



Capstone Project on Machine Learning

Presentation by – Natesh Kumar

Introduction to Dataset

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.

Business Understanding

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

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 Collection

Data Understanding:

The dataset consists of numerous features that describe the properties.

Key features include: ID, Date, no. of bedrooms, no. of bathrooms, living area, lot area, no. of floors, waterfront present, number of views condition of the house, grade of the house, area of the house and basement, Built year, renovation year, Postal code, Latitude and Longitude, Living area, lot area, Schools nearby Distance from the airport and Price.

Exploratory Data Analysis (EDA)



Initial data exploration reveals the following:

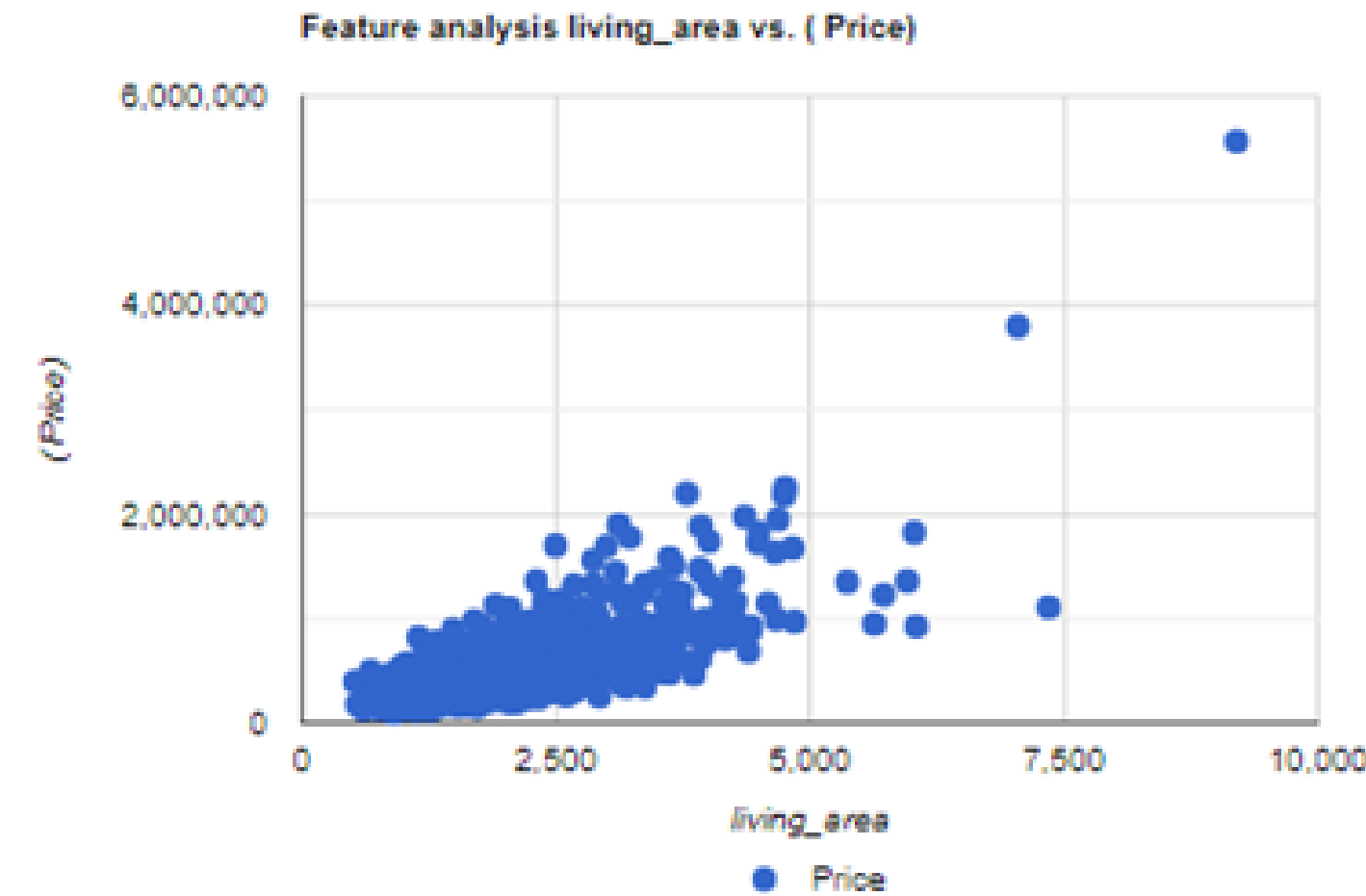
- The dataset contains 14620 rows and 23 columns.
- The target variable Outcome is numeric, with indicating house prices in INR.

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

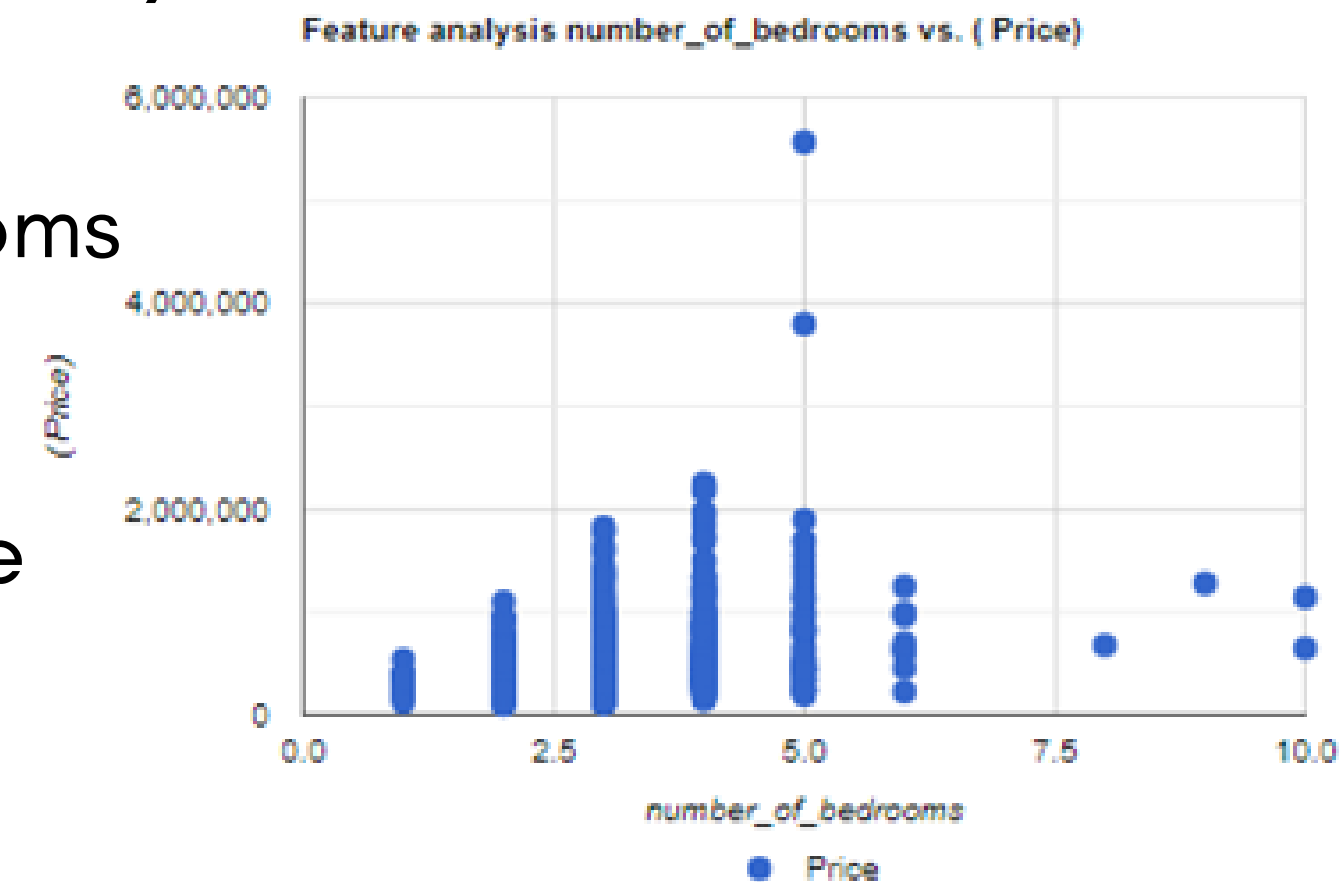
There is a visible positive correlation between living area and price; as the living area increases, the price tends to increase as well.



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.

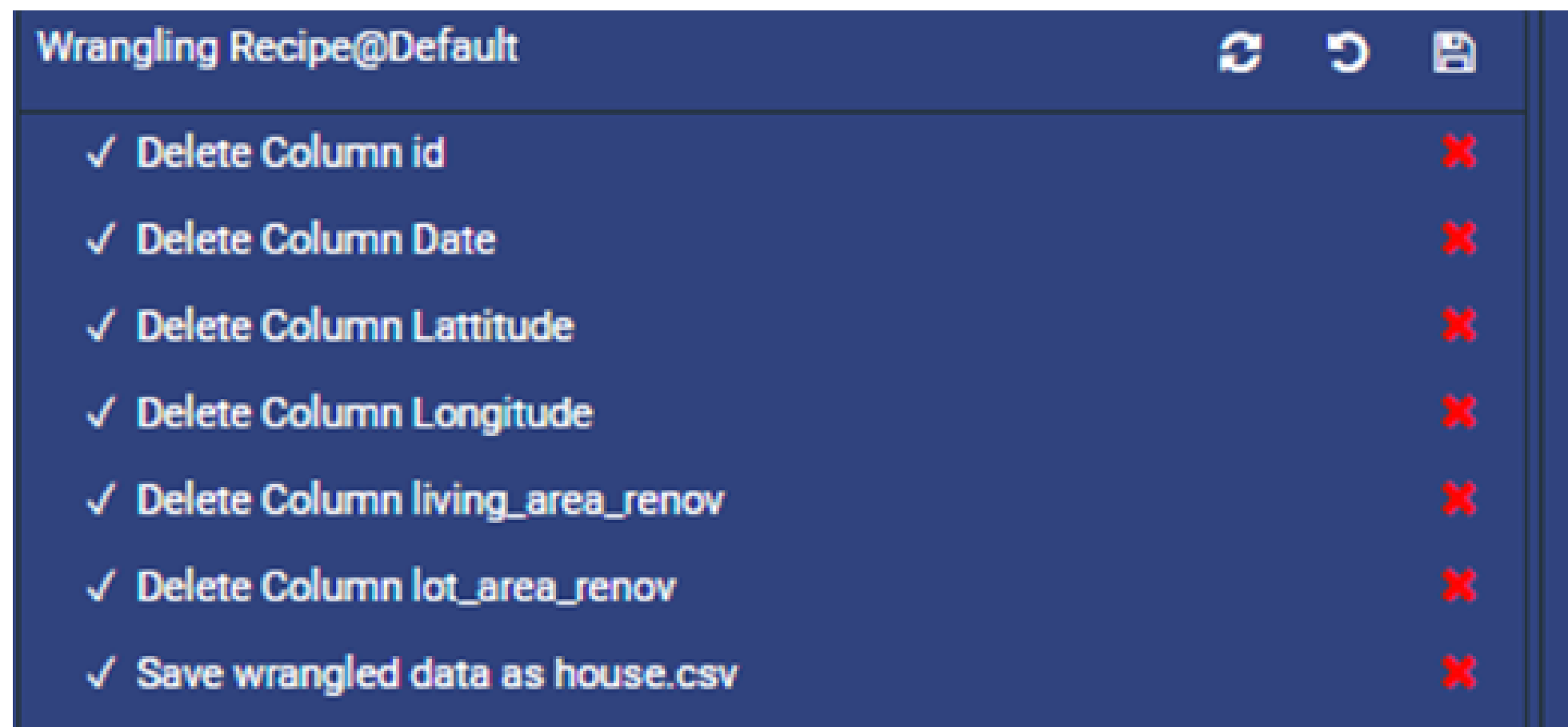


Data Preparation

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 latitude.
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:

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

Example

House Prices:

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

Calculation:

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

Example

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

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:

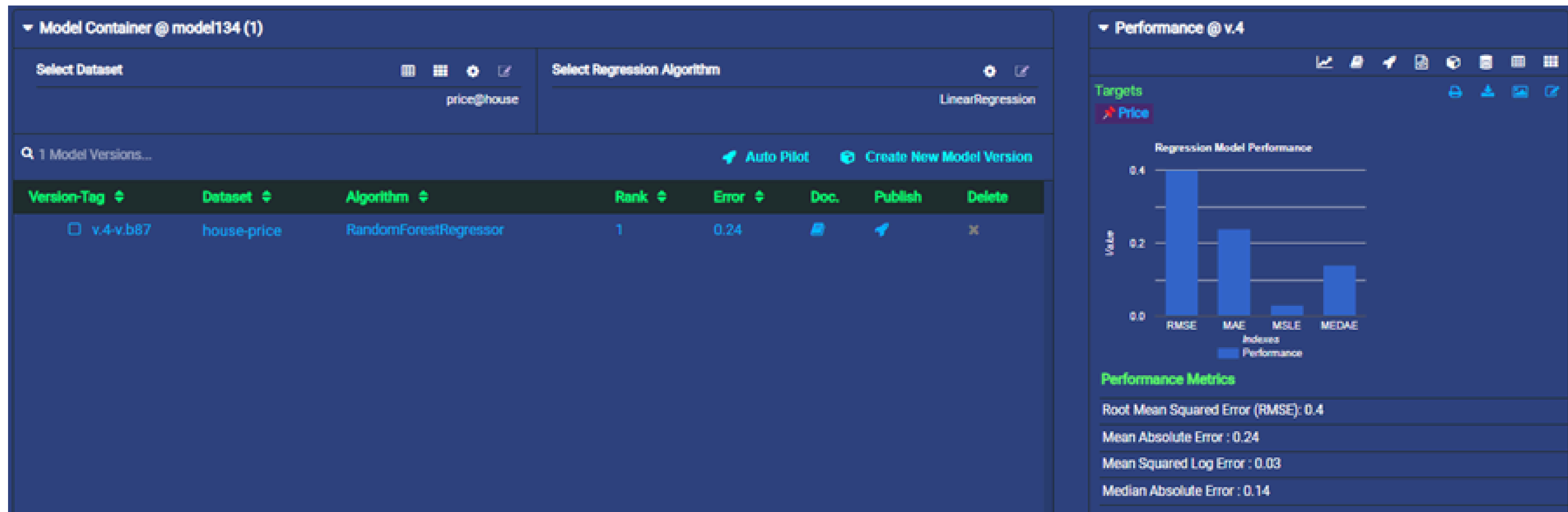
Mean Squared Log Error (MSLE), Mean Absolute Error (MAE)

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

Best Model

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



Model Accuracy:

The predicted vs. original plot confirms that the model's predictions are closely aligned with the actual values.



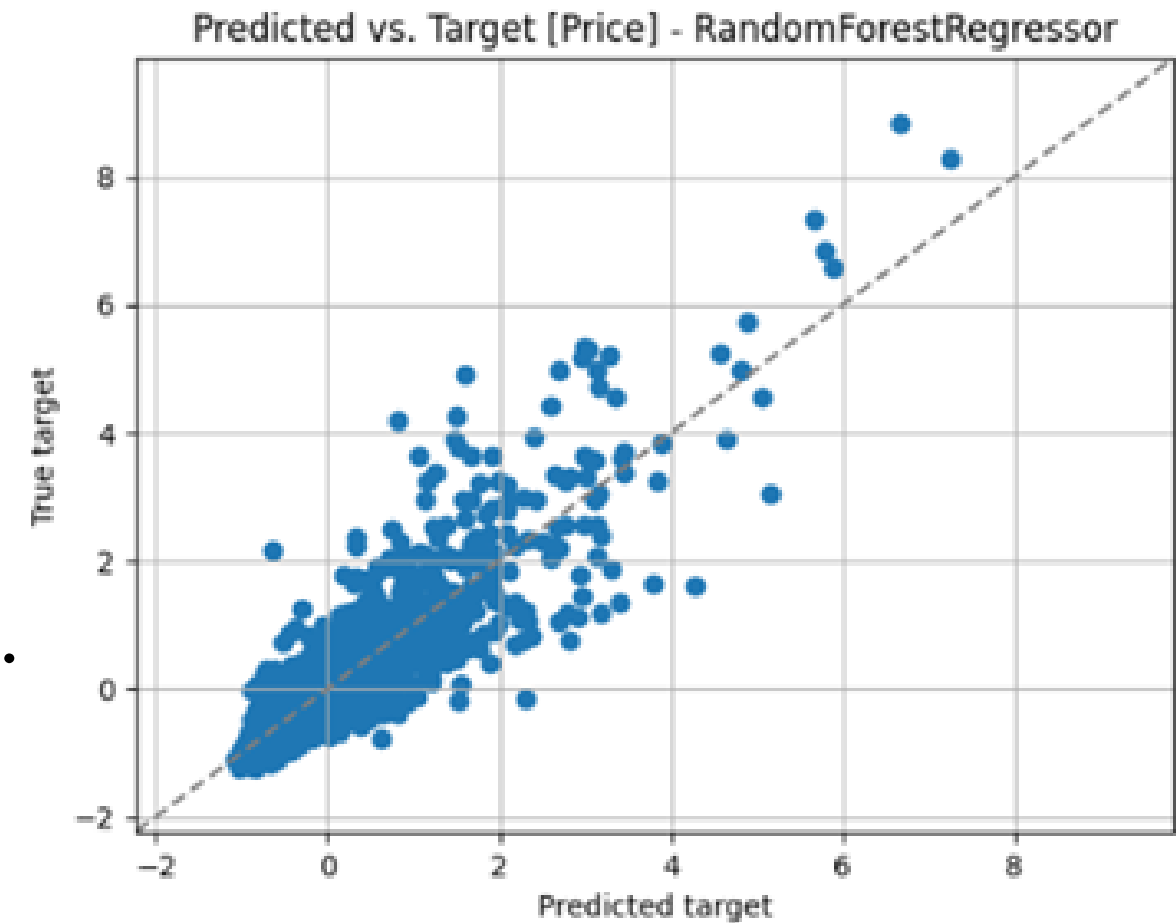
Performance Metrics:

- The model predictions are accurate, which means better decision-making and reduced financial risk.
- The model is likely to generalize well to new data, making it reliable for real-world applications.

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

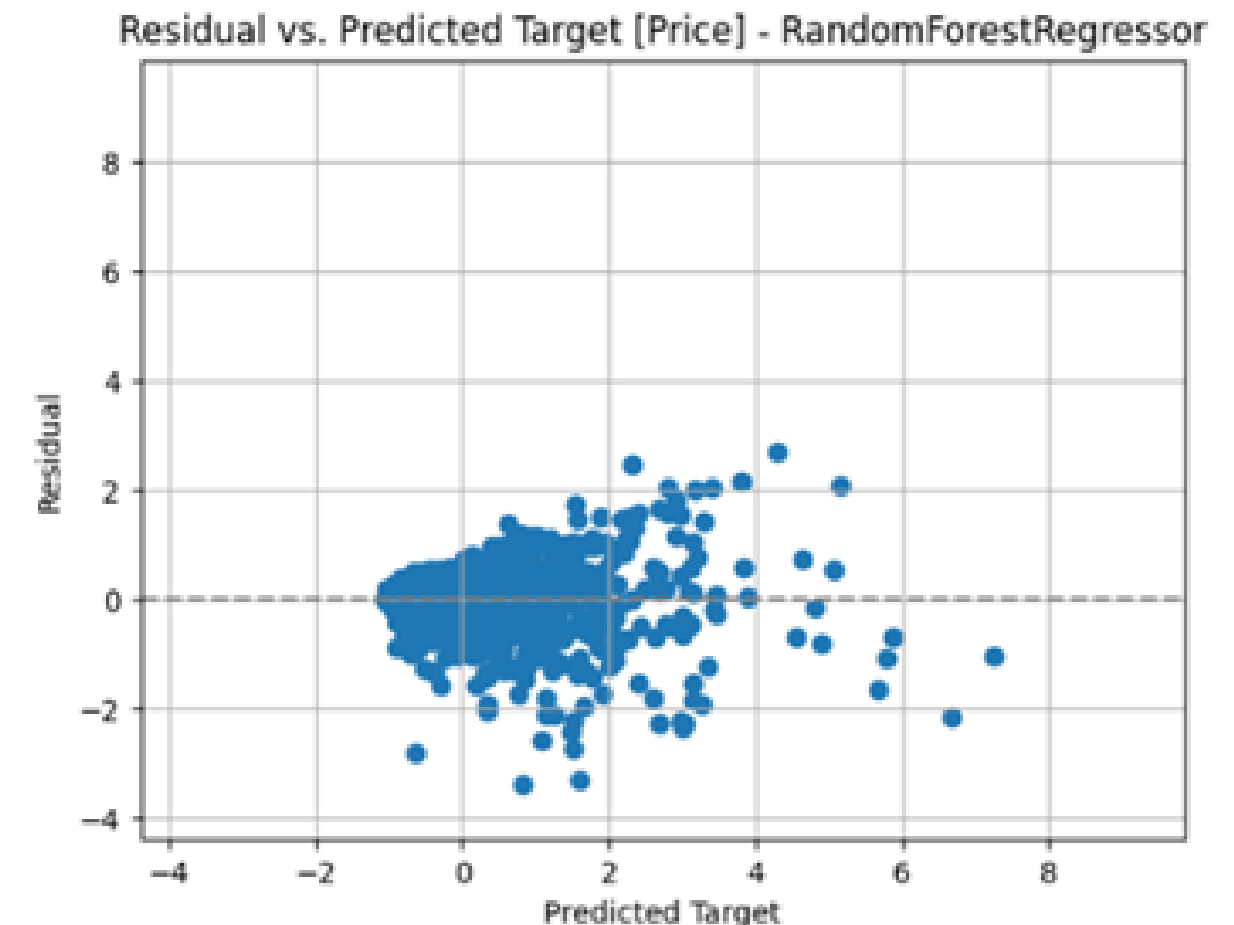
Predicted vs. Target - Random Forest Regressor

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.



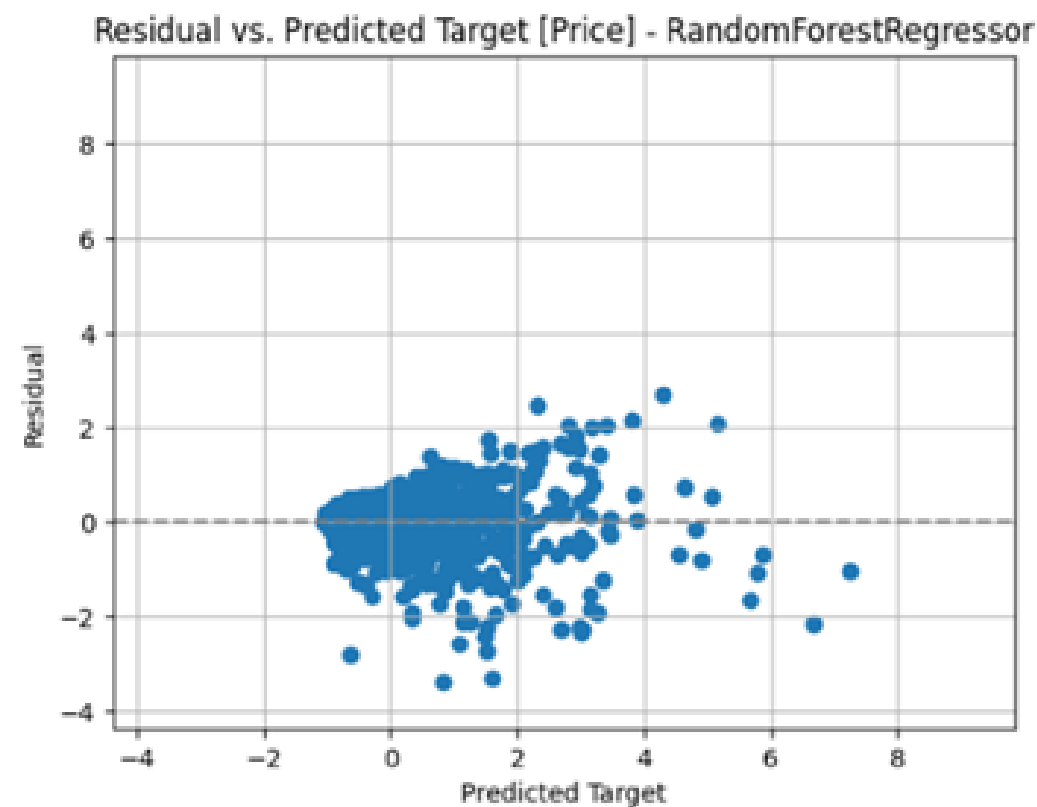
Residual vs. Predicted Target - Random Forest Regressor

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.



Model Governance

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.

Real Time Prediction

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



By setting up some random inputs here as we can see in image as well our model was running fine and gave an output.

Output (Predicted)

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

Conclusion

Implementation and Usage of the Solution

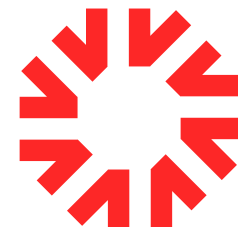
How could an organization or institution implement/use the solution?

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.

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.



Thank You