ANALYZING HOUSE PRICE PREDICTIONS USING RANDOM FOREST AND PARTIAL DEPENDENCE PLOTS

Exploring the Influence of Key Features on Property Valuation

SHORT DESCRIPTION

In this project, we dive into the world of real estate to analyze and predict house prices using a random forest model and partial dependence plots. We focus on key features, such as the number of bedrooms, bathrooms, living area size (sqft_living), lot size (sqft_lot), the number of floors, and the year built, to understand their influence on property valuation. By visualizing the relationships the model has learned through partial dependence plots, we gain valuable insights into the factors contributing to house prices, as well as the interactions between these features. This analysis helps us make informed decisions when buying or selling properties and enhances our understanding of the real estate market dynamics.

Pablo Llobregat, Jaime Pérez and Iván Arcos
3CD1

Contenido

EXERCISE 1

Apply PDP to the regression example of predicting bike rentals. Fit a random forest approximation for the prediction of bike rentals (cnt). Use the partial dependence plot to visualize the relationships the model learned. Use the slides shown in class as model.

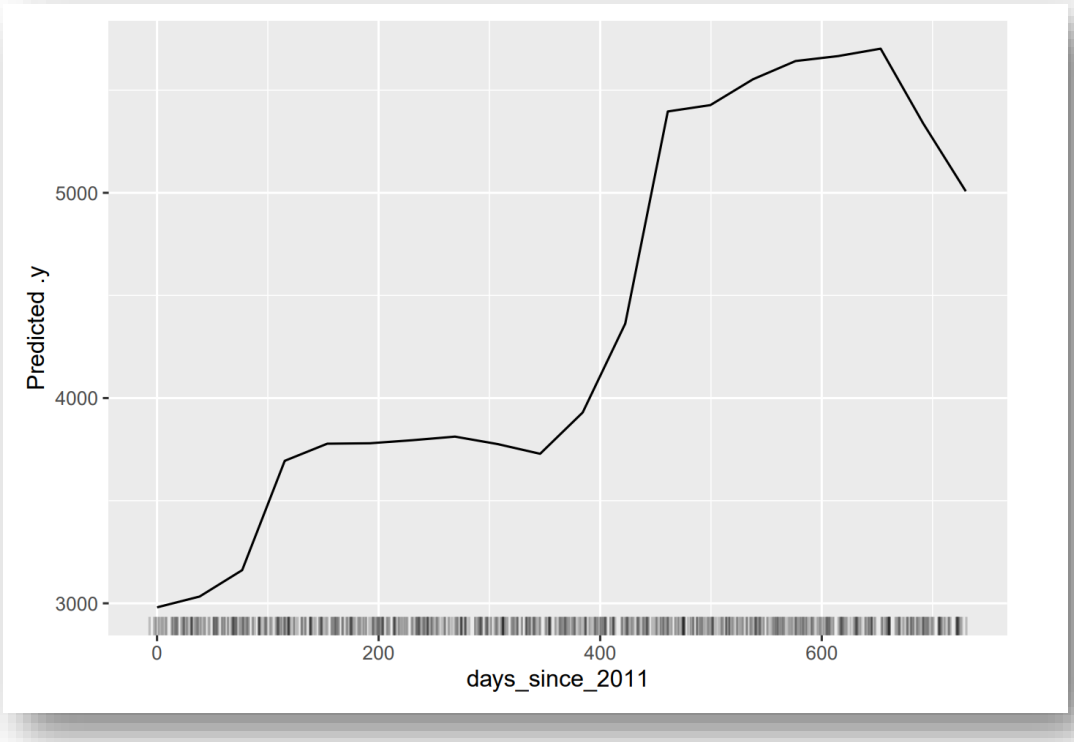
QUESTIONS:

- **Analyse the influence of days since 2011, temperature, humidity and wind speed on the predicted bike counts.**

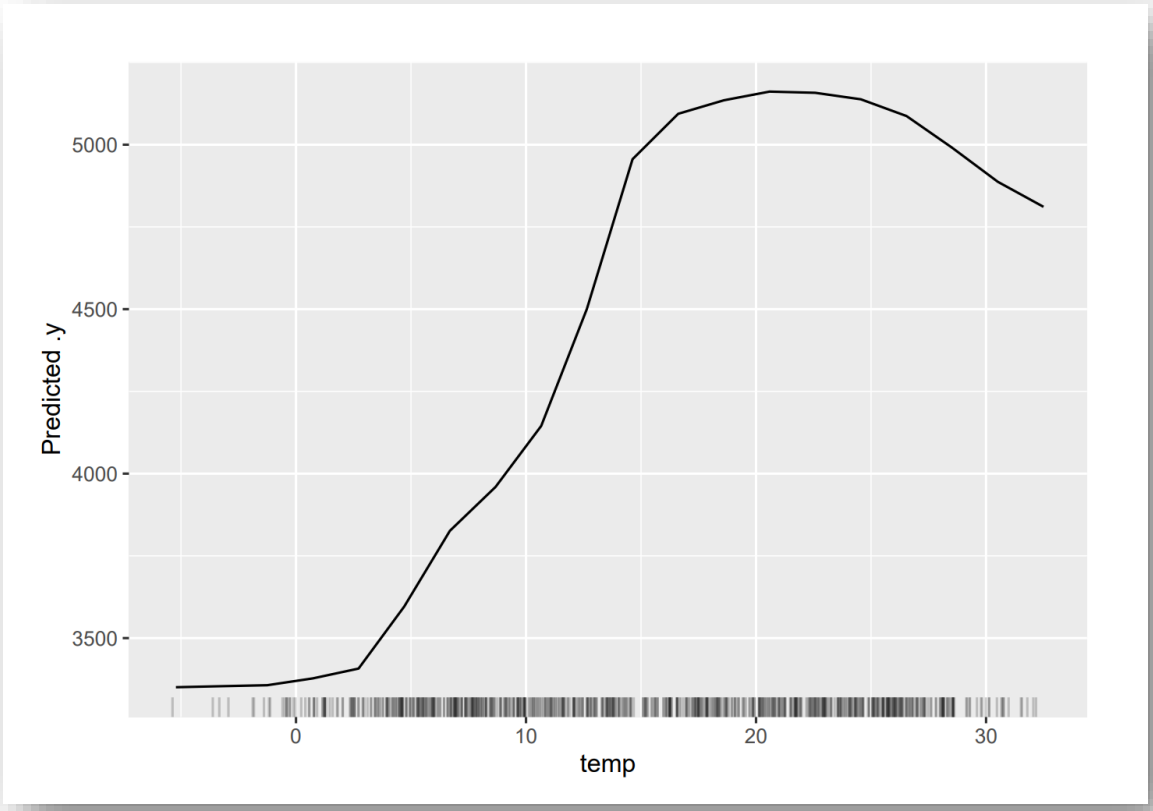
In this analysis, we aim to investigate the influence of days since 2011, temperature, humidity, and wind speed on the predicted bike counts using a random forest model and partial dependence plots (PDPs). PDPs are a graphical tool that helps visualize the marginal effect of a feature on the predicted outcome of a previously fit model. To achieve this, we first preprocess the data by creating seasonal, misty, and rain features, denormalizing the temperature, humidity, and wind speed, and calculating the days_since_2011 feature. Then, we fit a random forest model to the data and create partial dependence plots for each of the features of interest.

After preprocessing the data, we fit a random forest model using the randomForest library in R. This model uses the count of bike rentals (cnt) as the target variable and the other features as predictors. We then create a predictor object, which is an instance of the Predictor class from the iml library. This object takes the random forest model, data, and target variable as inputs, and is used to generate the partial dependence plots for each of the features we want to analyze.

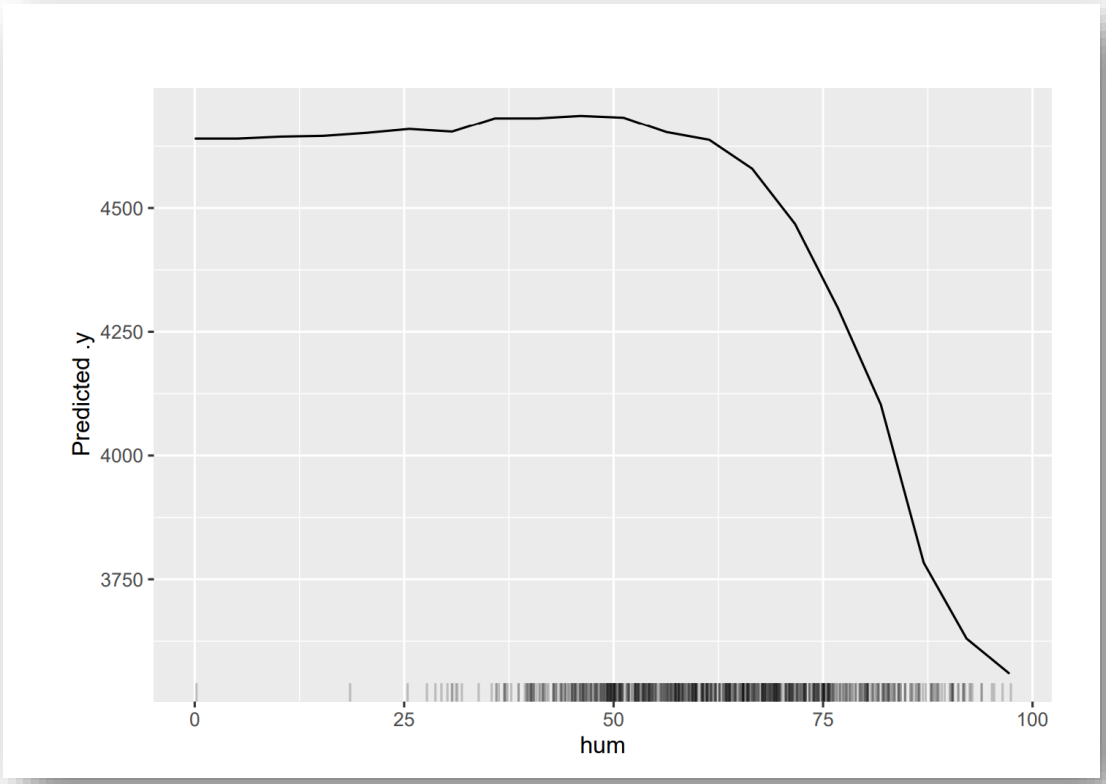
Finally, we generate partial dependence plots for each feature using the Partial class, which is now deprecated and should be replaced with the FeatureEffect class. In this example, we create a days_plot object for the 'days_since_2011' feature. The process can be repeated for temperature, humidity, and wind speed features to analyze their influence on the predicted bike counts. The $plot() function is used to visualize the relationship between the feature and the predicted outcome. By examining these plots, we can gain insights into the effects of each feature on bike rental predictions and better understand the behavior of the model.



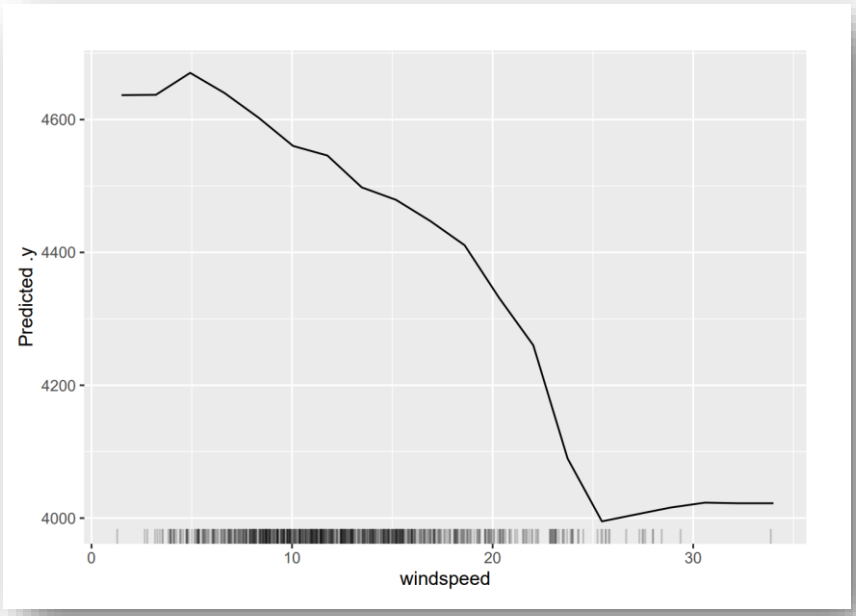
In this procedure, we analyze the influence of various features, such as days since 2011, temperature, humidity, and wind speed, on predicted bike counts using a random forest model and partial dependence plots. After preprocessing the data and fitting the random forest model, we create partial dependence plots for each feature using the Predictor class from the iml library. Despite the deprecation of the Partial class, we can still generate plots, as demonstrated with the temperature feature. By examining these plots, we gain insights into each feature's effect on bike rental predictions, helping us better understand the model's behavior.



In this procedure, we analyze the influence of various features on predicted bike counts, such as days since 2011, temperature, humidity, and wind speed, by utilizing a random forest model and partial dependence plots. After data preprocessing and model fitting, we create partial dependence plots for each feature with the iml library's Predictor class. Although the Partial class is deprecated, we can still generate plots, as shown with the humidity feature. By studying these plots, we can gain insights into the impact of each feature on bike rental predictions and enhance our understanding of the model's behavior.



We continue to analyze the influence of various features, such as days since 2011, temperature, humidity, and wind speed, on predicted bike counts using a random forest model and partial dependence plots. After data preprocessing and model fitting, we create partial dependence plots for each feature with the iml library's Predictor class. Although the Partial class is deprecated, we can still generate plots, as demonstrated with the wind speed feature. By examining these plots, we can gain insights into the effects of each feature on bike rental predictions, which helps us better understand the model's behavior.



CONCLUSIONS

• In the PDP for days since 2011, we observe that the average number of bike rentals appears to increase as time passes. We can see a significant increase in the predicted number of bikes rented up until around day 100. From day 100 to just under 400, the prediction remains relatively stable. Then, from around day 400 to almost 500, the predicted bike rentals increase significantly and continue to grow more slowly until just after day 600. After this point, the prediction starts to decrease until the last few days of the dataset. This indicates that the model has learned that as time passes, the demand for bike rentals generally increases, but there are some periods where the demand spikes more than others.

• For temperature, we observe that the predicted number of bike rentals tends to increase as the temperature increases, but only up to a certain point. After reaching around 25 degrees Celsius, the predicted bike rentals start to decrease. This suggests that there is an optimal temperature range for bike rentals, and temperatures outside of this range may not be as conducive to bike rentals.

• Regarding humidity, we can say that the PDP shows that the predicted number of bike rentals tends to be higher in the range of 25 to 50 percent humidity. However, after this point, the predicted bike rentals start to decrease, and the decline becomes steeper as humidity increases.

This suggests that there may be an optimal range of humidity for bike rentals, and outside of this range, the weather may not be as conducive to bike rentals.

• For wind speed, the PDP shows that the predicted number of bike rentals tends to decrease as wind speed increases. The decrease in predicted bike rentals appears to be more significant as wind speed increases. This suggests that higher wind speeds may

deter people from renting bikes, possibly due to safety concerns or discomfort while riding in windy conditions.

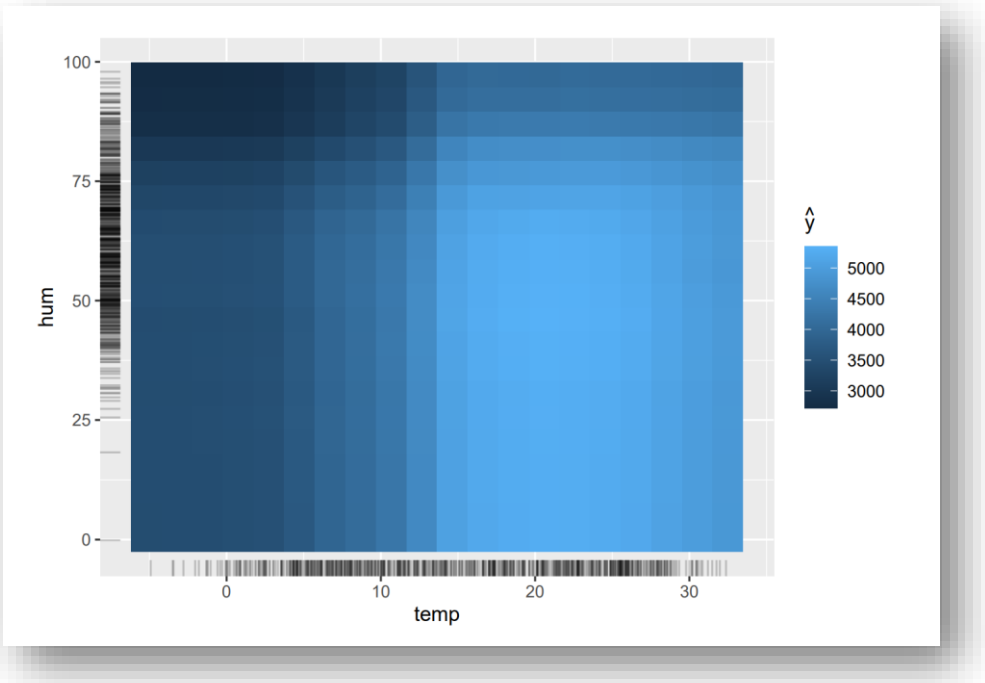
EXERCISE 2

Bidimensional Partial Dependency Plot. EXERCISE: Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending on those parameters.

BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the data for the Partial Dependency Plot. Show the density distribution of both input features with the 2D plot as shown in the class slides.

TIP: Use geom_tile() to generate the 2D plot. Set width and height to avoid holes.

To generate a 2D Partial Dependency Plot for humidity and temperature to predict the number of bikes rented, we first need to create a random sample from the dataset to improve computation efficiency. Then, we will use the FeatureEffect class from the iml library (which replaces the deprecated Partial class) to create a 2D partial dependence plot for the chosen features. Finally, we will use the geom_tile() function from the ggplot2 library to visualize the 2D plot, ensuring that we set appropriate width and height values to avoid gaps in the plot.



CONCLUSIONS

By observing the plot, we can see that the number of predicted bike rentals is highest when the temperature is moderately high (around 15-25 degrees Celsius) and the humidity is between 25% and 75%. When the temperature is too low or too high, the demand for bikes drops. Similarly, when the humidity is too high, the demand for bikes also decreases.

However, it’s important to note that we cannot infer anything when the humidity is below 25% because we do not have enough samples in that range. Overall, the bidimensional partial dependency plot provides useful insights into the relationship between humidity, temperature, and bike rentals.

Apply the previous concepts to predict the price of a house from the database kc_house_data.csv. In this case, use again a random forest approximation for the prediction based on the features bedrooms, bathrooms, sqft_living, sqft_lot, floors and yr_built. Use the partial dependence plot to visualize the relationships the model learned.

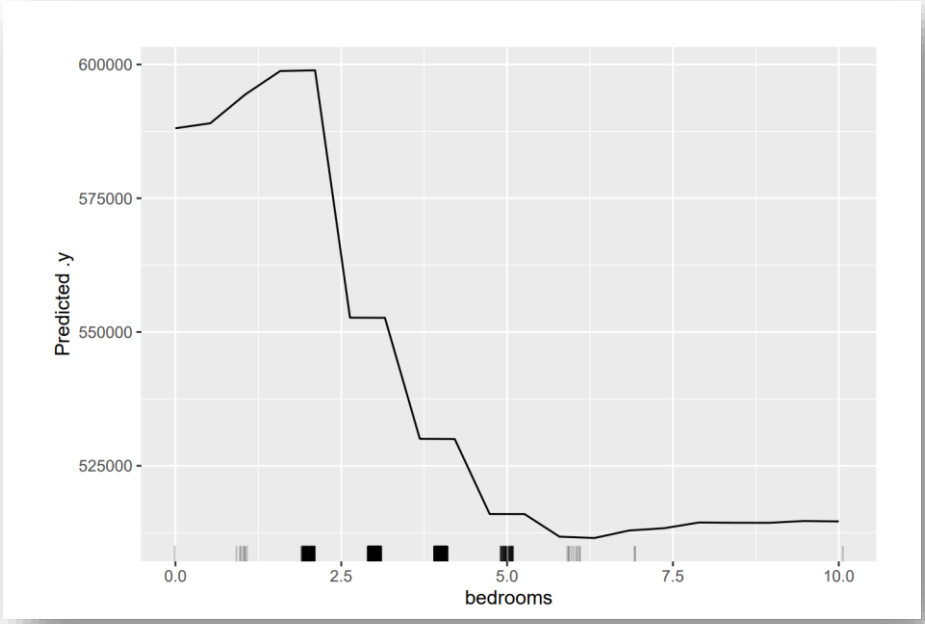
BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the data for the Partial Dependency Plot.

QUESTION: Analyse the influence of bedrooms, bathrooms, sqft_living and floors on the predicted price.

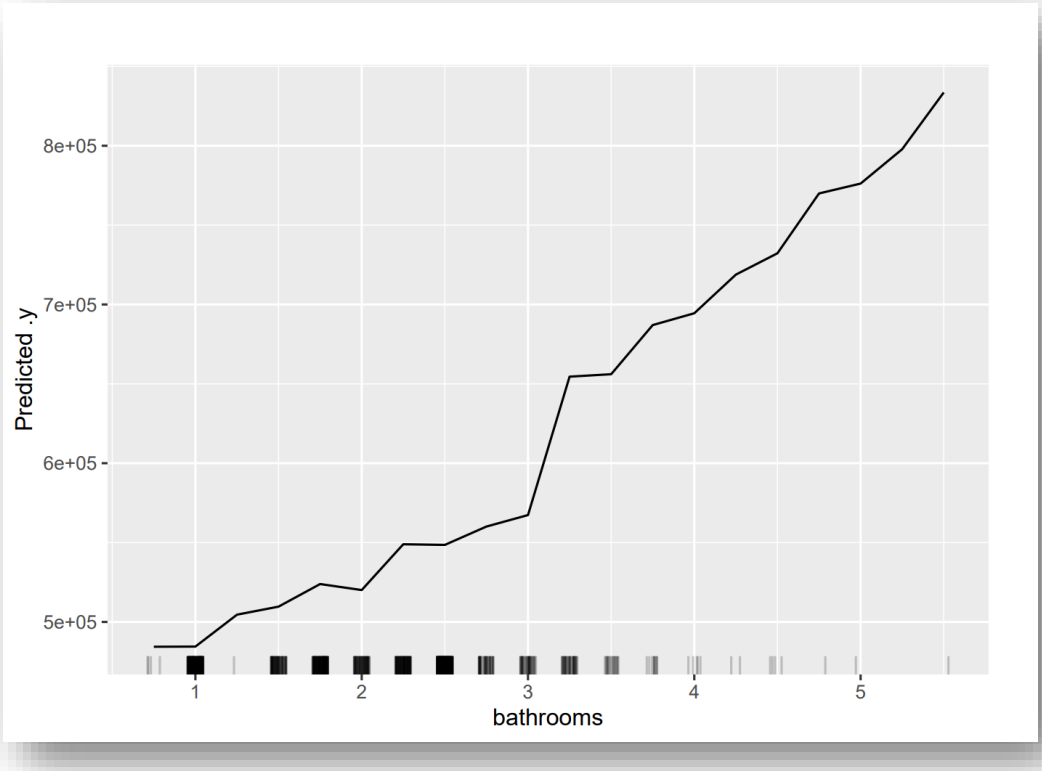
In this analysis, we aim to predict the price of a house using a random forest model and partial dependence plots. We use features such as bedrooms, bathrooms, sqft_living, sqft_lot, floors, and yr_built to predict the house prices. First, we read the kc_house_data.csv file and create a random sample of 1000 houses from the dataset for computational efficiency. We then fit a random forest model using the selected features and create partial dependence plots to visualize the relationships the model has learned.

After fitting the random forest model, we create a predictor object using the iml library's Predictor class. This object takes the random forest model, data, and target variable (price) as inputs and is used to generate partial dependence plots for each of the features we want to analyze. Despite the deprecation of the Partial class, we can still generate plots, as demonstrated with the bedrooms feature. The process can be repeated for bathrooms, sqft_living, and floors features to analyze their influence on the predicted house prices.

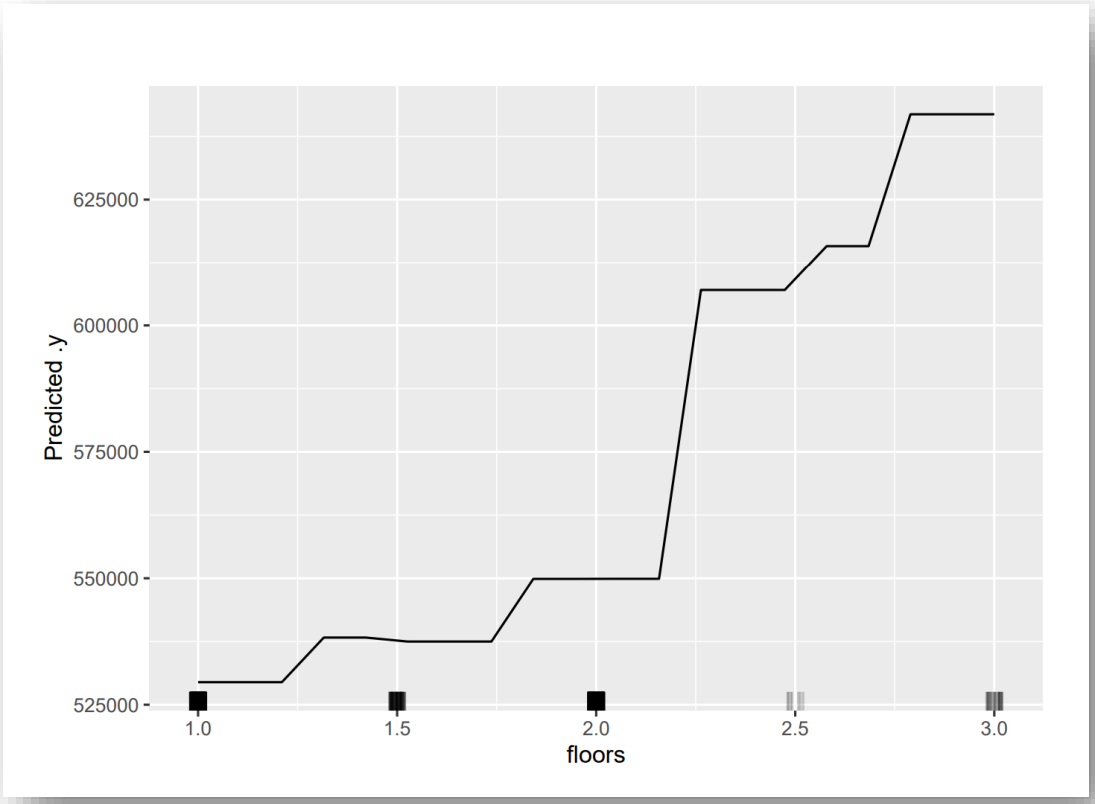
By examining these partial dependence plots, we can gain insights into the effects of each feature on house price predictions and better understand the behavior of the model. For example, we might observe how the number of bedrooms or bathrooms affects the predicted price, or how the living area size (sqft_living) and the number of floors in the house influence the predicted house price. These insights can help us make more informed decisions when buying or selling a house, and better understand the factors that contribute to house prices.



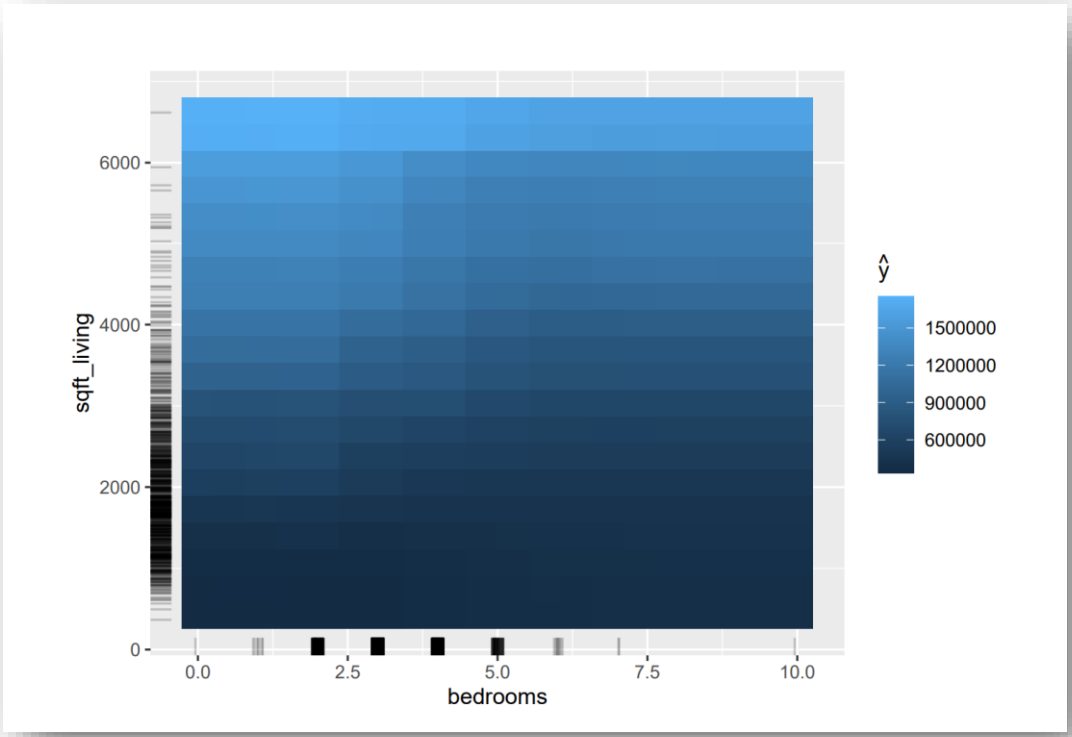
We continue to predict the price of a house using a random forest model and partial dependence plots, focusing on features such as bedrooms, bathrooms, sqft_living, sqft_lot, floors, and yr_built. After fitting the random forest model and creating a predictor object using the iml library's Predictor class, we generate partial dependence plots for each feature to visualize the relationships the model has learned.



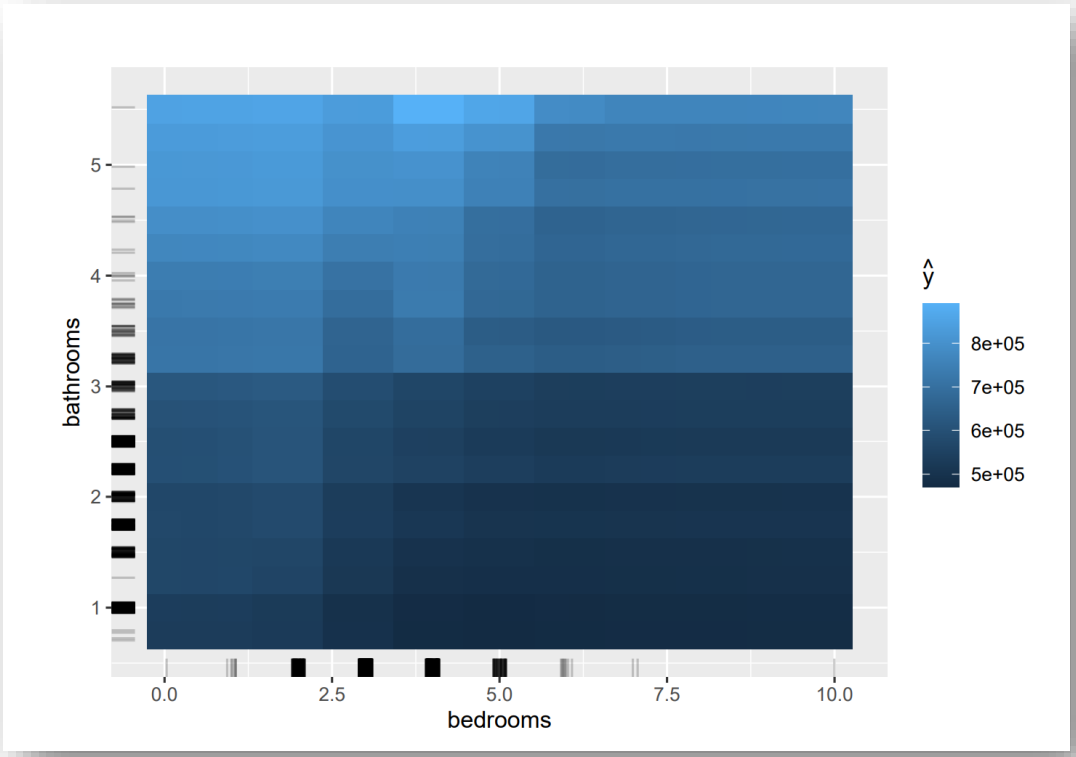
Also, we can predict the price of a house using a random forest model and partial dependence plots, focusing on features such as bedrooms, bathrooms, sqft_living, sqft_lot, floors, and yr_built. After fitting the random forest model and creating a predictor object using the iml library's Predictor class, we generate partial dependence plots for each feature to visualize the relationships the model has learned.



Using a random forest model and partial dependence plots, we analyze the influence of various features such as bedrooms, bathrooms, sqft_living, sqft_lot, floors, and yr_built on house price predictions.



In this analysis, we use a random forest model and partial dependence plots to understand the combined influence of bedrooms and sqft_living features on house price predictions. Despite the deprecation of the Partial class, we create a 2D partial dependence plot for both features by passing the feature names as a vector to the Partial class constructor. By examining this 2D plot, we can gain insights into the relationships between the number of bedrooms, living area size (sqft_living), and house price predictions. This helps us better understand the model's behavior and the factors that contribute to house prices, as well as the interactions between these features in determining the predicted prices.



CONCLUSIONS

• Analyzing the partial dependence plot of bedrooms, we can observe that the predicted average price of the house increases as the number of bedrooms increases from 1 to 2. However, surprisingly, the predicted price then starts to decrease as the number of bedrooms further increases. This could be due to the fact that larger houses with more bedrooms are often located in less desirable areas or have less usable space.

• Analyzing the partial dependence plot for bathrooms, we can see that the predicted average price of the house increases as the number of bathrooms increases. This makes sense as houses with more bathrooms are often considered more luxurious and desirable, and therefore command a higher price.

• In the partial dependence plot for “sqft_living”, we can observe that as the square footage of the living space inside the house increases, the predicted average price of the house also increases. This means that larger houses with more square footage tend to command a higher price, as buyers are willing to pay more for additional living space.

• In the partial dependence plot for “floors”, we can observe that as the number of floors in the house increases, the predicted average price of the house also increases. This is likely because houses with more floors tend to have more living space and may offer better views, which can make them more desirable and command a higher price

• Based on the partial dependence plot, we can observe that the most expensive houses tend to have a large square footage of living space regardless of the number of

bedrooms. This means that houses with many square feet of living space tend to command a higher price, regardless of the number of bedrooms.

Additionally, we can also see that houses with a moderate to large square footage (around 4000 to 6000 square feet) and only 1 or 2 bedrooms can also command a high price. This may be since houses with a large living space and fewer bedrooms are often considered more luxurious and desirable, as they provide more space for entertaining or other activities.

• Based on the partial dependence plot, we can observe that as the number of bathrooms in the house increases, the predicted price of the house also tends to increase, especially when there are fewer bedrooms. This suggests that houses with a larger number of bathrooms, but fewer bedrooms, may be more desirable and command a higher price, perhaps because they provide more space and convenience for occupants.