

Multimodal Framework for Satellite Imagery-Based Property Valuation

1. Overview: Approach and Modeling Strategy

This project implements a multimodal deep learning framework to estimate residential property values by fusing traditional structural attributes with geospatial environmental context. By integrating high-resolution satellite imagery, the model captures latent value drivers—such as canopy cover and infrastructure density—that are frequently omitted in standard regression analyses.

Pipeline & Methodology

- **Data Acquisition:** A custom utility fetches 224×224 satellite patches via the Mapbox Static API, synchronized with property coordinates.
- **Feature Engineering:** Data preprocessing includes :**log-transforming** prices to normalize skewness (reduced from 1.71 to 0.43), calculating property "age" based on sale dates, etc.
- **Spatial Analysis:** Models are enriched with engineered features like `dist_from_center` to capture location premiums.
- **Vision Branch:** An **EfficientNetB0** backbone extracts high-dimensional visual embeddings from satellite imagery.
- **Gated Fusion:** A learned sigmoid mechanism dynamically weights the importance of visual versus tabular data for each specific property.

Modeling & Optimization

The system utilizes a two-phase training strategy:

- **Phase 1:** Freezes the CNN backbone to optimize the fusion layers and regression head.
- **Phase 2:** Unfreezes top CNN layers for fine-tuning, allowing the model to adapt to real-estate-specific visual cues.
- **Loss Function:** Employs **Huber Loss** to ensure robustness against market outliers.

2. Grad-CAM(visuals and financial insights):

To interpret how satellite imagery influences property valuation, **Gradient-weighted Class Activation Mapping (Grad-CAM)** was applied to the trained multimodal model. This provides visual explanations by highlighting spatial regions in satellite images that most strongly affect the predicted price.

2.1 Methodology:

- The trained multimodal model receives both the satellite image and corresponding tabular features for a given property.
- Gradients of the predicted log-price are computed with respect to the final convolutional feature maps of the EfficientNetB0 backbone.
- These gradients are globally averaged to obtain importance weights for each feature map.
- A weighted sum of the feature maps is passed through a ReLU activation to generate a localization map.
- The resulting heatmap is upsampled and overlaid on the original satellite image, ensuring alignment with the true multimodal prediction.

2.2 Heatmap Color Interpretation:

Color	Meaning
Red / Yellow	Strong influence → Higher predicted price
Green	Moderate contribution
Blue	Low or negligible contribution

Important: Colors show *relative importance*, not absolute monetary value.

2.3 Key Observations:

2.3.1 Neighborhood Context Outweighs Individual Structures:

Key visual cues include:

- Road networks and connectivity patterns
 - Plot openness and available land area
 - Presence of vegetation and tree cover
 - Spatial separation between adjacent properties
- ➔ indicates that the model learns **neighborhood-level valuation signals**, which are fundamental drivers of real estate pricing.

2.3.2 Environmental Features Act as Primary Value Indicators:

Rather than emphasizing rooftops alone, the CNN captures broader environmental context such as:

- Accessibility and layout of surrounding roads
- Degree of openness and congestion
- Quality and organization of nearby infrastructure

➔ These features align closely with the intended role of satellite imagery—providing contextual information that complements structural property data.

2.3.3 Prediction Deviations Are Interpretable:

Discrepancies between predicted and actual prices can often be attributed to factors not observable from satellite imagery, including:

- Interior finishes and renovation quality
- Market-driven negotiation dynamics
- Temporal, regulatory, or policy-related influences

➔ Despite these limitations, Grad-CAM explanations remain **consistent and coherent**, suggesting that the model has learned stable and meaningful visual patterns rather than spurious correlations.

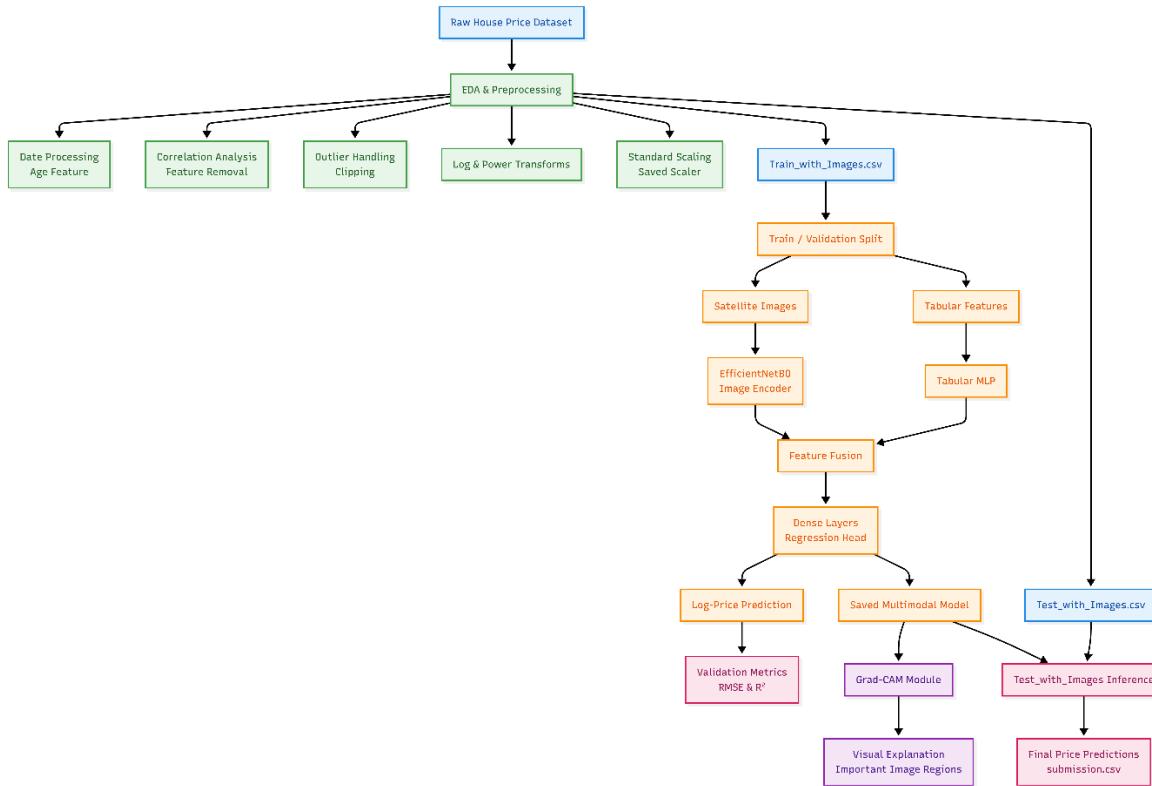


corresponds to a property with a significantly higher predicted price (₹572,629). Here, the Grad-CAM visualization highlights a much broader spatial context, including extensive tree cover, open spaces, and the surrounding road layout. The strong activation over greenery and neighborhood structure suggests that the model associates these environmental characteristics with higher property value. This demonstrates that the model is not only focusing on the building but also capturing neighborhood quality and visual appeal.



shows a property with a relatively lower predicted price (₹397,009). The Grad-CAM heatmap primarily concentrates around the immediate building footprint and limited surrounding greenery. The activation is localized and comparatively weaker, suggesting that the model relies mainly on the house structure itself, with minimal contribution from broader neighborhood features. This indicates a denser or less visually distinctive environment, which aligns with the lower predicted valuation.

3. Flow Chart Diagram:

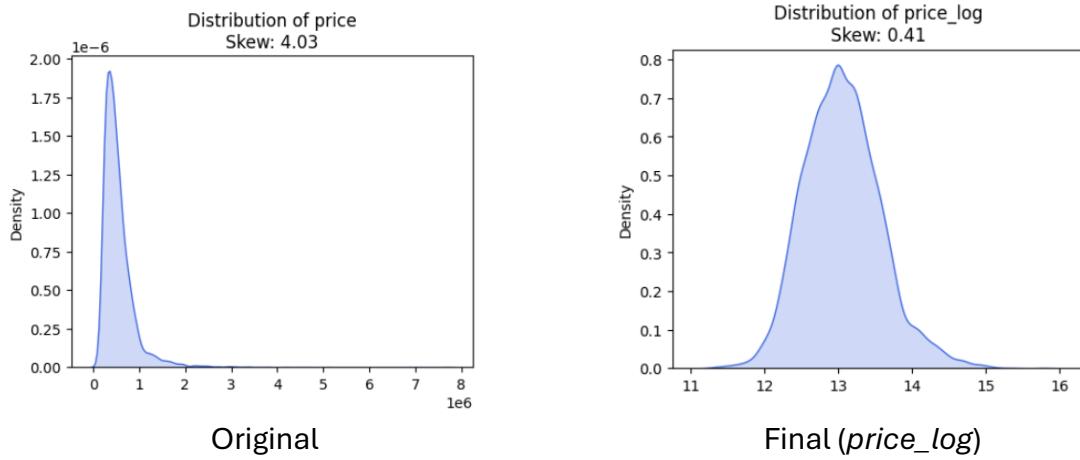


4. Exploratory Data Analysis(EDA):

4.1 Price Distribution and Target Transformation:

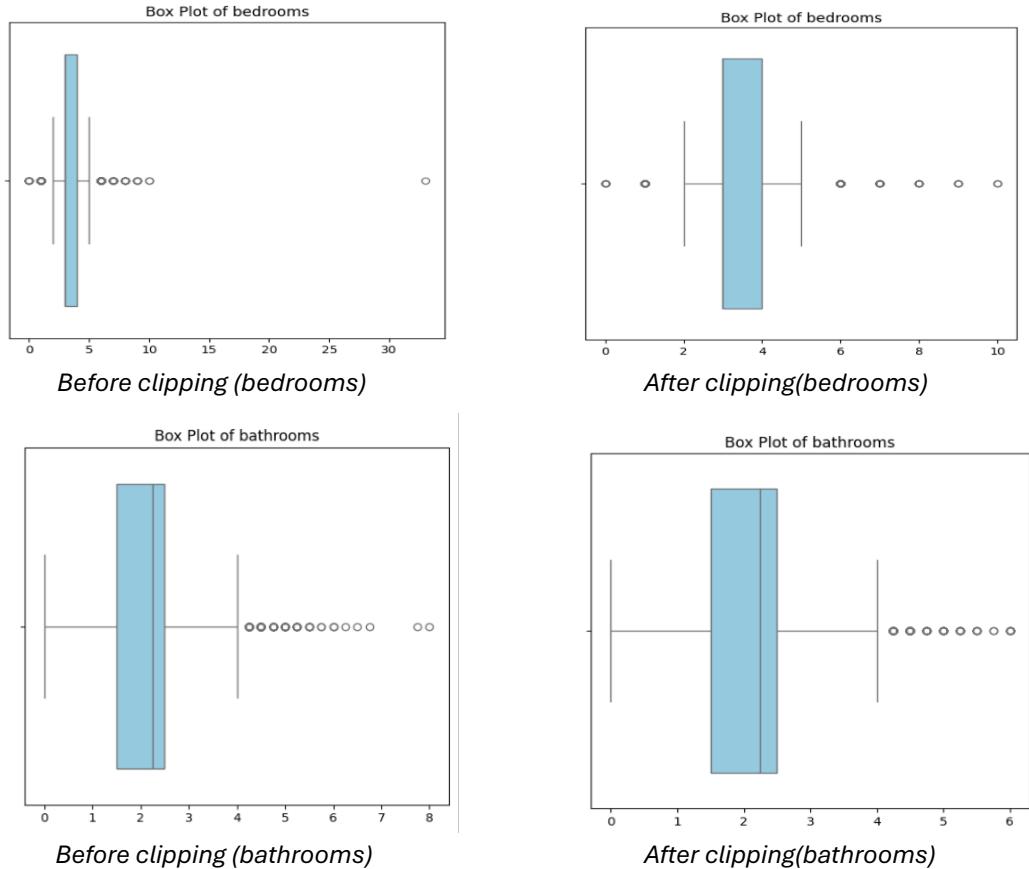
The raw price distribution exhibited strong **right skewness**, indicating the presence of high-value outliers. To stabilize variance logarithmic transformation was applied to price. Skewness was reduced significantly (from heavy right skew to near-normal distribution).

This transformation improves regression stability and ensures better error behavior during training.



4.2 Handling Skewness and Outliers:

- Outliers were handled using domain-aware clipping, where extreme values in bedrooms and bathrooms were capped at reasonable upper limits to reduce the influence of anomalous observations without removing data



→ Area-related features (`sqft_lot`, `sqft_lot15`, `sqft_basement`, etc) showed non-normal distributions and were normalized using the Yeo-Johnson power transformation to reduce skewness while preserving zero values.

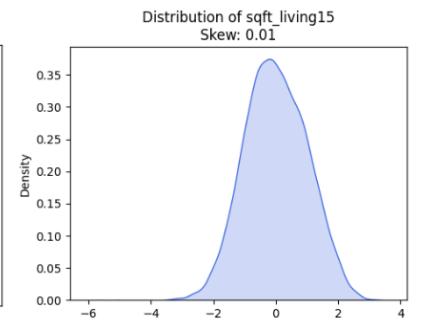
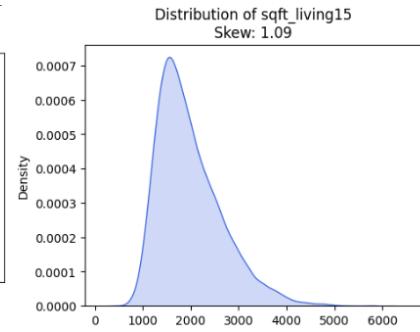
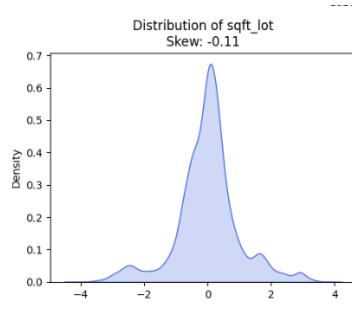
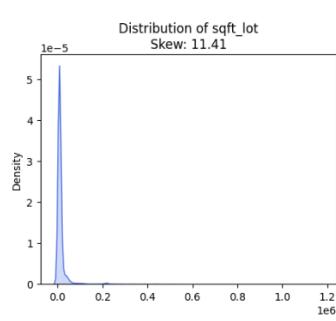
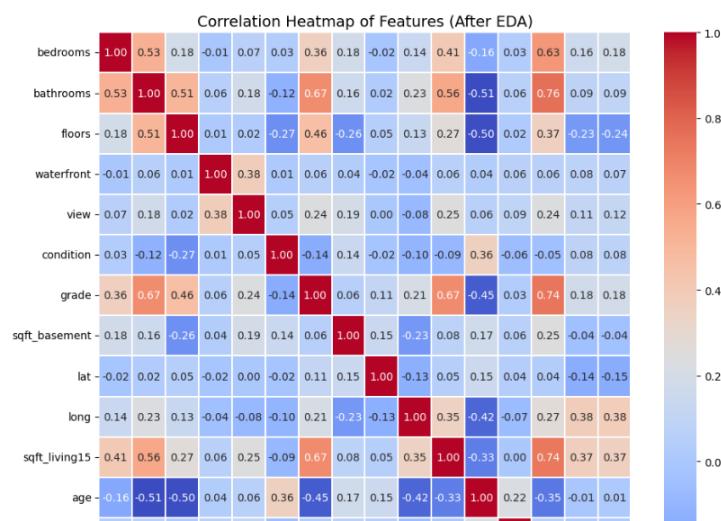
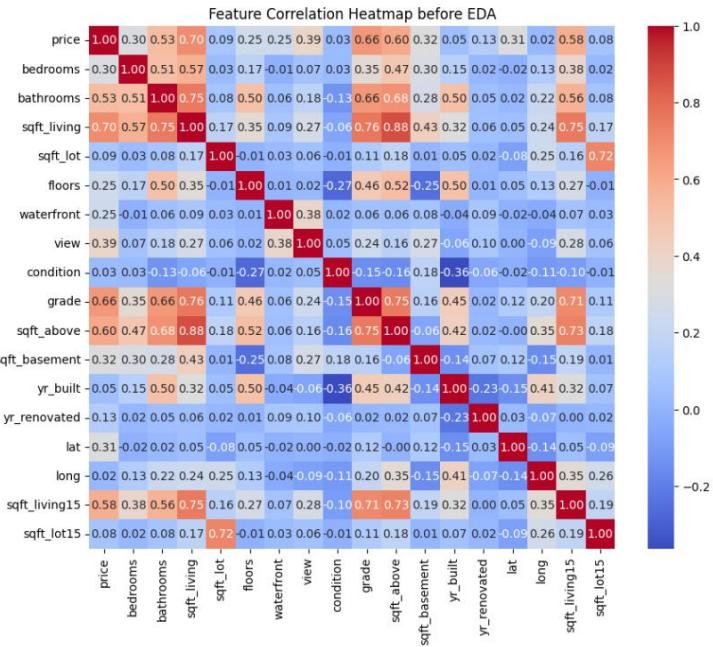


Figure 1: Left: before transformation
(`sqft_lot`) Right: after transformation

Figure 2: Left: before transformation
(`sqft_living15`) Right: after transformation

4.3 Heatmap and its interpretation:

The initial correlation heatmap reveals significant multicollinearity among raw features, particularly between size-related variables like `sqft_living`, `sqft_above`, and `sqft_living15`, as well as `grade`. This redundancy with the target variable (price) increases the risk of model instability and overfitting. Furthermore, the weak correlations exhibited by raw spatial(`lat`, `long`) and temporal (`yr_builtin`, `yr_renovated`) attributes suggest that these features, in their original form, fail to capture critical location or age-related trends effectively.

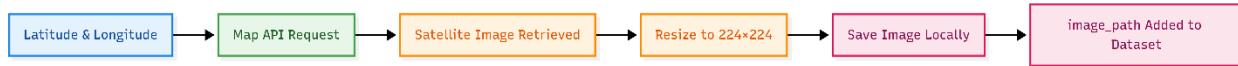


The post-EDA correlation heatmap illustrates relationships among numerical features after handling skewness, outliers, and redundant variables. Size- and quality-related attributes such as `sqft_living_log`, `bathrooms`, and `grade` show strong positive correlations, confirming their central role in property valuation. Log-transformed area features

(sqft_lot_log and sqft_lot15_log) remain highly correlated, reflecting shared spatial characteristics with improved distributional stability. The engineered age feature exhibits negative correlations with several size and quality indicators, capturing depreciation effects in older properties. Location-based features (lat and long) display moderate correlations, indicating spatial pricing patterns. Features such as renovated contribute complementary information with weaker but meaningful relationships. Overall, the correlation structure is more balanced than in the raw data, demonstrating reduced multicollinearity and improved feature conditioning for downstream modeling.

Although bedrooms and bathrooms exhibit a relatively high correlation, neither feature was removed during preprocessing. Both variables independently represent critical aspects of a house's functional utility and are well-known determinants of property value. Removing either feature could result in loss of meaningful information relevant to buyer preferences and pricing dynamics. Therefore, both were retained in the final feature set despite their correlation.

5. Satellite Image Acquisition and Integration:



Workflow for satellite image acquisition and integration. Geographic coordinates are used to retrieve satellite images, which are standardized and linked to tabular records through file paths for multimodal learning.

6. Comparison of Tabular-Only and Multimodal Approaches

1. Tabular-Only Models

Multiple classical and ensemble machine learning models were evaluated using **only structured tabular features** derived from property attributes and engineered spatial variables.

- **Linear Regression** served as a baseline model and achieved the weakest performance ($R^2 \approx 0.78$), indicating that linear assumptions are insufficient to capture the complex relationships governing house prices.
- **KNN** and **Decision Tree** models showed moderate improvements but were limited by sensitivity to noise and overfitting, as evidenced by relatively higher RMSE values.
- **Random Forest** significantly improved test performance ($R^2 \approx 0.88$) but exhibited a large train–test RMSE gap, indicating overfitting due to its high capacity.
- **XGBoost** and **LightGBM** achieved the best performance among all tabular-only models.
 - LightGBM attained the highest test R^2 score (**0.904**), with strong generalization and a small train–test gap.
 - These gradient boosting models effectively captured non-linear interactions and feature importance within structured data.

Overall, tabular-only ensemble models demonstrated strong predictive capability when high-quality feature engineering and careful regularization were applied.

2. Multimodal Model (Tabular + Satellite Images)

The multimodal approach combined:

- A CNN (EfficientNetB0) for satellite imagery
- A neural network for tabular features
- A fusion layer for joint learning

The multimodal model achieved:

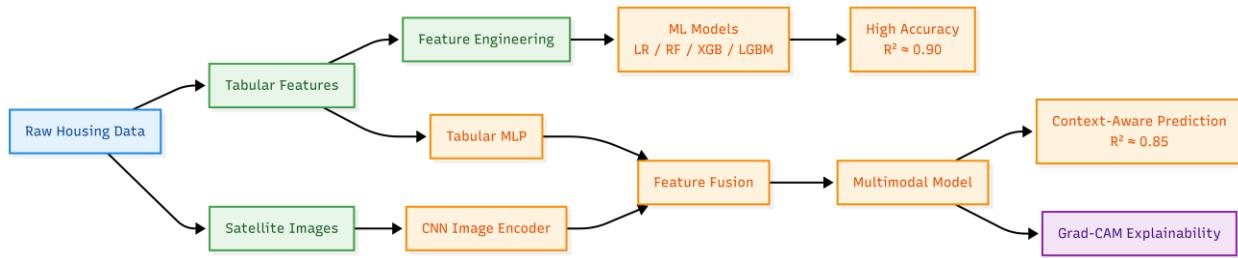
- **Log RMSE ≈ 0.206**
- **Log $R^2 \approx 0.851$**

While this performance is lower than the best tabular-only gradient boosting models, it provides **additional qualitative advantages**:

- Incorporates **visual neighborhood context** (greenery, road density, urban structure)
- Learns information not explicitly present in tabular features

- Enables **model interpretability via Grad-CAM**, which is not possible with tree-based models

The slightly reduced numerical performance can be attributed to limited image resolution, fixed zoom levels, and the challenge of jointly optimizing heterogeneous data sources.



3. Key Observations

- **Best numerical performance:** LightGBM (tabular-only)
- **Best interpretability and contextual understanding:** Multimodal model
- **Tree-based ensembles** excel when rich, engineered tabular features are available
- **Multimodal learning** offers complementary insights and better real-world interpretability, even if raw metrics are slightly lower

This comparison highlights a trade-off between **pure predictive accuracy** and **context-aware, explainable modeling**.