducibility.
2. **Training a specific model**: We fit a Random Forest model using only temperature and humidity as predictors, focusing specifically on the interaction between these two features.
3. **Preparation for interpretation**: We encapsulated the model using the `Predictor` class from the `iml` package, specifying the two features of interest.
4. **Two-dimensional PDP generation**: We created a `FeatureEffect` object that considers both features simultaneously, using the "pdp" method to calculate partial dependence values on a two-dimensional grid.

5. **Advanced visualization**: We used `ggplot2` to create a heatmap visualization where:
    - The x-axis represents normalized temperature
    - The y-axis represents normalized humidity
    - The color represents the predicted number of bike rentals
    - Rug plots on the margins show the distribution of observations
6. **Enhancing interpretability**: We applied a viridis color scale for better visual perception and added descriptive labels to facilitate interpretation.

4.2 Results and Interpretation

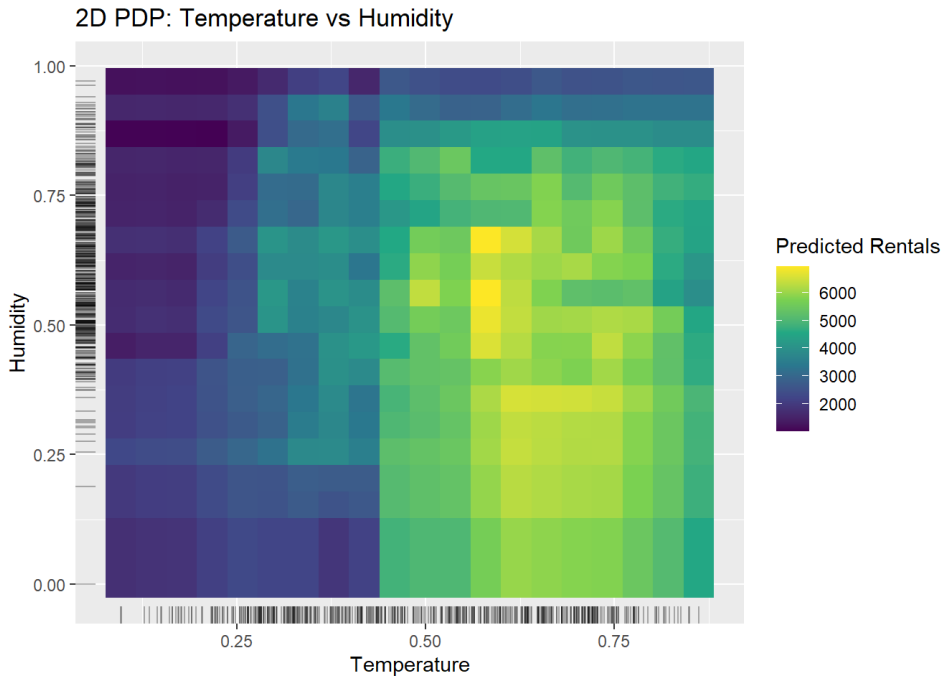

Figure 2: Two-dimensional Partial Dependence Plot showing the interaction between temperature and humidity in predicting bike rentals. The color represents the predicted number of rentals, with darker blue indicating higher values. Rug plots on the margins show the distribution of observations.

Figure 2 shows the two-dimensional PDP for temperature and humidity. This visualization reveals complex interactions between these two features in predicting bike rentals.

4.2.1 Optimal Conditions

The highest predicted rental counts occur when:

- Temperature is moderate to warm (approximately 0.55-0.75 on the normalized scale)
- Humidity is moderate (approximately 0.35-0.60 on the normalized scale)

This region, represented by the darkest blue area in the plot, represents the "sweet spot" for bike rentals, where weather conditions are most conducive to cycling. The predicted rental counts in this optimal region are substantially higher than in other regions of the feature space.

4.2.2 Limiting Factors

Several regions of the plot show significantly reduced predicted rentals:

- **Low temperature effect**: Even when humidity is optimal, low temperatures (below approximately 0.30) suppress predicted rentals significantly. This is visible in the light-colored region at the left side of the plot. This suggests that cold weather is a strong deterrent regardless of humidity levels.
- **High humidity effect**: High humidity (above approximately 0.80) strongly reduces predicted rentals across all temperature ranges, as shown by the light-colored region at the top of the plot. This indicates that high humidity is consistently unfavorable for cycling, even when temperatures are otherwise ideal.
- **Combined extremes**: The lowest predicted rentals occur when both unfavorable conditions coincide—low temperature and high humidity—as seen in the lightest region in the top-left corner of the plot.

4.2.3 Interaction Effects

The 2D PDP reveals interaction effects that are not visible in the one-dimensional PDPs:

- The negative effect of high humidity is more pronounced at moderate temperatures than at low temperatures, as indicated by the steeper color gradient in the middle-top region compared to the left-top region.
- The positive effect of increasing temperature is stronger at low humidity levels than at high humidity levels, shown by the more dramatic color change along the horizontal axis at lower y-values.

These interactions suggest that the effects of temperature and humidity are not simply additive but have complex interdependencies that affect bike rental behavior.

4.2.4 Data Distribution

The rug plots along the margins reveal the distribution of observations:

- Most observations cluster in the moderate temperature and humidity ranges, as shown by the denser rug marks in these regions.
- Fewer observations exist in the extreme regions (very low/high temperature, very low/high humidity), indicated by sparser rug marks.

This distribution information is crucial for assessing the reliability of predictions across different regions of the feature space. Predictions in regions with sparse data should be interpreted with caution, as they are based on fewer observations and may be less reliable.

5 Partial Dependence Plots for House Price Prediction

5.1 Implementation Process

We applied the same PDP methodology to a house price prediction problem using the King County housing dataset. We followed these steps:

1. **Data loading and sampling**: We imported the `kc_house_data.csv` dataset and, due to its size, performed random sampling of 1000 observations to reduce computational load.

2. **Feature selection**: We focused on six key features for price prediction: bedrooms, bathrooms, living area (`sqft_living`), lot size (`sqft_lot`), floors, and year built (`yr_built`).

3. **Model training**: We fit a Random Forest model using these features to predict house price. We used 100 trees to balance accuracy and computational efficiency.

4. **Preparation for interpretation**: We encapsulated the trained model using the `Predictor` class from the `iml` package, specifying the relevant features.

5. **PDP generation for selected features**: We focused on four key features: bedrooms, bathrooms, living area, and floors. For each, we created a `FeatureEffect` object with the "pdp" method.

6. **Visualization configuration**: We adjusted axis limits and labels to ensure consistency and facilitate comparison between plots. We set consistent limits for the y-axis (predicted price) to allow direct comparisons of effect magnitude.

7. **Results organization**: We used an apply function (`lapply`) to generate PDPs for all features of interest with consistent settings, then combined the plots into a 2x2 grid using `arrangeGrob` and `grid.arrange`.

5.2 Results and Interpretation

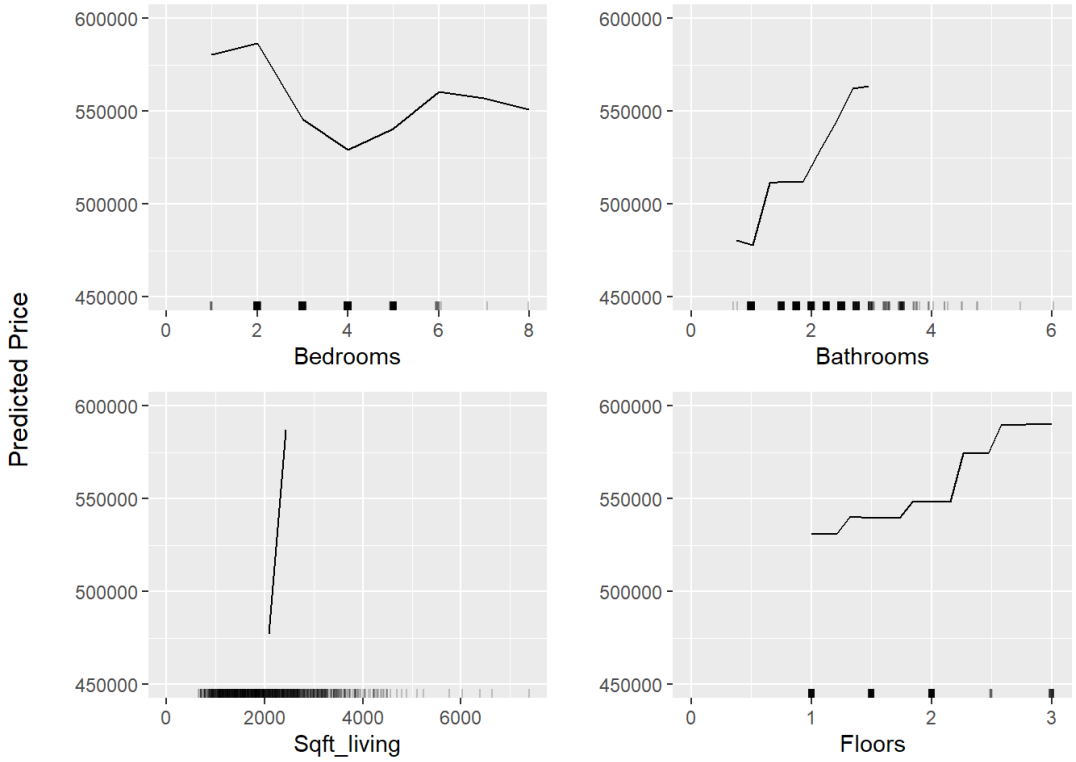

Figure 3: One-dimensional Partial Dependence Plots for house price prediction showing the influence of bedrooms, bathrooms, living area, and floors on predicted house prices.

Figure 3 shows the one-dimensional PDPs for the four selected features. These plots reveal how each feature individually influences the predicted house prices.

5.2.1 Number of Bedrooms

The PDP for the number of bedrooms reveals a complex, non-linear relationship with predicted house prices:

- **Initial increase**: A modest increase in predicted price from 1 to approximately 4 bedrooms
- **Plateau and fluctuation**: Between 4 and 7 bedrooms, the predicted price shows small fluctuations without a clear trend
- **Overall range**: The effect spans approximately $560,000 to $590,000, a relatively narrow range compared to other features

This pattern suggests that while additional bedrooms add some value to a house, this effect has diminishing returns and eventually plateaus. The wavy pattern may indicate interactions with other features or reflect market segments where very large homes have different valuation dynamics. The relatively small effect size suggests that the number of bedrooms alone is not a strong predictor of house price when other features are considered.

5.2.2 Number of Bathrooms

The bathroom PDP shows a clearer positive relationship with predicted house prices:

- **Strong positive trend**: A consistent increase from 1 to approximately 3 bathrooms
- **Substantial effect size**: Moving from 1 to 3 bathrooms increases the predicted price from approximately $540,000 to $585,000
- **Diminishing returns**: The curve begins to flatten beyond 3 bathrooms

This pattern aligns with real estate valuation principles, where bathrooms are typically considered high-value additions to a property. The diminishing returns beyond 3 bathrooms likely reflect practical limitations—most households don't require more than 3 bathrooms, so additional bathrooms add less marginal value. The effect size of approximately $45,000 indicates that bathrooms are a significant factor in house price prediction.

5.2.3 Living Area (sqft_living)

The living area PDP exhibits the strongest and most consistent relationship with predicted house prices:

- **Strong positive correlation**: A steep, nearly linear increase across the entire range
- **Large effect size**: Predicted price increases from approximately $450,000 at 1,000 square feet to over $550,000 at 2,500 square feet
- **Continued growth**: Unlike bedrooms and bathrooms, the curve shows little evidence of plateauing within the observed range

This strong relationship reflects the fundamental importance of size in real estate valuation. Square footage is a primary driver of house prices, as it directly relates to the utility and capacity of the property. The effect size of approximately $100,000 over the observed range makes living area the most influential feature among those analyzed.

5.2.4 Number of Floors

The floors PDP shows a step-wise positive relationship with predicted house prices:

- **Discrete jumps**: Clear increases at each additional floor
- **Moderate effect size**: Each additional floor adds approximately $20,000 to the predicted price
- **Limited range**: The data primarily contains houses with 1-3 floors

This pattern suggests that multi-story homes command premium prices, possibly due to architectural prestige, better space utilization, or correlation with newer construction styles. The step-wise pattern is particularly interesting, showing distinct jumps at integer values, which aligns with the discrete nature of this feature.

5.3 Comparative Analysis of Housing Features

Comparing the influence of the four features on predicted house prices:

1. **Living area (sqft_living)**: Shows the strongest influence with the steepest curve and largest effect size (approximately $100,000)
2. **Bathrooms**: Demonstrates a clear positive relationship with substantial effect size (approximately $45,000)
3. **Floors**: Exhibits a moderate, step-wise positive relationship (approximately $40,000 total)
4. **Bedrooms**: Shows the weakest and most complex relationship with the smallest effect size (approximately $30,000)

This ranking provides valuable insights for real estate stakeholders:

- For homeowners considering renovations, increasing living area or adding bathrooms may yield better returns than adding bedrooms
- For developers, optimizing the balance of these features could maximize property values
- For buyers, understanding these relationships helps in evaluating property values and identifying potentially undervalued properties

6 Discussion

6.1 Methodological Considerations

Several methodological aspects should be considered when interpreting the results:

- **Subsampling effects**: For computational efficiency, we subsampled the datasets before generating PDPs. This may introduce sampling variability, particularly for the 2D PDP and house price analysis.
- **Feature correlation**: PDPs assume feature independence, which may not hold in reality. For example, bedrooms and square footage are likely correlated, potentially affecting the interpretation of their individual PDPs.
- **Model accuracy**: The interpretations are based on the relationships learned by the Random Forest models, which may not perfectly capture the true relationships in the data.
- **Feature scaling**: The features in the bike rental dataset appear to be normalized, while those in the house price dataset are in their original scales. This affects the interpretation of the x-axes in the PDPs.

6.2 Practical Implications

6.2.1 Bike Rental Service Optimization

The insights from the bike rental PDPs could inform several practical strategies:

- **Seasonal planning**: Allocate more bikes during periods with optimal temperature and humidity conditions
- **Weather-based pricing**: Implement dynamic pricing based on weather conditions to optimize revenue
- **Marketing strategies**: Target promotional campaigns during favorable weather periods
- **Infrastructure development**: Prioritize expansion in areas with climate profiles conducive to cycling

6.2.2 Real Estate Valuation and Development

The house price PDPs provide actionable insights for various stakeholders:

- **Property developers**: Optimize the mix of features in new constructions to maximize value
- **Homeowners**: Make informed decisions about which home improvements might yield the best return on investment
- **Real estate appraisers**: Refine valuation models by accounting for the non-linear relationships revealed by the PDPs
- **Buyers and sellers**: Better understand how specific features contribute to property values

6.3 Limitations and Future Work

6.3.1 Limitations

- **Limited feature set**: Our analysis included only a subset of available features. Other important factors may influence the predictions.
- **Temporal aspects**: The bike rental data likely contains seasonal patterns that are not fully captured in our analysis.
- **Geographical factors**: The house price data may contain spatial dependencies that are not accounted for in the model.
- **Model-specific insights**: While PDPs are model-agnostic, the specific relationships revealed depend on the underlying model. Different models might learn different relationships.

6.3.2 Future Work

Several directions for future research emerge from this analysis:

- **Feature interaction analysis**: Extend the 2D PDP analysis to other feature pairs to identify additional interaction effects.
- **Comparison with other XAI methods**: Compare PDP insights with those from other techniques such as SHAP values or Accumulated Local Effects (ALE) plots.
- **Temporal analysis**: Incorporate time-series analysis to understand how feature relationships change over time.
- **Model comparison**: Apply PDPs to different model types (e.g., gradient boosting, neural networks) to assess consistency of learned relationships.
- **Causal inference**: Explore methods to move from correlation to causation in interpreting feature relationships.

7 Conclusion

This study demonstrates the utility of Partial Dependence Plots as a model-agnostic method for interpreting complex machine learning models. By applying PDPs to bike rental and house price prediction problems, we have gained valuable insights into how various features influence the predictions made by Random Forest models.

Key findings include:

- Temperature and humidity have strong, interactive effects on predicted bike rentals, with optimal conditions occurring at moderate levels of both
- Living area and number of bathrooms are the strongest predictors of house prices among the features analyzed
- Many feature relationships are non-linear and exhibit diminishing returns or threshold effects
- Feature interactions can reveal patterns not visible in one-dimensional analyses

These insights not only enhance our understanding of the models but also provide actionable information for stakeholders in the respective domains. The methodology demonstrated here can be applied to a wide range of machine learning models and problem domains, contributing to the broader goal of making AI systems more transparent and interpretable.

8 References

1. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
2. 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.

3. Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

4. Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B, 82(4), 1059-1086.