Satellite Imagery-Based Property Valuation using Multimodal Deep Learning

Introduction:

Accurate real estate valuation requires more than just property attributes such as size, number of bedrooms, or location coordinates. Environmental and neighborhood context such as greenery, road connectivity, proximity to water, and urban density — plays a critical role in determining property value.

This project develops a multimodal regression framework that combines structured tabular data with satellite imagery to predict house prices. By fusing visual context with classical housing attributes, the model captures "curb appeal" and spatial characteristics that are not directly available in tabular form.

EXPLORATORY DATA ANALYSIS (EDA) & DATA CLEANING

Exploratory Data Analysis (EDA) was performed to understand the structure, quality, and statistical properties of the housing dataset before building the predictive model. This step also included systematic data cleaning and feature engineering, and the resulting cleaned dataset was saved for modeling.

Dataset Inspection

The original dataset contained information about property size, rooms, location, condition, and sale price. Initial inspection revealed:

- Mixed data types (numerical, categorical, and date fields)
- Missing values in renovation and basement fields
- A few extreme outliers in bedroom count and lot size

The shape, data types, and null values were inspected using pandas.info() and summary statistics.

Handling Dates: The date column was converted to a datetime format and decomposed into:

- sale_year
- sale_month

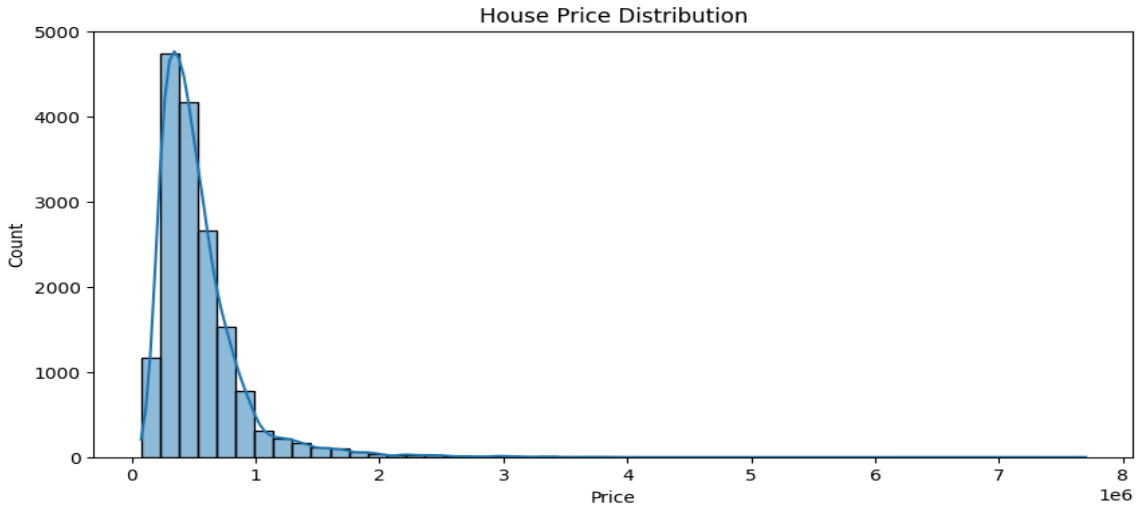
These features allow the model to capture seasonal and market trends.

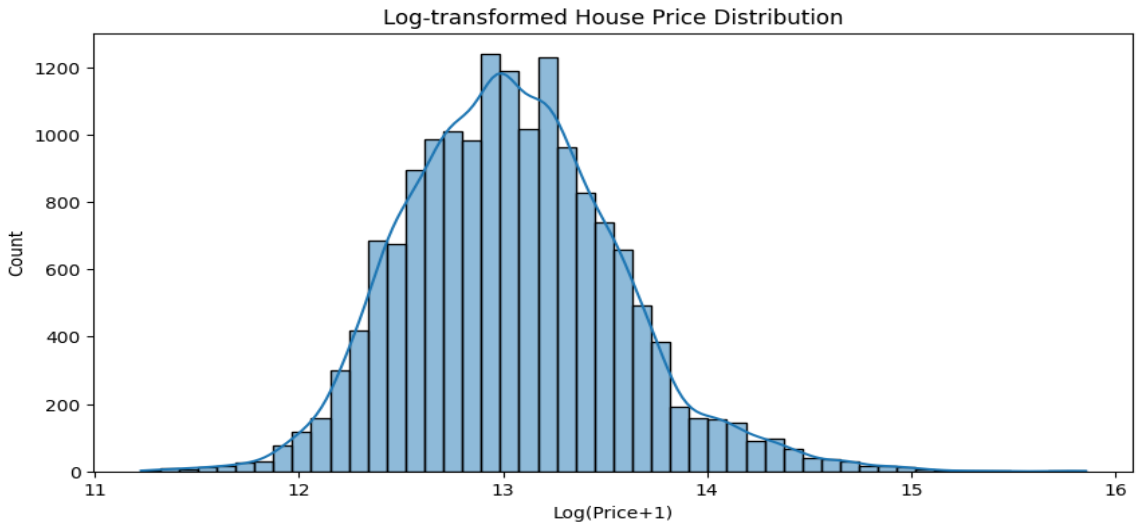
The original date column was then removed.

Feature Engineering: Several meaningful real-estate features were created:

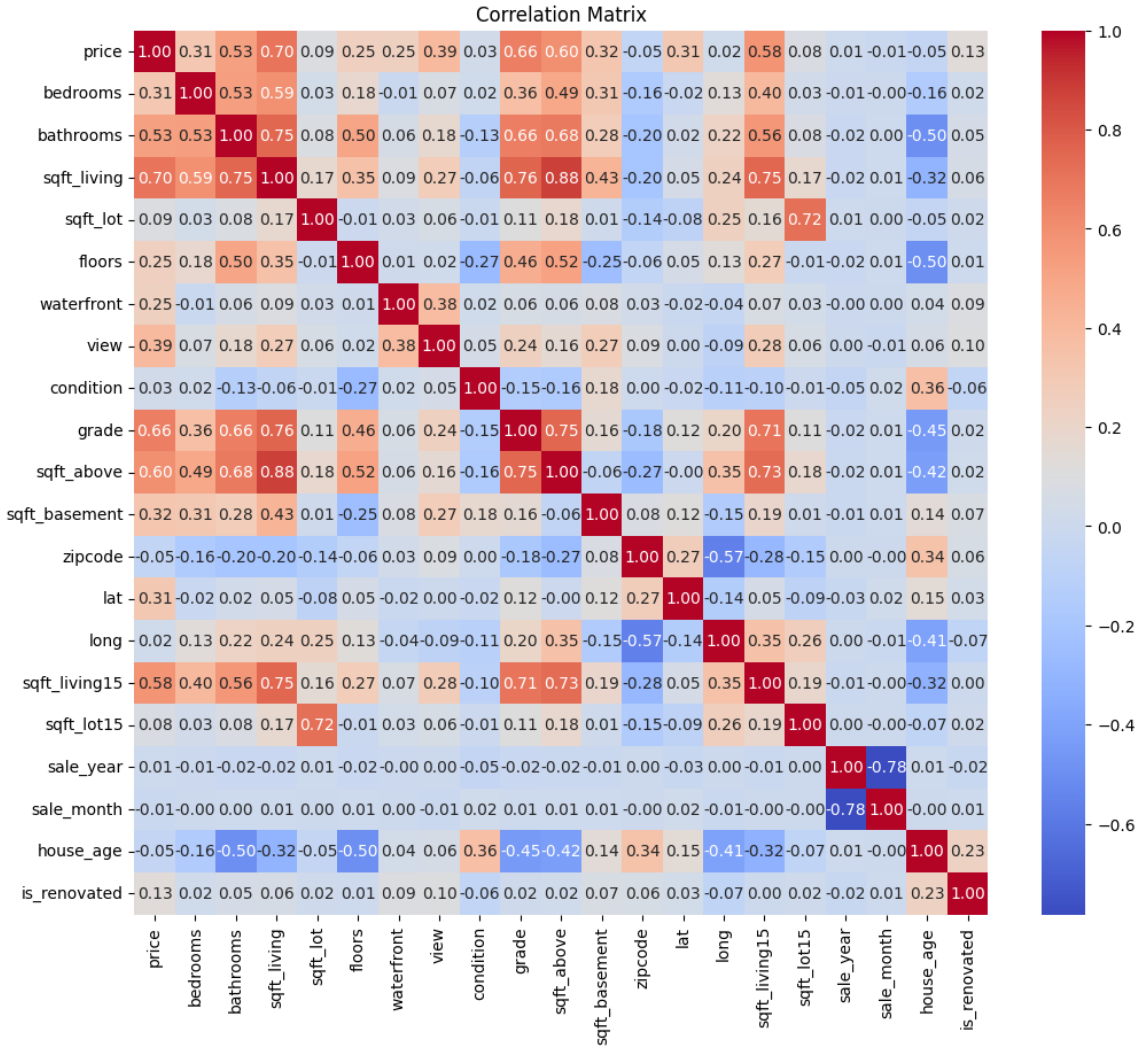
Feature	**Description**
house_age	Difference between sale year and year built
is_renovated	Binary flag indicating whether a renovation was done
log_price	Log-transformed target variable for stable regression

The original yr_built and yr_renovated fields were dropped after feature creation.





Correlation matrix to visualize how two features are correlated:



Outlier Handling:

Properties with unrealistically high bedroom counts were removed to prevent distortion of model learning.

Categorical Encoding: The zipcode column was converted into numerical form using categorical encoding to make it compatible with machine learning models.

Saving the Cleaned Dataset: After all cleaning and feature engineering, the processed dataset was saved as a new file and used for all modeling steps.

This ensured:

- Reproducibility
- Consistent train-test processing
- Separation between raw and clean data

Satellite Image Acquisition:

To incorporate environmental and neighborhood context into the housing price prediction model, high-resolution satellite images were programmatically fetched for each property using its geographic coordinates (latitude and longitude).

API Used

We used the Mapbox Static Images API with the Satellite map style, which provides high-resolution aerial imagery.

Mapbox was chosen because:

- It provides free tier access
- It allows programmatic downloads
- It supports high zoom levels suitable for individual properties

Image Download Pipeline

Each property contains:

- lat → Latitude
- long → Longitude

These were used to generate a URL request for Mapbox.

For each house:

1. The satellite image URL was created using: longitude, latitude, zoom, image_size
2. A zoom level of 18 was used to capture house-level details
3. Each image was downloaded in 512×512 resolution
4. The images were saved using a consistent naming scheme:

   house_<row_index>.jpg

sample of downloaded image: house_1

Baseline Selection:

Before introducing satellite imagery, we first established a strong baseline model using only the structured (tabular) housing data. This step is crucial because it allows us to measure how much predictive power is added by visual information.

The following regression models were evaluated:

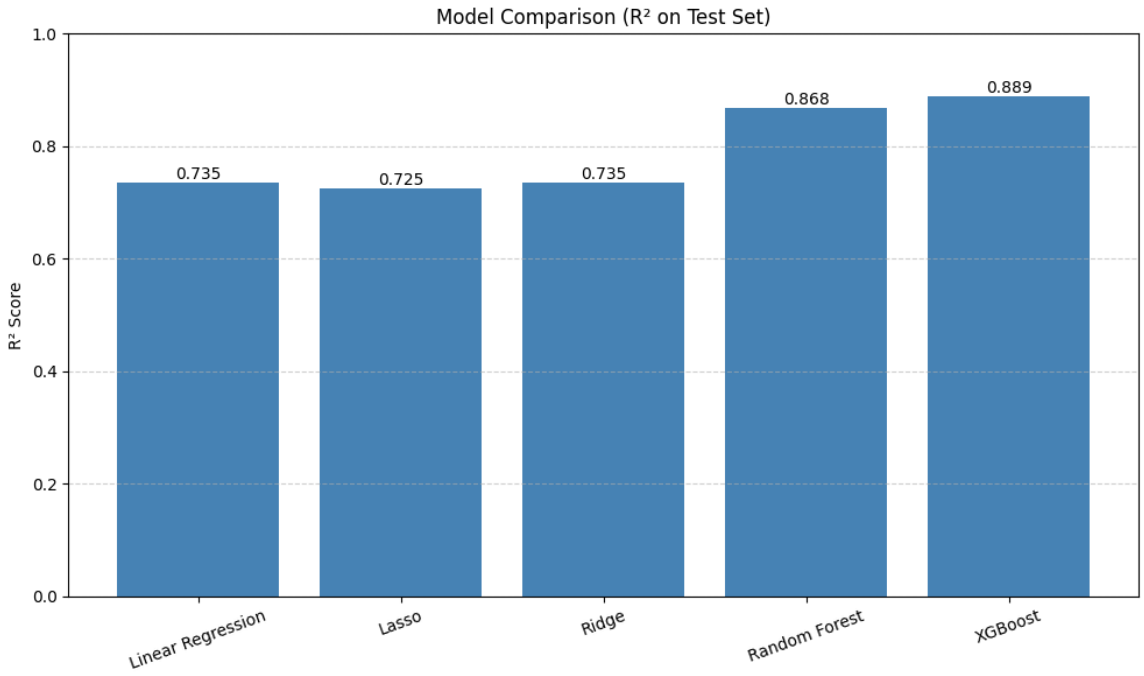
- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest
- XGBoost

Each model was trained to predict the log-transformed house price using engineered numerical features such as square footage, location, house age, grade, and neighborhood statistics.

Performance was evaluated using:

- R^2 score (explained variance)
- RMSE (prediction error in dollars after converting back from log-price)

These metrics were computed on a held-out test set to ensure fair comparison.



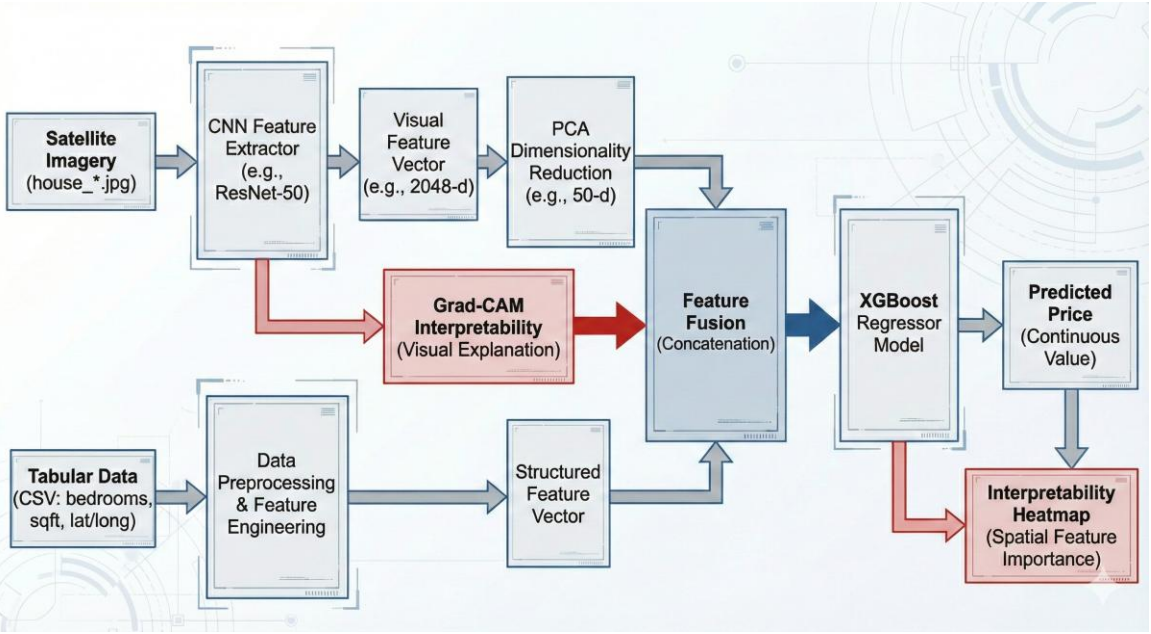
So we will choose XGBoost as the baseline model.

Multimodal Model Architecture:

To capture both structural and environmental factors that influence house prices, we designed a multimodal learning system that combines tabular data with satellite imagery.

This architecture allows the model to learn:

- What the house is (size, rooms, age, location)
- Where the house is (greenery, roads, water bodies, neighborhood layout)



1)Image Processing Branch:

Each property is associated with a satellite image centered on its latitude and longitude. These images are passed through a pretrained ResNet-50 CNN.

The CNN does not predict price directly. Instead, it works as a feature extractor.

It learns patterns such as:

- Green areas (parks, trees)
- Roads and infrastructure
- Water bodies
- Density of surrounding buildings

The CNN outputs a 2048-dimensional embedding representing the visual characteristics of the neighborhood.

To reduce dimensionality and remove redundancy, PCA compresses these features into 50 visual features while retaining most of the variance.

2)Tabular Data Branch:

The numerical data is processed separately using feature engineering:

- Date → Sale year & month
- House age
- Renovation flag
- Neighborhood statistics
- Location and size attributes

All categorical variables (like zipcode) are encoded numerically.

This creates a structured feature vector that captures the physical and financial attributes of the property.

3)Feature Fusion:

The tabular features and the visual features are concatenated:

Final Feature Vector = [Tabular Features | Image Features]

This creates a single multimodal representation of each property.

This fused vector contains:

- What the house looks like
- Where it is
- How big it is
- How good its neighborhood is

4)Final Regressor (XGBoost):

The fused features are fed into XGBoost, a powerful gradient-boosted decision tree model.

XGBoost was chosen because:

- It handles nonlinear interactions
- It works well with mixed-scale features
- It can learn complex relationships between visual and numeric data

The model predicts the log of house price, which is later converted back to real price.

Results:

```
Final Multimodal Model (Tabular + Satellite Images)
MAE:  $63,709
RMSE: $111,236
R²:   0.8898
```

Model Performance Comparison:

After training and evaluating multiple models, the final performance on the test dataset was as follows:

Model	R^2 Score
Tabular Data Only	0.8889
Tabular + Satellite Images (Multimodal)	0.8898

Analysis:

- Using tabular data only, the model achieved a strong R^2 of 88.89%, which indicates that the traditional features like number of bedrooms, square footage, location, and house condition explain most of the variance in house prices.
- Incorporating satellite imagery into the model slightly improved the R^2 to 89%, showing that visual features such as neighborhood density, greenery, and surrounding infrastructure provide additional but incremental predictive value.
- The modest improvement suggests that while tabular data captures the majority of variance, satellite imagery adds contextual environmental information that can refine predictions, especially for houses in similar neighborhoods or with subtle visual differences.

Model Explainability:

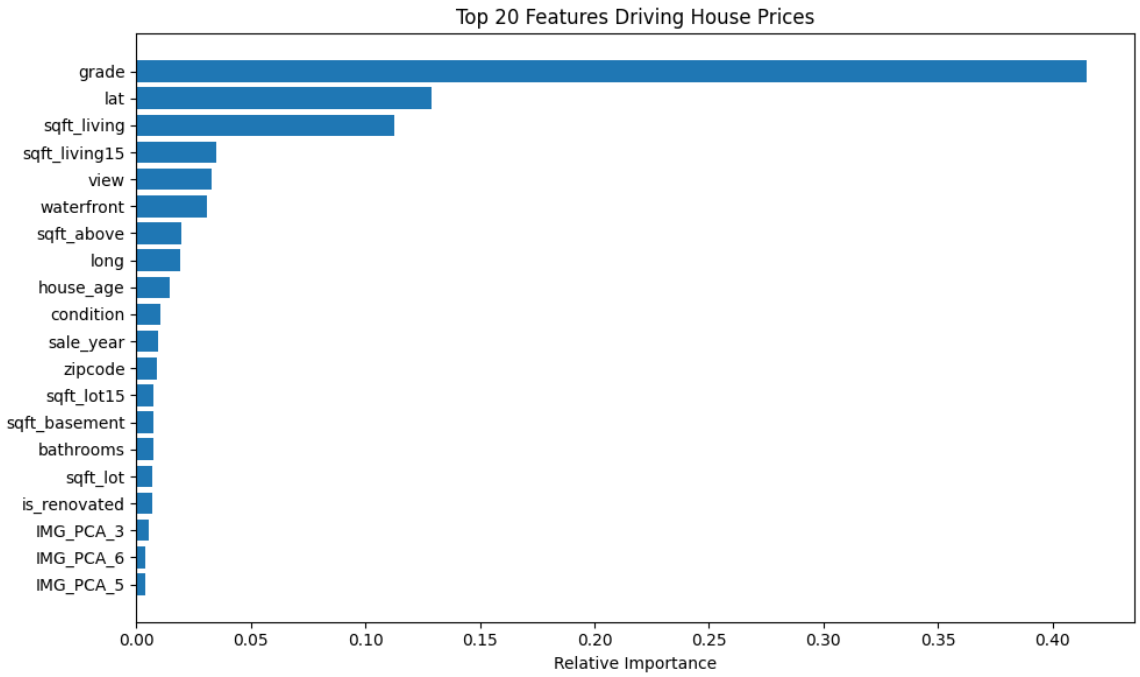
Feature Explainability using SHAP

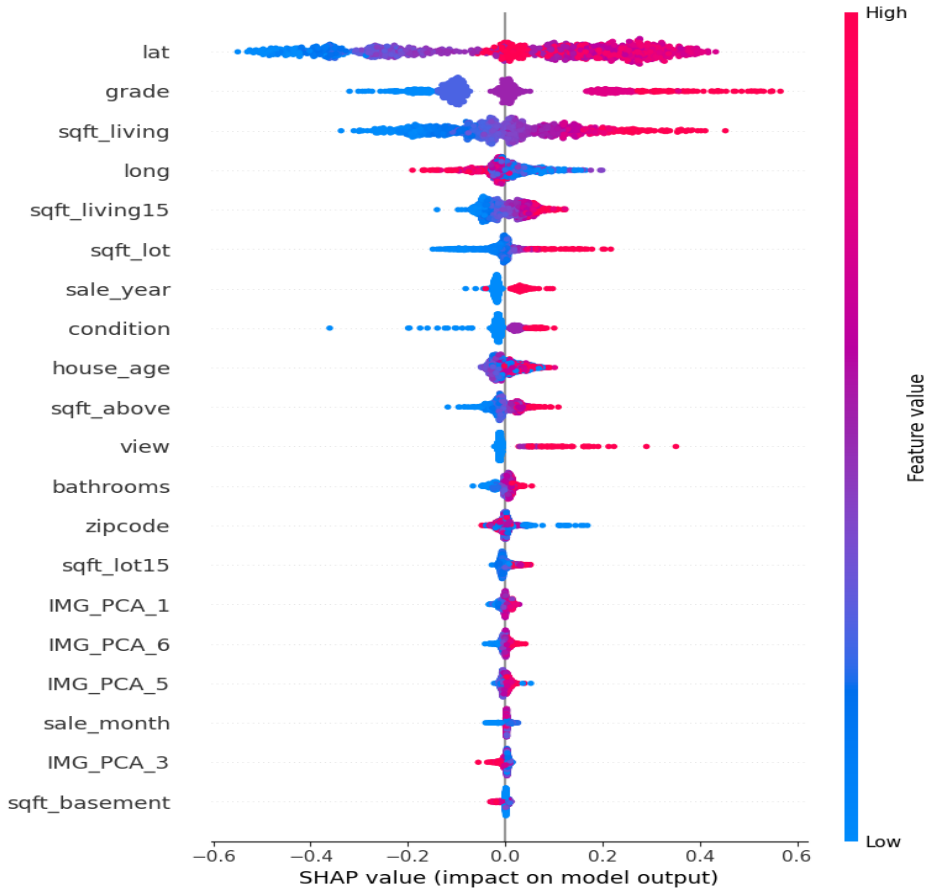
To understand how both tabular data and satellite image features influence house prices, we used SHAP (shapley Additive explanations) on the final XGBoost multimodal model.

SHAP assigns each feature a contribution value that explains how much it pushes a prediction up or down.

This allows us to explain:

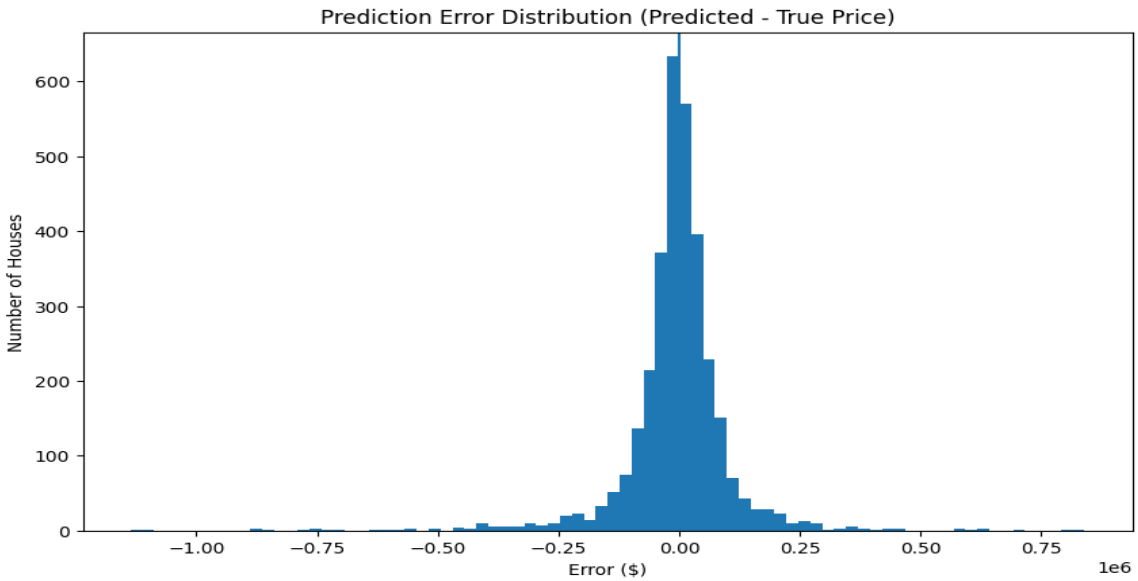
- Why one house is predicted to be expensive
- Why another is predicted to be cheaper
- Whether images truly add value beyond tabular data





Satellite Image Contribution: 5.82 %

Tabular Contribution: 94.18 %



Grad-CAM:

While the multimodal model achieves high accuracy, it is equally important to understand why it makes certain predictions. To make the CNN's decisions interpretable, we applied Grad-CAM (Gradient-weighted Class Activation Mapping) to the satellite imagery branch.

Grad-CAM allows us to visualize which parts of an image most influenced the model's price prediction.

Why Explainability Matters

House pricing is a high-stakes financial decision.
Black-box predictions are not acceptable without justification.

Grad-CAM provides:

- Transparency
- Trust in predictions
- Visual evidence of what the model learns

It allows us to verify whether the model is focusing on meaningful visual cues instead of noise.

How Grad-CAM Works

Grad-CAM operates by analyzing the gradients flowing through the last convolutional layer of the CNN (ResNet-50).

Steps:

1. A house image is passed through the CNN.
2. The model computes the feature maps in the final convolutional layer.
3. The gradient of the predicted price with respect to these feature maps is calculated.
4. Important regions receive higher gradient values.
5. These values are projected back onto the image as a heatmap.

Red areas → Strong influence
Blue areas → Weak influence

What the Model Learns Visually

Grad-CAM reveals that the CNN focuses on:

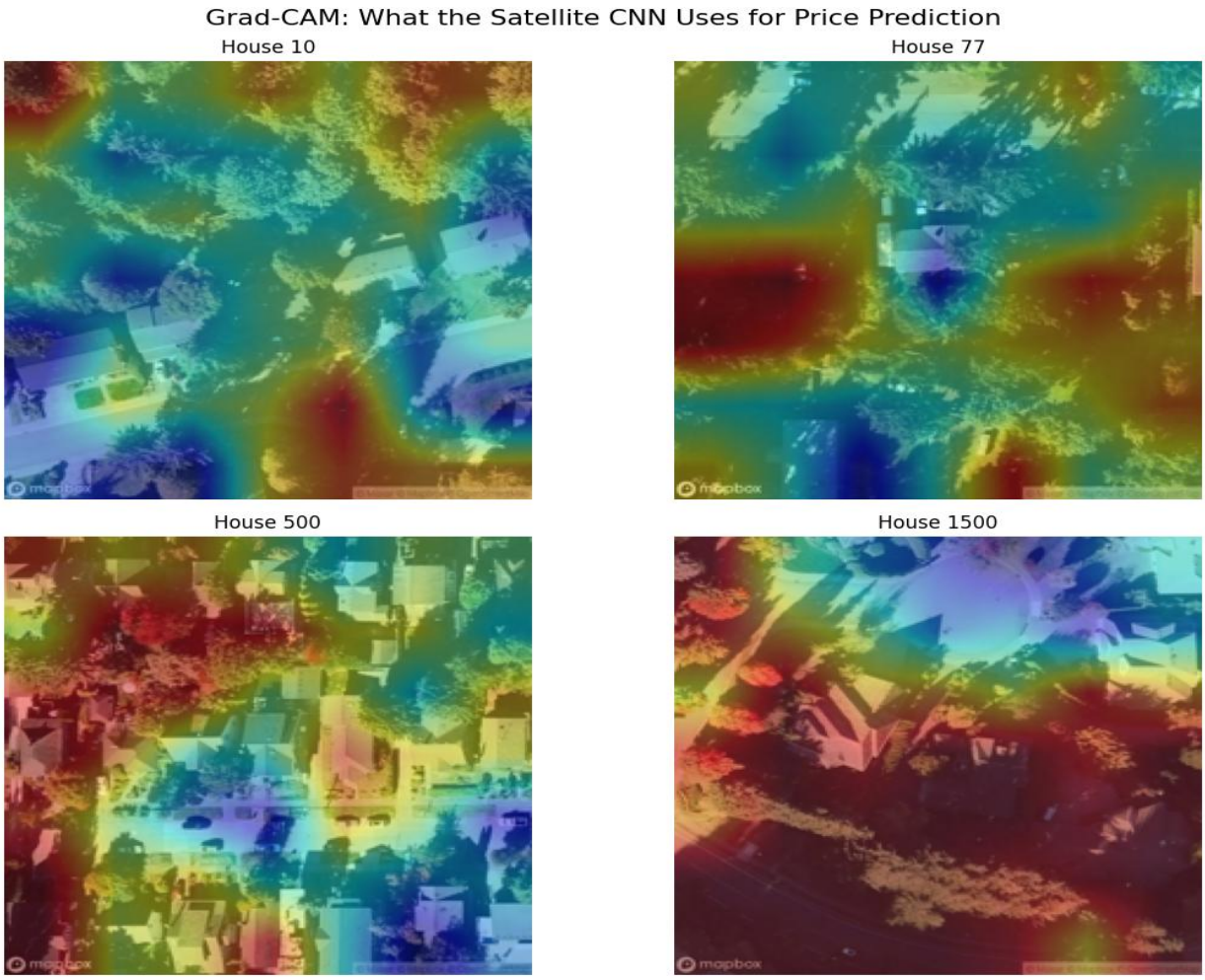
High-value regions

- Water bodies (lakes, rivers, coastline)
- Green spaces and trees
- Open landscapes
- Well-organized road layouts

Low-value regions

- Dense construction
- Industrial zones
- Sparse or poorly developed areas

This aligns strongly with real-estate logic: environment quality affects property price.



Applying the Multimodal Model to the Test Dataset:

After training and validating the multimodal house price prediction system, the same pipeline was applied to the unseen test data to generate final price predictions. To ensure fairness and prevent data leakage, the exact preprocessing, feature engineering, and transformation steps used during training were reused for the test set.

Final Price Prediction

The trained XGBoost model produced predictions in log-price space (for numerical stability). These predictions were converted back to real house prices using exponential transformation.

The final output was stored in both:

- final_predictions.csv
- final_predictions.xlsx

containing:

- House ID
- Predicted market price

Conclusion:

1. Multimodal Learning Adds Value:
   Integrating CNN-derived visual features with tabular data allows the model to capture nuanced environmental characteristics that influence property prices.
2. Tabular Data Remains Dominant:
   Standard property attributes (size, age, location, condition) remain the primary drivers of price prediction.
3. Explainability:
   Grad-CAM visualizations highlighted areas in the satellite images that the CNN focused on, such as dense greenery, nearby roads, or water bodies, helping interpret the model's decisions.
4. Future Improvements:
   - Use higher-resolution satellite images or multiple angles for richer visual context.
   - Incorporate temporal changes (e.g., neighborhood development over years).

- Explore more advanced fusion architectures for better integration of tabular and image features.

Overall, the project demonstrates that multimodal learning is a practical and explainable approach for real estate valuation, capturing both numeric and environmental signals effectively.

Thanking you

B.N.S. PAVAN KUMAR,

23116023,

ECE 3Y.