XAI: Model-agnostic methods

José Milán Server and Juan José Prades García

May 16, 2025

1 Introduction

This report presents an interpretability analysis of Random Forest regression models applied to two datasets: bike rental counts and house prices. While both models achieved satisfactory predictive performance, understanding how individual features influence predictions is crucial for practical insights and trustworthiness. To this end, we use model-agnostic explainability techniques, primarily Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots, to visualize and interpret the marginal effects and heterogeneity of key predictors. These methods allow us to explore the complex, often nonlinear relationships learned by the models, providing valuable understanding beyond traditional performance metrics.

2 Random Forest Model for Bike Rental Prediction

A Random Forest regression model was trained to predict the target variable *cnt* (count of bike rentals) using a set of predictor variables including denormalized temperature (*temp_ denorm*), denormalized humidity (*hum_ denorm*), denormalized windspeed (*windspeed_ denorm*), seasonal indicators, weather conditions (fog, rain), and the time variable *days_ since_ 2011*. The model used 200 trees and achieved an explained variance of 88.27% with a reasonably low mean squared residual error, indicating a well-fitted model.

3 Analysis of Partial Dependence Plots

To better understand how the Random Forest model predicts bike rental counts, Partial Dependence Plots (PDPs) were employed. PDPs offer a clear

visualization of the average effect each feature has on the model's predictions by isolating the influence of a single variable while averaging out the others. This approach is especially useful for complex, non-linear models like Random Forests, where direct interpretation of model parameters is not feasible. Through PDPs, we can gain insight into the direction, magnitude, and shape of each feature's impact on predicted rental counts, while bearing in mind that these plots reflect the model's learned relationships rather than definitive causal effects in the real world.

3.1 Effect of Days Since 2011

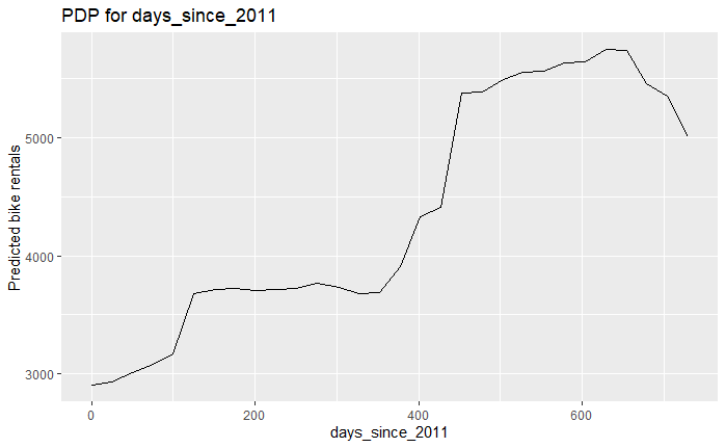

Figure 1: Partial Dependence Plot for *days_since_2011*.

The PDP for *days_since_2011* reveals a generally increasing trend in predicted bike rentals over time, which likely captures both a growing user base and seasonal or temporal patterns. Notably, the relationship is non-linear: periods of relative stability alternate with sharp increases, possibly reflecting event-driven spikes or seasonality. Towards the end of the timeline, a slight decline suggests seasonal downturns or other temporal factors impacting demand.

This behavior exemplifies how the model captures complex temporal dynamics beyond simple linear trends, providing interpretable insights into how time influences rental activity.

3.2 Effect of Temperature

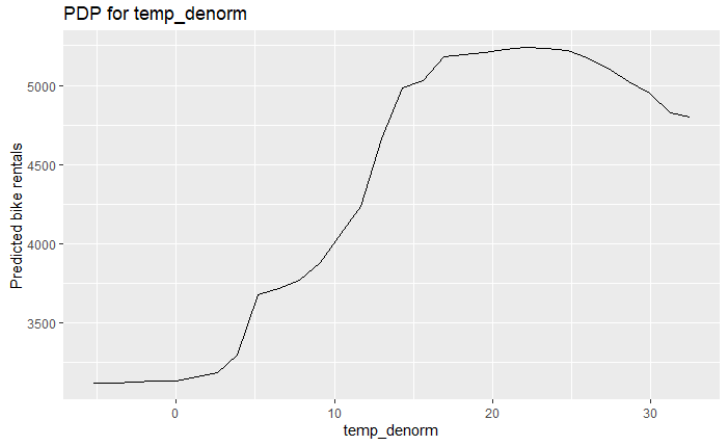

Figure 2: Partial Dependence Plot for *temperature*.

The temperature PDP displays a clear non-linear relationship with predicted rentals. From sub-zero temperatures up to around 20–25°C, predicted bike usage rises substantially, consistent with greater rider comfort in mild to warm weather. Beyond this range, predicted rentals plateau and slightly decline, which could indicate reduced biking under excessively hot conditions.

This pattern is intuitively plausible and highlights the model's ability to learn non-monotonic effects. However, it is important to consider that PDPs represent average marginal effects; potential interactions with other features or heterogeneous responses among subgroups may be masked by this aggregation.

3.3 Effect of Humidity

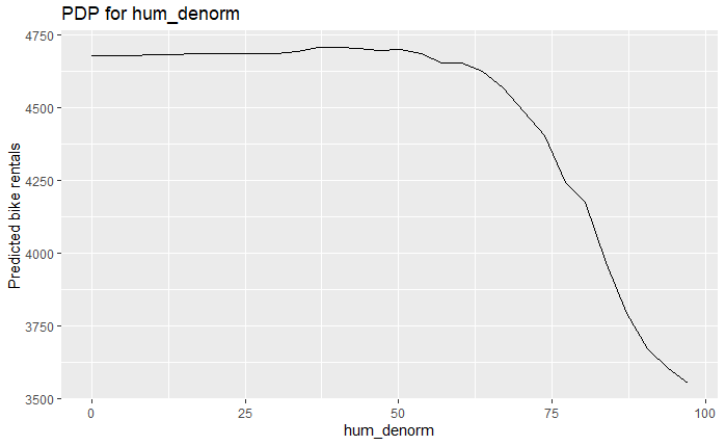

Figure 3: Partial Dependence Plot for *humidity*.

The humidity PDP shows that predicted bike rentals remain fairly stable and relatively high at low to moderate humidity levels (up to around 50%). However, as humidity increases beyond this threshold, the predicted rentals decrease markedly, dropping sharply near 100% humidity.

This indicates that the model associates high humidity—likely indicative of rainy or uncomfortable conditions—with a substantial reduction in bike usage. The non-linear decline aligns with domain knowledge, and this insight can help anticipate demand fluctuations related to weather.

It is worth noting that PDPs assume feature independence; since humidity may correlate with other variables such as temperature or weather conditions, interpretations should be made cautiously.

3.4 Effect of Windspeed

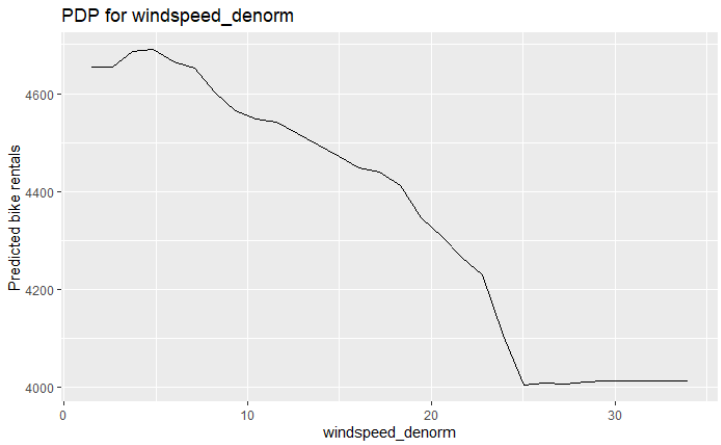

Figure 4: Partial Dependence Plot for *windspeed*.

The PDP for windspeed reveals a nuanced relationship: predicted rentals slightly increase at low wind speeds (0 to 5 units), suggesting mild breezes do not deter riders and may even encourage biking. From moderate to high wind speeds (5 to 23 units), predicted rentals decline progressively, reflecting the negative impact of stronger winds on bike usage. Beyond this, the rentals drop sharply and stabilize at a lower level, highlighting a critical windspeed threshold affecting rider behavior.

This result demonstrates the model's ability to capture non-linear and threshold effects, consistent with expectations that strong winds discourage outdoor cycling.

3.5 Usefulness of Individual Conditional Expectation (ICE) Plots

While Partial Dependence Plots provide an average effect of a feature on model predictions, they can sometimes mask important heterogeneity or interactions present in the data. Individual Conditional Expectation (ICE) plots complement PDPs by illustrating how the prediction changes for each individual instance as the feature varies, holding other features constant. This granular view allows us to detect variation in the effect of the feature across different observations, uncovering patterns that the averaged PDP might obscure.

In this study, an ICE plot was generated specifically for the feature *days_since_2011* to examine the individual-level variability in how time

affects bike rental predictions. The plot displays multiple lines, each representing one data point's predicted response as *days_since_2011* varies, with the overall average trend shown for reference.

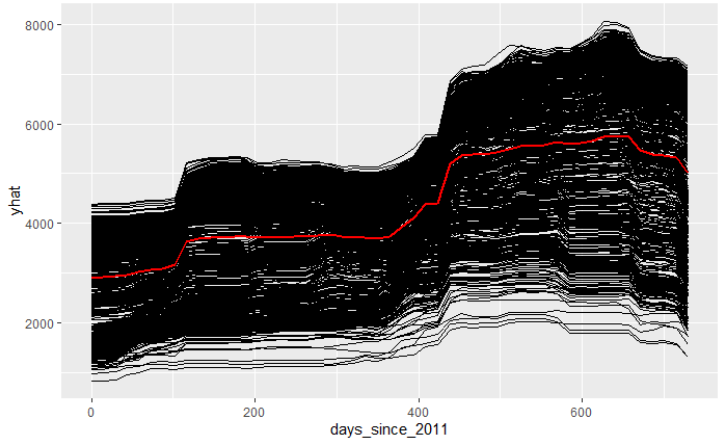

Figure 5: Individual Conditional Expectation (ICE) plot for *days_since_2011* with the average PDP trend shown in red.

The x-axis shows time in days since 2011, covering a range of approximately 0 to 750 days. The y-axis represents the predicted number of bicycles sold.

Initially, predictions range between 2,000 and 4,000 units. Between roughly 100 and 350 days, the predicted sales remain relatively stable, forming a plateau. After around 350 to 400 days, there is a noticeable increase in predicted sales, with the average rising to approximately 5,500 to 6,000 bicycles. Towards the end of the period (600 to 700 days), a slight decline in predicted sales is observed.

The spread of the individual lines indicates variability in how different observations respond to changes in "days_since_2011," suggesting interactions with other features in the dataset.

Overall, the variable "days_since_2011" exhibits a nonlinear temporal effect on bicycle sales, with phases of stability, growth, and slight decline. This confirms that the Random Forest model captures complex, time-dependent patterns in the data.

3.6 Variable Importance

Understanding which features contribute most significantly to the predictive performance of a Random Forest model is crucial for model interpretabil-

ity and practical decision-making. Variable importance measures provide a quantification of each feature's contribution to reducing prediction error, often reflecting how frequently and effectively the feature is used to split the data within the trees.

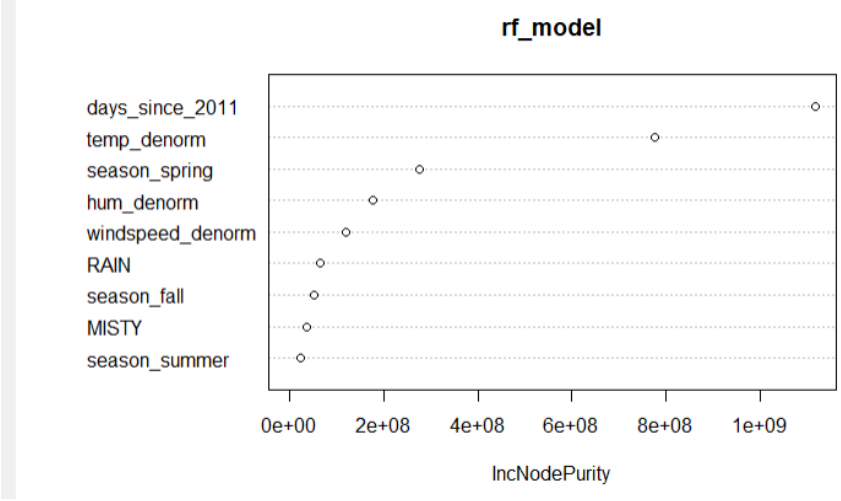

Figure 6: Variable importance based on Increase in Node Purity in the Random Forest model.

The plot displays the variable importance derived from the Increase in Node Purity metric for a Random Forest model predicting bike rentals. This metric quantifies how much each feature contributes to reducing the prediction error by improving node purity in the decision trees.

- **X-axis (IncNodePurity):** Represents the increase in node purity gained when a variable is used for splitting the data. Higher values indicate greater importance for the model's predictive performance.
- **Y-axis (Variables):** Lists the features used in the model.

Most Important Variables

1. `days_since_2011`: This is by far the most important feature, indicating that the elapsed time since 2011 strongly influences bike rental demand. It likely captures long-term trends and seasonality.
2. `temp_denorm` (Temperature): The second most important variable, which is expected since weather conditions heavily affect bike rentals.

3. `season_spring` (Spring season): Shows the seasonal effect, suggesting higher rentals during spring.

4. `hum_denorm` (Humidity): Moderately important, reflecting the impact of humidity on rental behavior.

5. `windspeed_denorm` (Wind speed): Also relevant, though less so than temperature or seasonality.

6. `RAIN`, `season_fall` (Fall season), `MISTY` (Fog), `season_summer` (Summer season): These features have lower but noticeable importance, likely reflecting adverse weather conditions reducing rentals and additional seasonal effects.

Conclusion

The Random Forest model relies primarily on temporal and weather-related variables to predict bike rentals, with clear seasonal patterns. This aligns well with domain knowledge since bike usage is sensitive to weather and time trends.

4 Two-Dimensional Partial Dependence Plot: Temperature and Humidity

To further investigate the combined effect of environmental factors on bike rental predictions, a two-dimensional Partial Dependence Plot (PDP) was generated using *temperature* and *humidity* as input features. Given the computational intensity of 2D PDPs on large datasets, a random subset of the original data was extracted to ensure efficient processing without compromising representativeness.

The 2D PDP illustrates how varying both temperature and humidity simultaneously influences the predicted number of bike rentals by the Random Forest model. Additionally, the joint density distribution of these two features in the dataset is presented to contextualize the regions with higher data concentration, which helps avoid over-interpretation in areas with sparse observations.

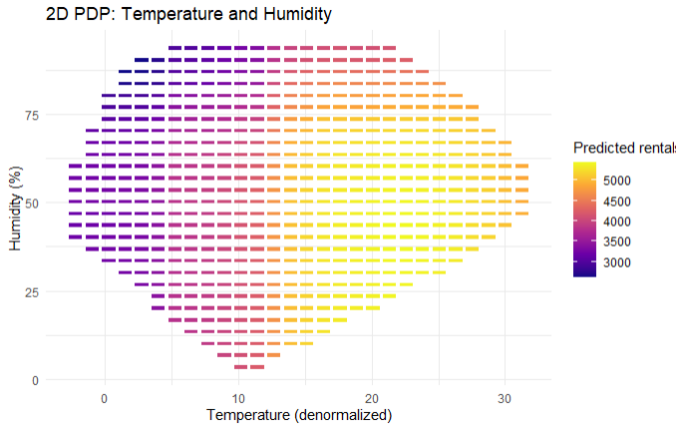

Figure 7: Two-dimensional Partial Dependence Plot for *temperature* and *humidity*.

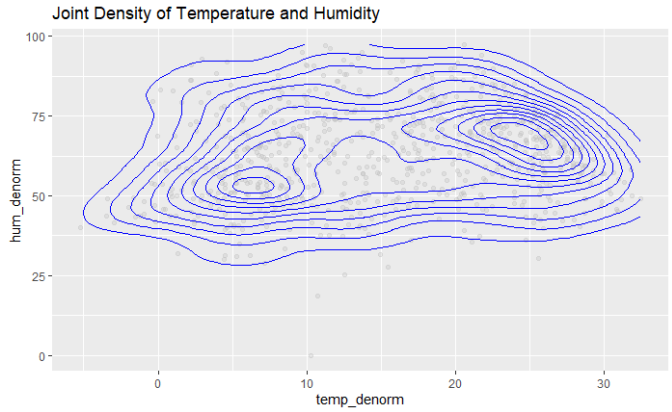

Figure 8: Joint density distribution of *temperature* and *humidity* in the dataset.

The first figure, the two-dimensional Partial Dependence Plot (PDP), illustrates how the predicted bike rentals vary simultaneously with temperature and humidity according to the Random Forest model. It can be seen that the predicted rentals peak at moderate to warm temperatures, roughly between 15 and 30 degrees Celsius, combined with moderate humidity levels around 40% to 60%. Under these conditions, the model predicts the highest rental demand, exceeding 5000 rentals. On the other hand, low temperatures or very high humidity (above 75%) are associated with a significant drop in predicted rentals, indicating lower user activity during colder or highly hu-

mid weather. Similarly, low humidity combined with low temperatures also results in reduced predicted demand.

The second figure displays the joint density distribution of temperature and humidity in the dataset, revealing where the majority of observations lie. The blue contour lines indicate regions with higher data concentration, mostly between 5 and 25 degrees Celsius and humidity levels from 30% to 80%. This density information is crucial to contextualize the PDP results, as it highlights the areas where the model's predictions are more reliable due to sufficient data support. Conversely, areas with extreme temperature or humidity values show sparse data points, suggesting caution when interpreting the model's behavior in these less populated regions.

Together, these visualizations demonstrate that the model predicts higher bike rental activity in weather conditions that are commonly observed in the dataset, specifically moderate temperatures and humidity. Predictions in less frequent environmental conditions are lower and less certain, emphasizing the importance of considering data density alongside model outputs to avoid over-interpreting results in regions with limited observations.

5 Partial Dependence Plots for House Price Prediction

The Random Forest models fitted for both datasets yielded satisfactory performance metrics.

For the bike rental prediction, the model was trained using 200 trees with three variables considered at each split. The resulting mean squared residual (MSR) was approximately 439,460, while the model explained 88.27% of the variance in the target variable. This indicates a strong fit with relatively low prediction error, reflecting the model's capability to capture the complex relationships influencing bike rental counts.

In the case of house price prediction, the Random Forest model also employed 200 trees but used two variables at each split. The model produced a higher mean squared residual of roughly 51.5 billion, with a percentage of variance explained equal to 62.34%. Although the fit is less precise compared to the bike rental model, it remains reasonably informative, given the inherent complexity and variability of housing prices. The lower variance explained may reflect greater noise or non-modeled factors in the housing dataset.

These metrics provide an initial quantitative assessment of model quality before delving into interpretability analyses via Partial Dependence Plots.

5.1 Effect of Bedrooms

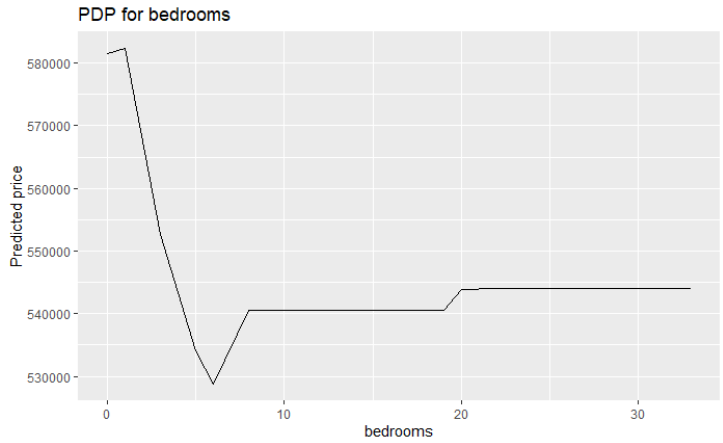

Figure 9: Partial Dependence Plot for *bedrooms*.

5.2 Effect of Bathrooms

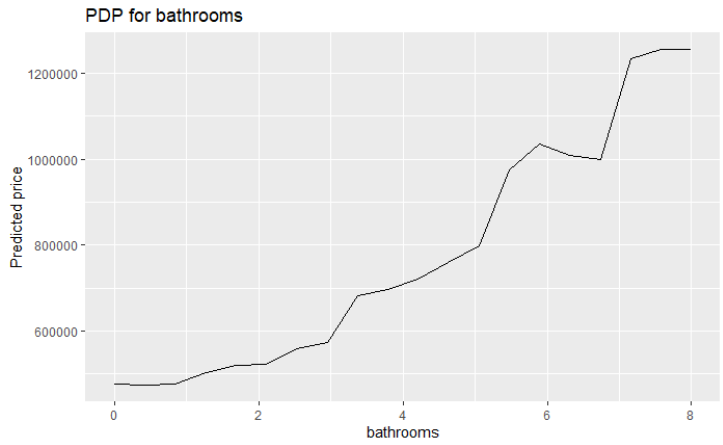

Figure 10: Partial Dependence Plot for *bathrooms*.

The PDP for *bathrooms* shows a clear positive relationship between the number of bathrooms and the predicted house price. As the number of bathrooms increases from 0 to 8, the predicted price rises steadily, with some noticeable jumps especially after 3 and 5 bathrooms. This pattern suggests that bathrooms have a strong and mostly monotonic effect on house prices, reflecting their important role in valuation. The non-linear increments and jumps may

indicate threshold effects where additional bathrooms lead to discrete value increases.

In contrast, the PDP for *bedrooms* reveals a different pattern. The predicted price initially decreases as the number of bedrooms increases from 0 up to about 7, after which the predicted price stabilizes and remains almost constant for higher bedroom counts. This suggests that having more bedrooms beyond a certain point does not significantly increase the predicted price and might even slightly reduce it at low bedroom counts. This could reflect market preferences or diminishing returns on adding bedrooms in the dataset. The flat region indicates a saturation effect where extra bedrooms no longer add value according to the model.

Together, these PDPs provide insights into how the Random Forest model perceives the influence of bathrooms and bedrooms on house prices. Bathrooms appear to have a stronger and more consistently positive impact, while bedrooms show a more complex and non-monotonic effect. These findings align with the overall model performance described, where the model captures some non-linearities and threshold behaviors in the housing market.

5.3 Effect of Square Footage of Living Area (*sqft_ living*)

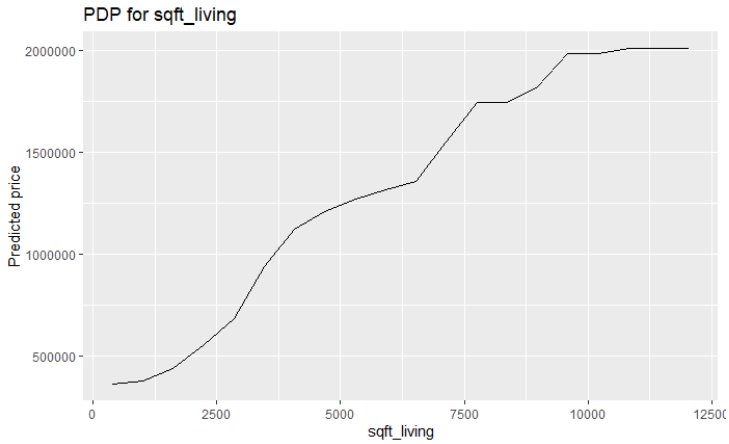

Figure 11: Partial Dependence Plot for *sqft_ living*.

The PDP for *sqft_ living* shows a strong positive relationship between the size of the living area and the predicted house price. As the living area increases from around 500 sqft to over 10,000 sqft, the predicted price rises substantially. The increase is not strictly linear; distinct steps or jumps

appear, likely reflecting threshold effects where price increments occur more abruptly at certain size intervals.

There is also evidence of a saturation effect near the upper end of the range (around 9,000 to 10,000 sqft), where further increases in living area no longer produce substantial price increases. This plateau could indicate market limits or model extrapolation boundaries for very large properties.

Overall, the model captures a mostly monotonic positive effect of living area on house price, consistent with typical real estate market expectations.

5.4 Effect of Floors

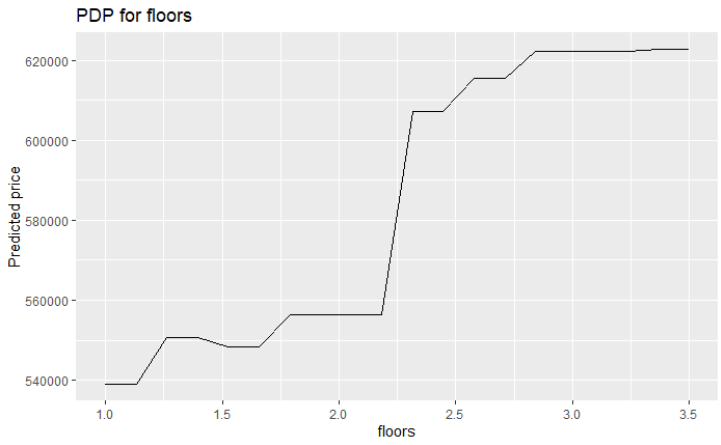

Figure 12: Partial Dependence Plot for *floors*.

The PDP for *floors* indicates a generally positive but more moderate influence on the predicted house price. Predicted prices tend to increase with the number of floors, although the magnitude of change is smaller compared to *sqft_living*. The plot shows noticeable jumps between some floor counts, particularly between 2 and 3 floors, suggesting discrete increments in value associated with additional floors.

Since *floors* is a discrete variable with a limited range, the PDP reflects stepwise changes rather than a smooth continuous effect. The model interprets the number of floors as a factor positively affecting house price but with less impact and more nuanced behavior compared to living area.

Together, these PDPs provide valuable insights into the Random Forest model's learned relationships. The *sqft_living* variable exhibits a strong, mostly monotonic effect with non-linear thresholds and saturation, highlighting its importance in house price prediction. In contrast, *floors* shows a pos-

itive but smaller effect, with discrete jumps revealing threshold effects in the number of floors.

These findings align with the overall model performance, where the model explains about 62% of variance in house prices, reflecting the complexity and inherent variability of the housing market. The PDPs demonstrate the model's ability to capture non-linear and threshold effects in key features affecting price.

5.5 Variable Importance

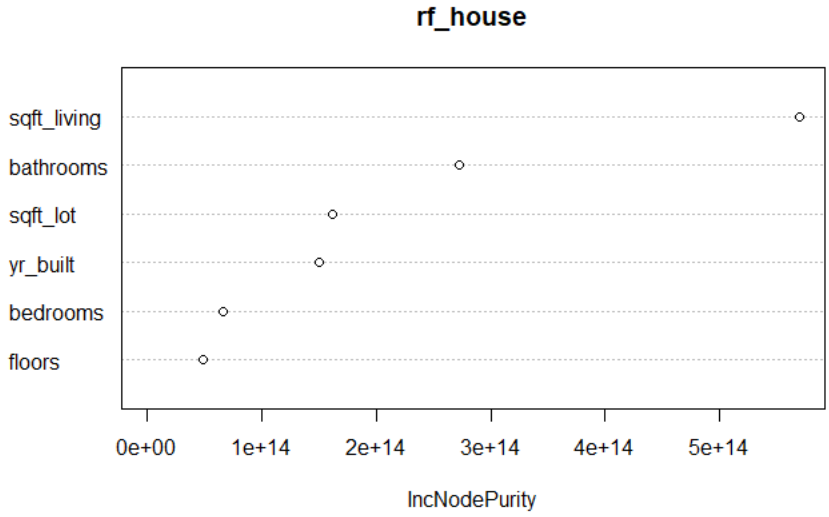

Figure 13: Variable importance based on Increase in Node Purity in the Random Forest model.

The plot displays the relative importance of features used by the Random Forest model to predict house prices, measured by the increase in node purity (IncNodePurity).

- `sqft_living` (living area in square feet) is by far the most important variable, showing a significantly higher IncNodePurity value than all others. This indicates that living area is the strongest predictor and contributes most to reducing prediction error in the model.
- `bathrooms` also shows considerable importance, ranking second among the features.

- `sqft_lot` (lot size) and `yr_built` (year built) have moderate importance.
- `bedrooms` and `floors` have relatively low importance, being the least influential predictors among the variables considered.

Overall, the model primarily relies on living area and number of bathrooms for its predictions, while bedrooms and floors have a smaller impact.

5.6 Individual Conditional Expectation (ICE) Plot for *sqft_living*

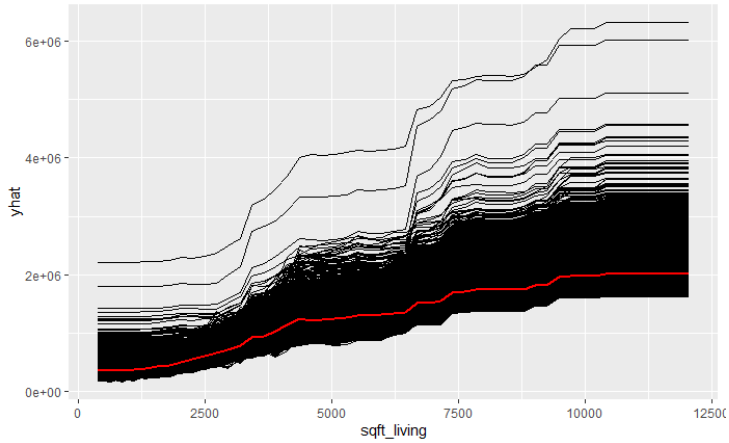

Figure 14: ICE plot for *sqft_living*. The red line shows the average PDP trend.

This plot shows the Individual Conditional Expectation (ICE) curves for the `sqft_living` variable, illustrating how it affects predicted house prices ($\hat{y}$) individually for each observation in the dataset:

- Each black line represents how the predicted price for a single observation changes as `sqft_living` varies, while keeping all other features fixed.
- Generally, the curves display a clear upward trend: as the living area increases, the predicted house price also increases, confirming a positive relationship.

- The lines are not parallel and show considerable variability, indicating that the effect of `sqft_living` depends on the context of other house characteristics.
- Some noticeable steps or jumps suggest threshold effects or nonlinearities in this relationship.
- The red line corresponds to the average effect (Partial Dependence Plot), reinforcing the positive trend with visible step-like increments.
- The horizontal axis covers a wide range of living areas, from very small to very large houses, with the largest price increases occurring in the mid-to-upper range.

In summary, the ICE plot highlights that living area is a key driver in the model's house price predictions, with a generally positive but context-dependent influence.

6 Conclusions

The application of Partial Dependence and Individual Conditional Expectation plots has provided meaningful insights into the behavior of Random Forest models for both bike rental and house price predictions. For bike rentals, temporal and weather-related variables showed intuitive and nuanced effects, highlighting how factors such as time, temperature, humidity, and windspeed jointly influence demand. The two-dimensional PDP further revealed important interaction effects between temperature and humidity.

In the housing price model, living area and number of bathrooms emerged as dominant predictors with mostly positive impacts, while bedrooms and floors had more complex or moderate effects. The ICE plots underscored individual variability in the influence of living area, reflecting contextual dependencies captured by the model.

Overall, these interpretability tools complement quantitative performance metrics by uncovering nonlinearities, threshold effects, and feature interactions learned by the models. Such insights are crucial for informed decision-making and demonstrate the value of explainable machine learning in practical applications.