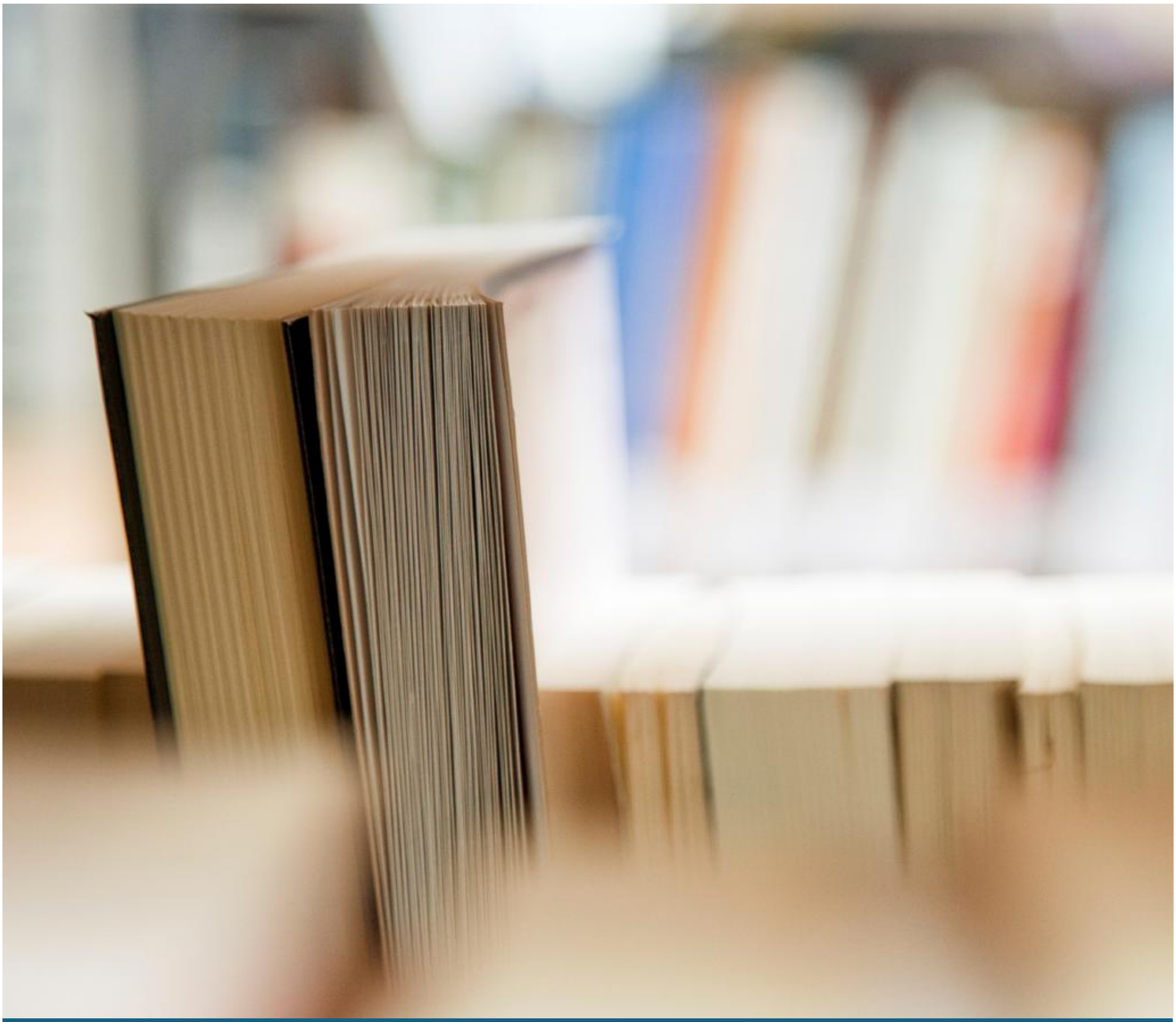


Supervised Learning - Regression

Team Nexus

Abhishek

Meit

Mitul

Ramkumar

3/8/25

Supervised Learning 046

Paris Housing Dataset – Problem

1. **What problem are you solving?**

- The goal of this analysis is to predict house prices in Paris based on various property features such as size, number of rooms, amenities, and location. Accurate price predictions can help buyers, sellers, and real estate agents make informed decisions. The dataset consists of property listings with attributes including square meters, number of rooms, presence of amenities like pools and garages, and historical data such as the year built and number of previous owners.

2. **Why is it worth solving?**

- Real estate pricing is crucial for market efficiency. Overpricing can lead to long sale durations, while under-pricing results in financial loss. A predictive model can assist stakeholders in making data-driven decisions, improving transparency, and optimizing pricing strategies. This benefits homeowners, buyers, investors, and policymakers.

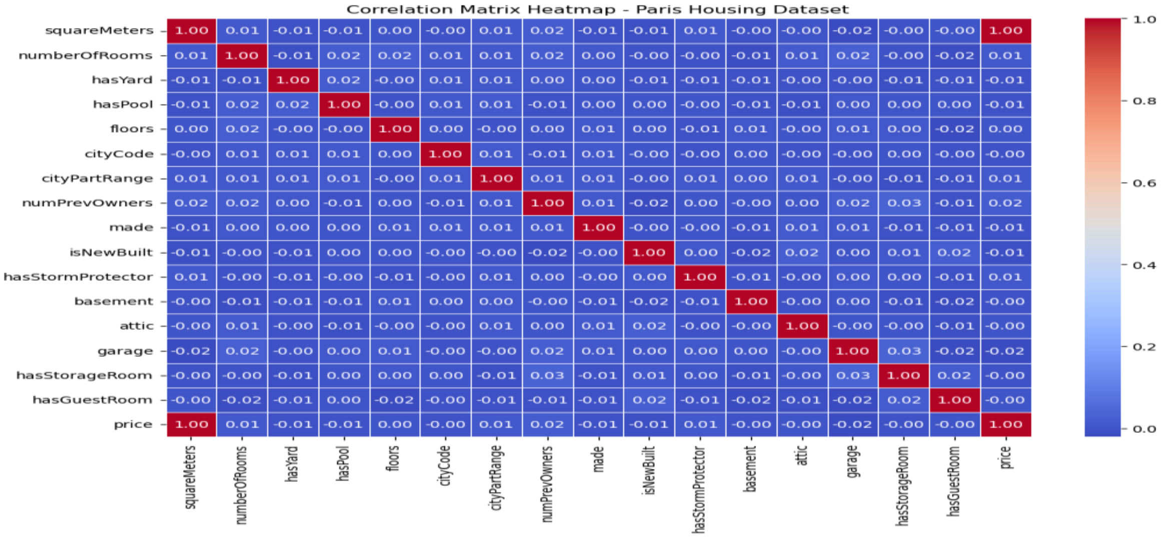
3. **What is the source of your data and what kinds of data are you using?**

- The dataset originates from real estate listings in Paris. It includes structured numerical data such as square meters, number of rooms, and property features (e.g., presence of a pool, garage, or storage room). Additionally, historical and categorical attributes such as the year built, number of previous owners, and city codes are included to enhance prediction accuracy.

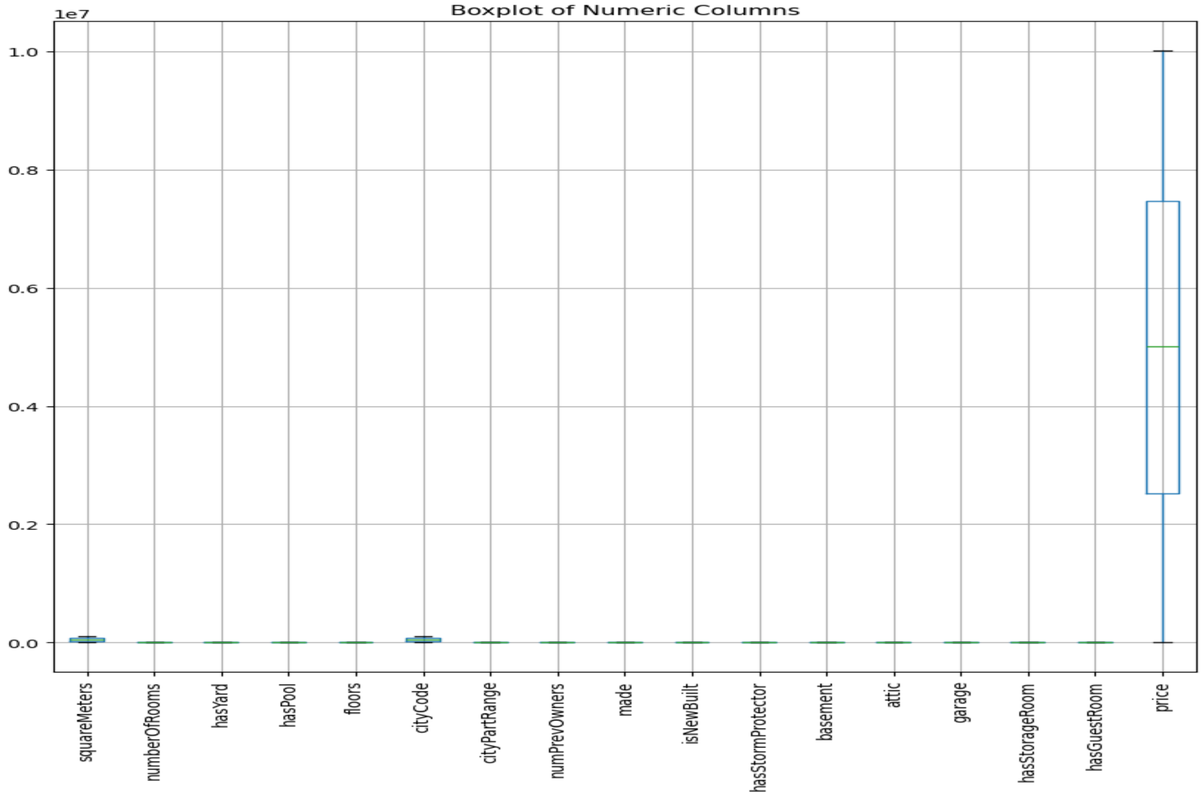
Methodology

1. **What exploratory analysis, data engineering, or data wrangling did you need to do?**

- A **correlation heatmap** was used to identify relationships between features and the target variable (price). Square meters, number of rooms, and number of floors showed strong correlations.
- No missing values were found, eliminating the need for imputation.
- Numerical features were standardized to ensure better model performance.



- The **box plot** below provides insights into the distribution of house prices and helps identify potential outliers

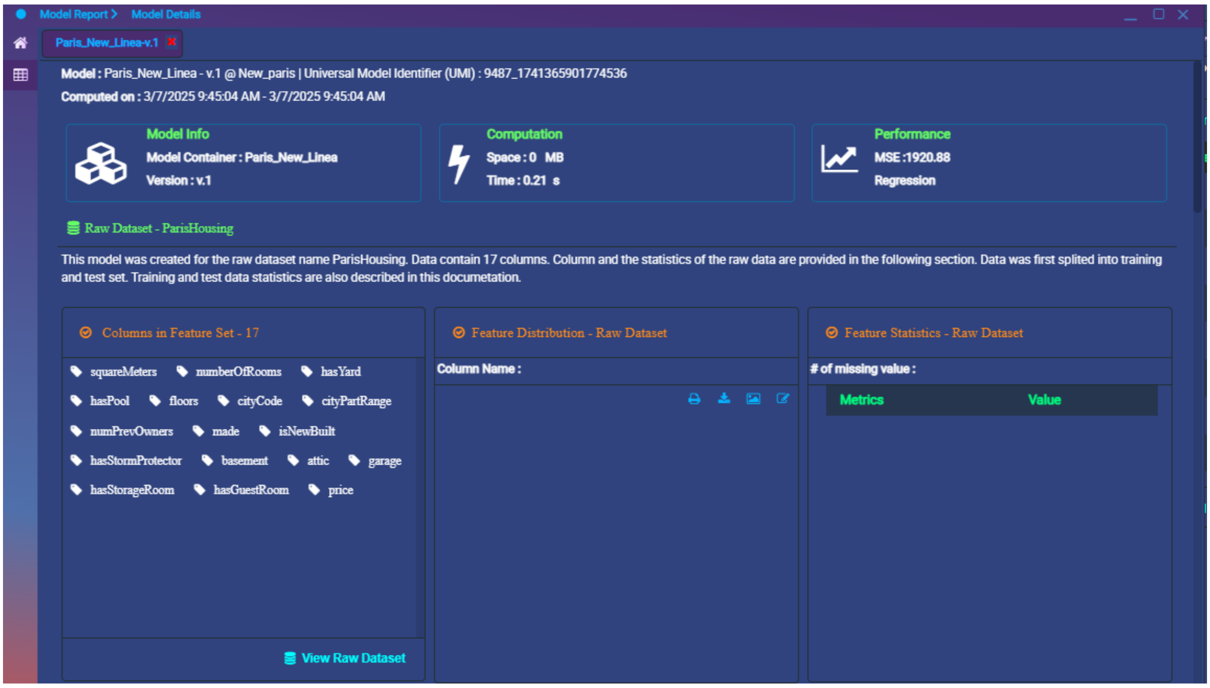


2. **How did you prepare the data for modelling?**

- Standardized numerical variables to bring all features to a similar scale.
- Split the dataset into **80% training** and **20% testing** to assess model performance.

3. **What was your modelling process? Specifically, which algorithms and parameters did you use and why?**

- Used Linear Regression as it effectively captures the linear relationship between house price and property attributes.
- Linear Regression was chosen due to its interpretability and strong correlation patterns in the dataset.
- Evaluated the model using regression metrics such as MAE, MSE, RMSE, and R^2 score.



Model : Paris_New_Linea - v.1 @ New_paris | Universal Model Identifier (UMI) : 9487_1741365901774536

Computed on : 3/7/2025 9:45:04 AM - 3/7/2025 9:45:04 AM

Model Info
Model Container : Paris_New_Linea
Version : v.1

Computation
Space : 0 MB
Time : 0.21 s

Performance
MSE :1920.88
Regression

Raw Dataset - ParisHousing

This model was created for the raw dataset name ParisHousing. Data contain 17 columns. Column and the statistics of the raw data are provided in the following section. Data was first splited into training and test set. Training and test data statistics are also described in this documetation.

Columns in Feature Set - 17

squareMeters numberOfRooms hasYard
hasPool floors cityCode cityPartRange
numPrevOwners made isNewBuilt
hasStormProtector basement attic garage
hasStorageRoom hasGuestRoom price

Feature Distribution - Raw Dataset

Column Name :

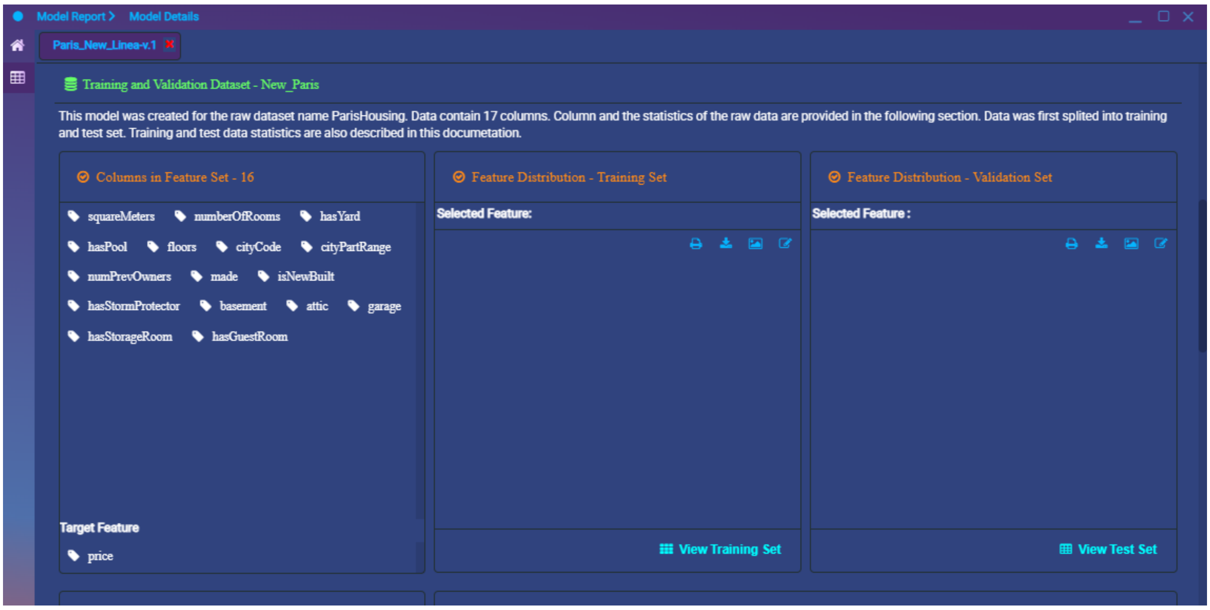
Feature Statistics - Raw Dataset

# of missing value :

Metrics	Value

Modelling Process: Algorithms and Parameters

- **Algorithm Used**: Linear Regression



Training and Validation Dataset - New_Paris

This model was created for the raw dataset name ParisHousing. Data contain 17 columns. Column and the statistics of the raw data are provided in the following section. Data was first splited into training and test set. Training and test data statistics are also described in this documetation.

Columns in Feature Set - 16

squareMeters numberOfRooms hasYard hasPool floors cityCode cityPartRange numPrevOwners made isNewBuilt hasStormProtector basement attic garage hasStorageRoom hasGuestRoom

Target Feature

price

Feature Distribution - Training Set

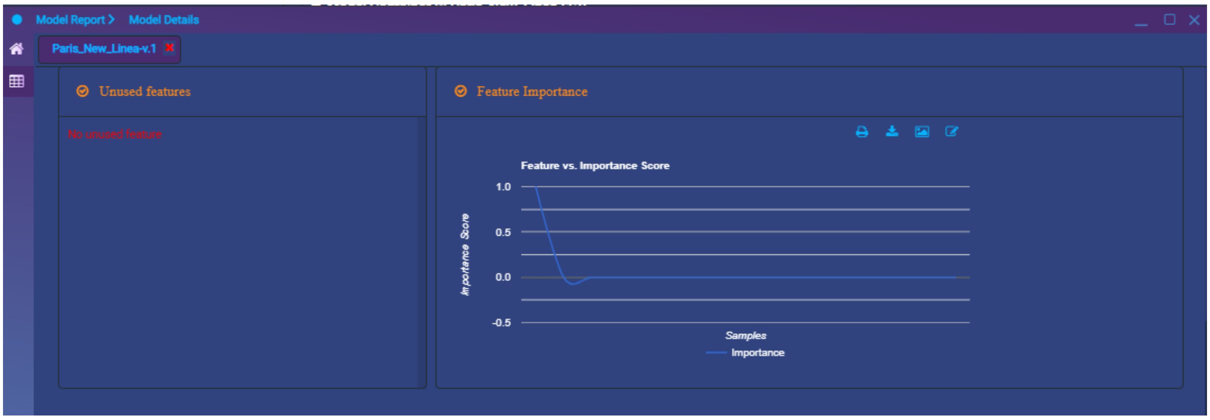
Selected Feature:

View Training Set

Feature Distribution - Validation Set

Selected Feature :

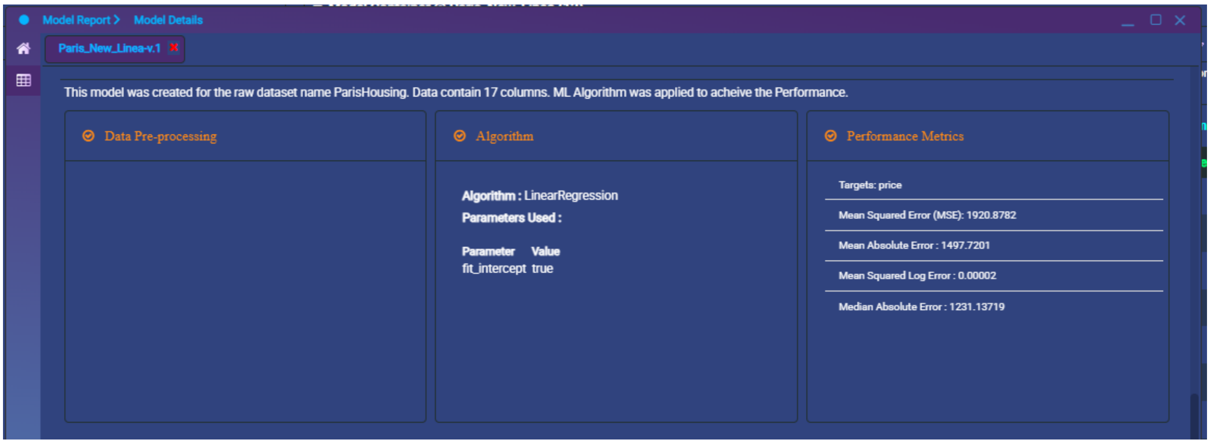
View Test Set



Unused features

No unused feature

Feature Importance



This model was created for the raw dataset name ParisHousing. Data contain 17 columns. ML Algorithm was applied to acheive the Performance.

Data Pre-processing

Algorithm

Algorithm : LinearRegression

Parameters Used :

Parameter	Value
fit_intercept	true

Performance Metrics

Targets: price

Mean Squared Error (MSE): 1920.8782

Mean Absolute Error : 1497.7201

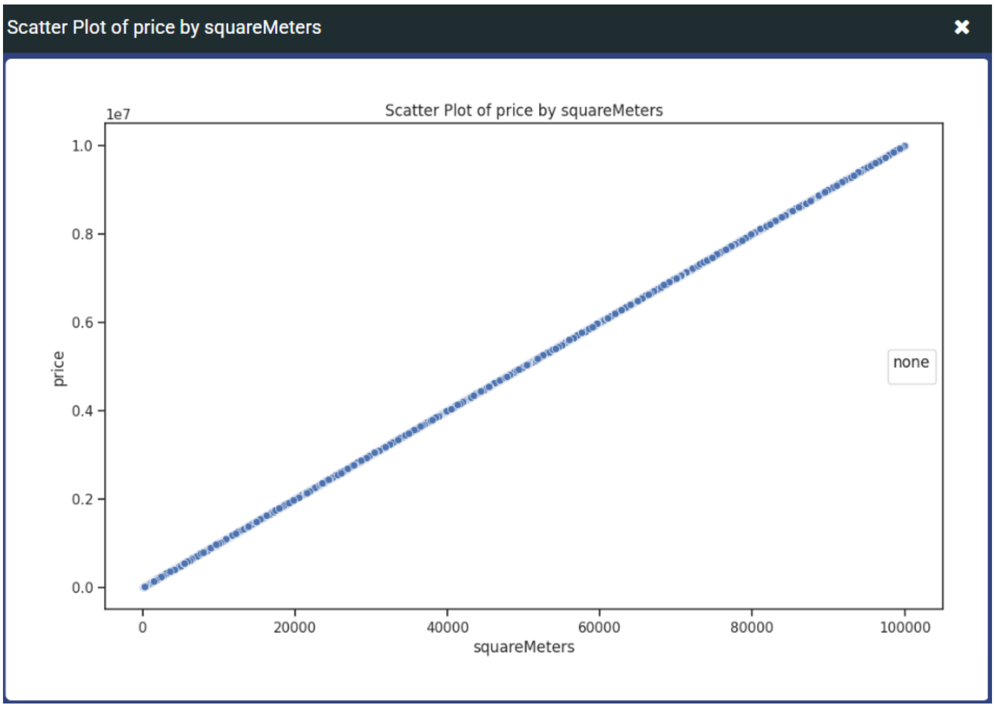
Mean Squared Log Error : 0.00002

Median Absolute Error : 1231.13719

Model Results

- **Mean Squared Error (MSE)**: 1920.8782
- **Mean Absolute Error (MAE)**: 1497.7201
- **Mean Squared Log Error (MSLE)**: 0.00002
- **Median Absolute Error**: 1231.13719

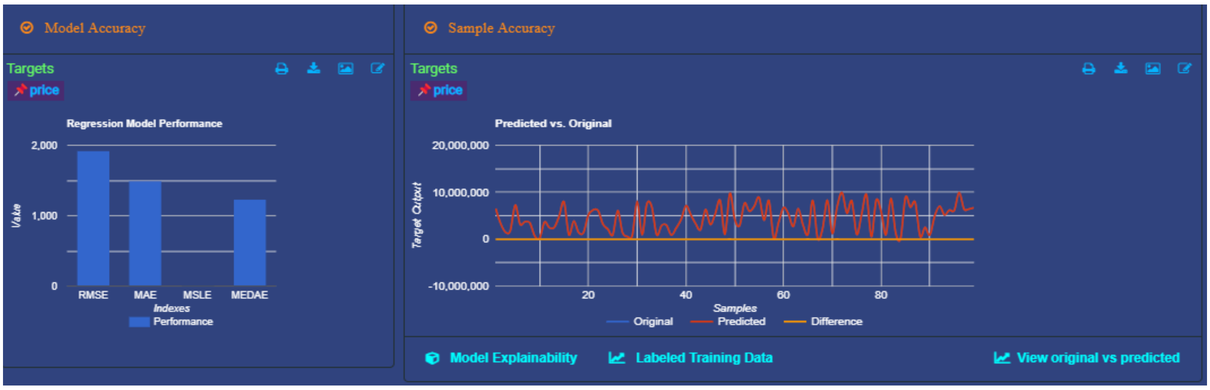
The scatter plot below shows the relationship between square meters and price, confirming the strong correlation:



Model Performance Evaluation

The performance was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Mean Squared Log Error (MSLE)**
- **Median Absolute Error**



Conclusions

1. **Future Improvements**:

- Experimenting with advanced models like Random Forest or XGBoost to compare results.
- Incorporating categorical variables such as cityCode using encoding techniques.
- Feature selection to reduce multicollinearity and improve interpretability.

2. **Real-World Applications**:

- Real estate agencies can use this model to estimate house prices based on property attributes.
- Buyers and sellers can leverage this tool for fair price negotiations.
- Investors can assess property value trends in different city parts.

3. **Business Value to Client**:

- Reduces the risk of overpricing or under-pricing properties.
- Helps in urban planning and real estate development decisions.

4. **Key Takeaways**:

- Square meters, number of rooms, and number of floors are primary drivers of house prices.
- Linear Regression performed exceptionally well, likely due to the strong linear relationships in the dataset.
- Data preprocessing, particularly feature scaling, played a crucial role in improving model performance.