



---

# MACHINE LEARNING REGRESSION MODEL BUILDING AND DEPLOYMENT ON MLOS

---

AUTHOR - NATESH KUMAR

## List of Content

- 1. Introduction**
- 2. Business Understanding**
  - a. Business Goal
  - b. Model Objective
- 3. Problem Statement**
  - a. Why is it worth solving?
  - b. Importance of the Problem
  - c. Data Sourcing
- 4. Data Collection**
  - a. Data Understanding
- 5. Exploratory Data Analysis (EDA)**
  - a. Univariate Analysis
- 6. Data Preparation**
  - a. Handling missing values
  - b. Data Wrangling
  - c. Normalization
  - d. Input and Target features
  - e. Data Splitting
- 7. Methodology**
  - a. Model Selection
    - i. Random Forest
    - ii. Extra Tree
    - iii. Decision Tree
  - b. Model Evaluation
- 8. Best model**
  - a. Feature Importance
  - b. Model Accuracy
  - c. Performance Metrics
  - d. Histograms
  - e. Feature Distribution
- 9. Result**
- 10. Real-time Prediction**
- 11. Conclusion**
  - a. How could an organization or institution implement/use the solution?
  - b. Lesson Learned

## 1. Introduction

House price dataset used in this project is a comprehensive collection of real estate data from India, consisting of 14,620 entries. Each entry represents a house and includes a variety of features that describe the characteristics and attributes of the house. These features range from basic information such as the number of bedrooms and bathrooms to more specific details like the presence of a waterfront, the condition and grade of the house, and its proximity to important amenities.

## 2. Business Understanding

### Business Goal

Objective is to build a predictive model for house prices to assist stakeholders in making data-driven decisions regarding property investments.

### Model Objective

Create a machine learning model that can accurately predict house prices based on input features such as the number of bedrooms, bathrooms, living area, lot area, and other attributes.

## 3. Problem Statement

### Why is it worth solving?

Predicting house prices accurately is crucial for various stakeholders in the real estate market. Accurate predictions can lead to better investment decisions, fair pricing, and improved customer satisfaction.

### Importance of the Problem

The problem is significant because it directly impacts financial decisions in the real estate market, influencing buyers, sellers, and investors. A reliable prediction model can enhance market efficiency and reduce the risk associated with property investments.

### Data Sourcing

The dataset was sourced from [Kaggle](#) and includes various features such as the number of bedrooms, bathrooms, living area, lot area, and other characteristics.

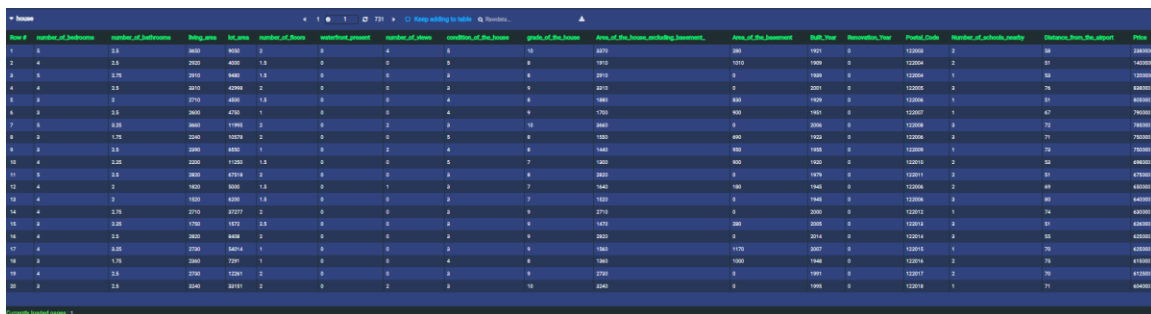
## 4. Data Collection

### Data Understanding

The dataset consists of numerous features that describe the properties.

Key features include:

1. **id**: A unique identifier for each house.
2. **Date**: The date on which the house information was recorded.
3. **number of bedrooms**: The total number of bedrooms in the house.
4. **number of bathrooms**: The total number of bathrooms in the house, including half-bathrooms.
5. **living area**: The total living area (in square feet) of the house.
6. **lot area**: The total area (in square feet) of the lot on which the house is built.
7. **number of floors**: The total number of floors in the house.
8. **waterfront present**: A binary indicator (0 or 1) denoting whether the house has a waterfront view.
9. **number of views**: The number of times the house has been viewed.
10. **condition of the house**: A numerical rating (1 to 5) of the overall condition of the house.
11. **grade of the house**: A numerical rating (1 to 13) of the construction and design quality of the house.
12. **Area of the house (excluding basement)**: The total area (in square feet) of the house, excluding the basement.
13. **Area of the basement**: The total area (in square feet) of the basement.
14. **Built Year**: The year in which the house was originally built.
15. **Renovation Year**: The year in which the house was last renovated (0 if never renovated).
16. **Postal Code**: The postal code where the house is located.
17. **Latitude**: The latitude coordinate of the house's location.
18. **Longitude**: The longitude coordinate of the house's location.
19. **living\_area\_renov**: The living area (in square feet) of the house after renovation.
20. **lot\_area\_renov**: The lot area (in square feet) after renovation.
21. **Number of schools nearby**: The number of schools located within a certain distance from the house.
22. **Distance from the airport**: The distance (in kilometers) from the house to the nearest airport.
23. **Price**: The price of the house in Indian Rupees.



id	number_of_bedrooms	number_of_bathrooms	living_area	lot_area	number_of_floors	waterfront_present	number_of_views	condition_of_the_house	grade_of_the_house	area_of_the_house_excluding_basement	area_of_the_basement	Built_Year	Renovation_Year	Postal_Code	Number_of_schools_nearby	Distance_from_the_airport	Price
1	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1901	0	122000	0	0	100000
2	4	2.5	2000	4000	1.5	0	0	0	0	1000	0	1900	0	122000	2	0	140000
3	1	0.5	1000	2000	1.0	0	0	0	0	500	0	1900	0	122000	1	0	120000
4	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1901	0	122000	0	0	80000
5	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	80000
6	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1901	0	122000	0	0	80000
7	0	0.5	500	1000	1	0	0	0	0	250	0	1900	0	122000	0	0	70000
8	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
9	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
10	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
11	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
12	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
13	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
14	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
15	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
16	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
17	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
18	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
19	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000
20	0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	1900	0	122000	0	0	70000

## 5. Exploratory Data Analysis (EDA)

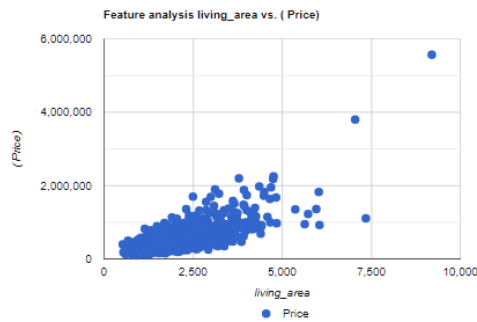
Initial data exploration reveals the following:

- The dataset contains 14620 rows and 23 columns.
- The target variable Outcome is numeric, with indicating house prices in INR.

### Univariate Analysis

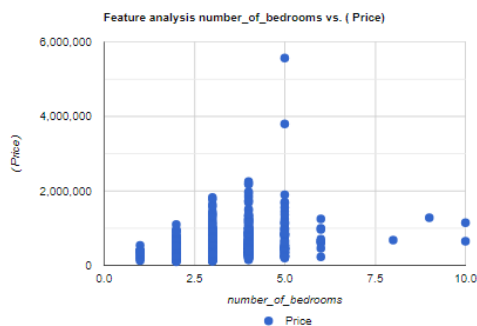
Univariate analysis was performed to understand the distribution of each feature and its potential influence on house prices.

#### Relationship between Living Area and Price (Moderate to strong positive correlation)



There is a visible positive correlation between living area and price; as the living area increases, the price tends to increase as well.

#### Number of Bedrooms vs. Price (Positive correlation)



There is a general trend where properties with more bedrooms tend to have higher prices.

Properties with 10 bedrooms are rare and show a wide price range.

## 6. Data Preparation

### Handling Missing Values

The dataset contains no missing values, but we did some wrangling and normalization to improve accuracy and reduce error rates.

### Data Wrangling

Data wrangling involved cleaning the data, removing outliers, and transforming variables to enhance model performance.



### Deleted columns

1. **Id:** The Id column is not used when building the model.
2. **Date:** The Date column is not used in the model building.
3. **Latitude:** We already have the postal code column, therefore we don't need the latitude.
4. **Longitude:** We already have the postal code column, therefore we don't need the longitude.
5. **Living\_area\_renovation:** There is no correct data; more than 90% of the data is 0 value.
6. **Lot\_area\_renovation:** There is no correct data; more than 90% of the data is 0 value.

## Normalization

Features were normalized to ensure they are on a similar scale, which is crucial for certain machine learning algorithms.



## Standard Scaling

Standard Scaling is a method used to adjust the values of data so that they have an average value of 0 and a spread (how much they differ from the average) of 1.

The formula looks like this:

$$\text{scaled value} = \frac{\text{original value} - \text{average of values}}{\text{standard deviation}}$$

## House Prices:

- House 1: 1,400,000 INR
- House 2: 838,000 INR

### Calculation:

1. **Calculate the Mean ( $\mu$ ):**
  - From the dataset, let's assume the mean price is 791,000 INR.
2. **Calculate the Standard Deviation ( $\sigma$ ):**
  - From the dataset, let's assume the standard deviation is 316,000 INR.
3. **Apply the Standard Scaling Formula:**
  - For House 1 (1,400,000 INR):  $z = \frac{1,400,000 - 791,000}{316,000} \approx 1.93$
  - For House 2 (838,000 INR):  $z = \frac{838,000 - 791,000}{316,000} \approx 0.15$

### Standard Scaled Prices:

- House 1: Approximately 1.93
- House 2: Approximately 0.15

Because the cost of a house exceeds 5 or 6 digits, I used ordinary scaling normalization on price (the goal variable). To achieve a reduced error rate, use conventional scaling, which compresses between 0 and 1 or -1.

Before Normalization

Price
2380000
1400000
1200000
838000
805000
790000
785000

After Normalization

Price
-0.0204968267
-0.5201527058
0.7891060147
-0.8942266259
-0.4827453138
-0.9049144522
0.0275983916



## Inputs and Target Features

All the columns excluding the Price are in Feature Input and the Price column is in Target Variable

Raw Data : house

Column Name...	Feature (Input)	Target (Output)	Data Type	Missing Values	Stat
number_of_floors	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
waterfront_present	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
number_of_views	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
condition_of_the_house	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
grade_of_the_house	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Area_of_the_house_excluding_basement	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Area_of_the_basement	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Built_Year	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Renovation_Year	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Postal_Code	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Number_of_schools_nearby	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Distance_from_the_airport	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	Int
Price	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Num	0	Int

☒ All Input Features
 View Raw Data

Feature Selection : house

Experiments:

Default

Analysis Name : Default

Target (Output)	Problem Type
Price	Regression(-174528.35)

Feature (Input)	Score
living_area	0.31
grade_of_the_house	0.31
Postal_Code	0.12
Built_Year	0.08
lot_area	0.04

Save Analysis as : ☒ Analysis name
 Save

Delete
Find Feature Importance

## Data Splitting

The dataset was split into training and testing sets:

- 80% of the data was used for training.
- 20% of the data was reserved for testing.

Raw Training and Test Dataset

Name...	Training/Test Set	Target	Info	Lock	Delete
price	80 % 20 %	Price			

Generate Cross Validation Dataset

Training and Validation Split

Training Dataset: 80%

Validation Dataset: 20%

☐ Sequence is important (e.g., time series data)

## 7. Methodology

### Model Selection

For this regression task, we used the following algorithms:

#### 1. Random Forest

**Definition:** Random Forest Regression is an ensemble learning method that builds multiple decision trees and merges them together to get a more accurate and stable prediction.

**Formula:**

$$y_{\text{pred}} = \frac{1}{T} \sum_{t=1}^T y_{\text{pred}}^{(t)}$$

Where  $T$  is the number of trees, and  $y_{\text{pred}}^{(t)}$  is the prediction of the  $t$ -th tree.

**Uses:**

- More accurate than individual decision trees.
- Reduces the risk of overfitting.
- Widely used in various domains like finance, healthcare, and marketing.

#### 2. Extra Trees

**Definition:** Extra Trees Regression is similar to Random Forest, but it adds more randomness to the model. It builds multiple trees and splits the data at random points rather than the most optimal points.

**Formula:**

$$y_{\text{pred}} = \frac{1}{T} \sum_{t=1}^T y_{\text{pred}}^{(t)}$$

Where  $T$  is the number of trees, and  $y_{\text{pred}}^{(t)}$  is the prediction of the  $t$ -th tree.

**Uses:**

- Similar benefits to Random Forest, but often faster to train.
- Good for large datasets with many features.

### 3. Decision Tree Regression

**Definition:** Decision Tree Regression is a type of model that predicts the target value by learning simple decision rules inferred from the data features. It splits the data into smaller and smaller subsets and then makes predictions based on the mean value of the target variable in these subsets.

**Formula:**

$$y_{\text{pred}} = \frac{1}{N} \sum_{i=1}^N y_i$$

Where  $N$  is the number of data points in the leaf, and  $y_i$  are the target values.

**Uses:**

- Simple and easy to interpret.
- Used in various applications where the relationships between features and the target variable are complex and non-linear.

#### Model Evaluation

Version-Tag ↕	Dataset ↕	Algorithm ↕	Rank ↕	Error ↕	Doc.	Publish	Delete
<input type="checkbox"/> v.4-v.b87	house-price	RandomForestRegressor	1	0.24			
<input type="checkbox"/> v.11-v.a42	house-price	DecisionTreeRegressor	2	0.33			
<input type="checkbox"/> v.12-v.9ef	house-price	ExtraTreeRegressor	2	0.33			

The performance of the models was evaluated using the following metrics:

**Mean Squared Log Error (MSLE)**

**Mean Absolute Error (MAE)**

**Mean Squared Error (MSE)**

**Median Absolute Error**

Regression	Mean Squared Log Error	MAE	MSE	Median Absolute Error
Random Forest	0.03	0.24	0.4	0.14
Extra Tree	0.06	0.33	0.55	0.19
Decision Tree	0.06	0.33	0.55	0.19

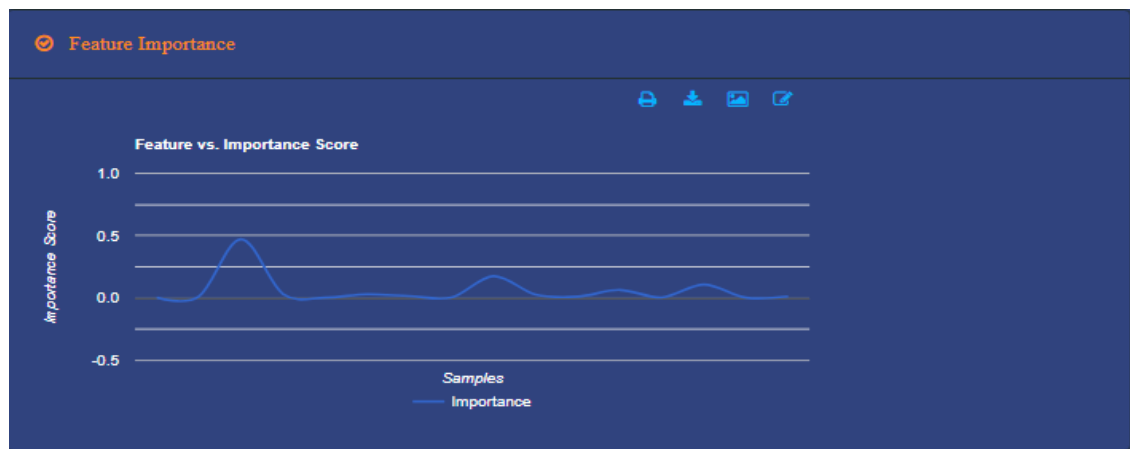
## 8. Best Model

The Random Forest Regressor performed the best with the highest accuracy and less accuracy.



### Feature Importance:

Some features have higher significance scores, indicating a greater impact on price prediction. Understanding feature importance helps in feature selection and improving model performance.



## Model Accuracy:

The metrics like RMSE, MAE, and MSLE provide a quantitative measure of the model's accuracy. The predicted vs. original plot further confirms that the model's predictions are closely aligned with the actual values.



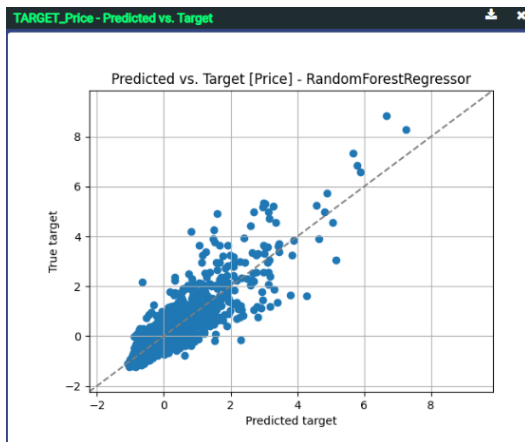
## Performance Metrics:

Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Squared Log Error (MSLE), and Median Absolute Error (MedAE) provide a comprehensive assessment of the model's performance. Lower values indicate better performance.

Performance Metrics
Targets: Price
Mean Squared Error (MSE): 0.40472
Mean Absolute Error : 0.23925
Mean Squared Log Error : 0.03277
Median Absolute Error : 0.13553

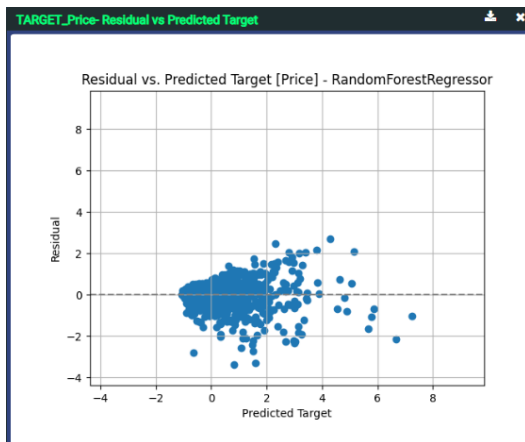
### Predicted vs. Target - Random Forest Regressor

The points closely follow the diagonal line ( $y = x$ ), showing that the predictions are very close to the actual values. This indicates a high level of accuracy in the model's predictions.



### Residual vs. Predicted Target - Random Forest Regressor

Most residuals are clustered around the zero line, indicating that the model predictions are generally accurate. There are few outliers, suggesting the model performs well with minor exceptions.



## Feature Distribution - Training and Validation Set (On Price)

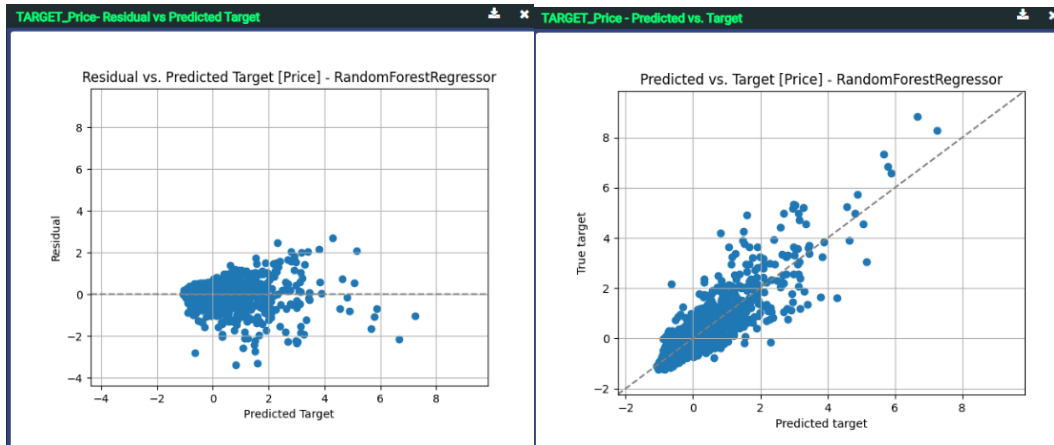
Both distributions are centered on the same range, indicating that the data is well-split and representative across both sets. This ensures that the model can generalize well to unseen data.



## 9. Result (Accept/Reject)

The model was accepted based on its strong performance metrics, less error rate and good accuracy.

Even as we can see below:



Most residuals are clustered around the zero line, indicating that the model predictions are generally accurate.

The points closely follow the diagonal line ( $y = x$ ), showing that the predictions are very close to the actual values.

### What were the results?

Random Forest had the highest accuracy with the lowest MSLE, MAE, and MSE. Feature importance analysis highlighted living area and number of bedrooms are crucial columns.

### How did you evaluate the performance of your model? What metrics did you use?

Performance was evaluated using MSLE, MAE, MSE, and Median Absolute Error, which provided a comprehensive view of model accuracy and error distribution.



## 10. Real-time Prediction

### Example 1:

Demonstrates the real-time prediction capabilities of the Random Forest Regressor for house prices based on various input features.



### Inputs

- **Number of Bedrooms:** 2
- **Number of Bathrooms:** 1
- **Living Area:** 1000 SQ FT
- **Lot Area:** 1800 SQ FT
- **Number of Floors:** 1
- **Waterfront Present:** 0 (No)
- **Number of Views:** 2
- **Condition of the House:** 5
- **Grade of the House:** 9
- **Area of the House Excluding Basement:** 600 SQ FT
- **Area of the Basement:** 0 (No basement)
- **Built Year:** 1956
- **Renovation Year:** 0 (There is no renovation)
- **Postal Code:** 122013
- **Number of Schools Nearby:** 1
- **Distance From The Airport:** 6.1 (in Kilometers)

### Output (Predicted)

- **Model Response:** The predicted standardized house price is approximately 0.535 which can be around 835000 INR.

## Example 2:

Demonstrates the real-time prediction capabilities of the Random Forest Regressor for house prices based on various input features.



## Inputs

- **Number of Bedrooms:** 4
- **Number of Bathrooms:** 3.5
- **Living Area:** 3890 SQ FT
- **Lot Area:** 9000 SQ FT
- **Number of Floors:** 3
- **Waterfront Present:** 0 (No)
- **Number of Views:** 0
- **Condition of the House:** 5
- **Grade of the House:** 13
- **Area of the House Excluding Basement:** 4210 SQ FT
- **Area of the Basement:** 980 SQ FT
- **Built Year:** 2001
- **Renovation Year:** 0 (There is no renovation)
- **Postal Code:** 122005
- **Number of Schools Nearby:** 3
- **Distance From The Airport:** 5.6 (in Kilometers)

## Output (Predicted)

- **Model Response:** The predicted standardized house price is approximately 1.892 which can be around 1240000 INR.

## 11. Conclusion

### Implementation and Usage of the Solution

#### How could an organization or institution implement/use the solution?

Organizations and institutions can leverage this house price prediction model in several impactful ways:

1. **Real Estate Agencies:** Can integrate the model into their platforms to provide accurate price estimates for properties, enhancing transparency and trust with clients.
2. **Financial Institutions:** Banks and mortgage lenders can use the model to assess the value of properties when approving loans, reducing the risk of overvaluation.
3. **Government and Urban Planners:** Can utilize the model to understand housing market trends and make data-driven decisions for urban development projects.
4. **Property Investors:** Investment firms can use the predictions to identify undervalued properties and make informed investment decisions.

#### Implementation Steps:

- **Data Integration:** Incorporate the model into existing databases and ensure seamless data flow for real-time predictions.
- **API Deployment:** Develop APIs to allow different systems and platforms to access the model's predictions.

#### Improvements for the Future:

1. **Incorporate Additional Features:** Adding more features like economic indicators, crime rates, and school quality could improve model accuracy.
2. **Advanced Techniques:** Using advanced machine learning techniques like ensemble learning, deep learning, or hybrid models could enhance prediction capabilities.
3. **Regular Updates:** Continuously updating the model with new data will help maintain its accuracy and relevance.
4. **User Feedback:** Incorporating user feedback can help refine the model and the prediction process.

## Lessons Learned and Improvement Areas

### Lessons Learned:

1. **Good Data Matters:** The quality and completeness of data are crucial for building accurate predictive models. Missing values, outliers, and irrelevant features can significantly impact model performance.
2. **Feature Engineering is important:** Creating new features and selecting the most relevant ones play a vital role in enhancing model accuracy. Understanding the domain and the data helps in identifying key features.
3. **Choosing the Right Model:** Different models have different strengths. Random Forest Regressor was chosen for its accuracy and robustness, but it's essential to compare multiple models.
4. **Interpreting Results:** It's not enough to just build a model; interpreting its results and understanding the feature importance are crucial for actionable insights.
5. **Practical Use:** Ensuring the model can be used in real-world applications involves considering factors like response time, scalability, and ease of integration.