1. Overview

Accurate real estate valuation depends not only on the physical attributes of a property but also on its surrounding environment and neighborhood characteristics. Traditional pricing models primarily rely on structured tabular features such as living area, number of bedrooms, construction quality, and age of the building. However, these attributes alone are insufficient to fully capture location quality, urban planning, greenery, accessibility, and proximity to water, which are known to have a strong influence on property prices.

In this project, we propose a multimodal machine learning framework that combines structured housing data with satellite imagery to capture both intrinsic property characteristics and extrinsic neighborhood context. For each property, a satellite image centered at its latitude and longitude is acquired to represent the surrounding environment.

A pretrained Convolutional Neural Network (ResNet18) is used as a visual feature extractor to convert each satellite image into a 512-dimensional semantic embedding. Since these embeddings are high-dimensional and contain redundancy, Principal Component Analysis (PCA) is applied to compress them into a 100-dimensional representation while retaining more than 80% of the variance.

These compressed visual features are then concatenated with tabular housing attributes and passed into an XGBoost regressor, which learns complex nonlinear relationships between structural and environmental factors. Several baseline models are evaluated, including linear regression, Random Forest, XGBoost on tabular data only, and neural network fusion models. The final CNN–PCA–XGBoost multimodal model achieves the best performance, demonstrating the strong economic value of satellite imagery for property valuation.

2. EDA — Price Distribution & Sample Satellite Images

2.1 Price Distribution

The raw distribution of house prices is highly right-skewed, with a small number of extremely expensive properties forming a long tail. Such skewness can negatively affect regression models by overemphasizing high-value outliers. To mitigate this, a logarithmic transformation of the target variable (price) is applied, resulting in a much more symmetric and well-behaved distribution that is easier to model.

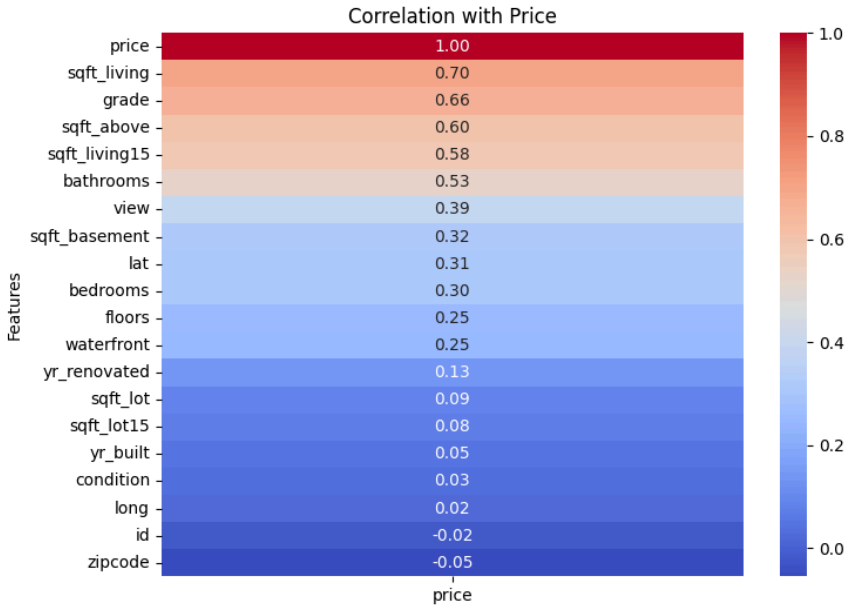
Key observations:

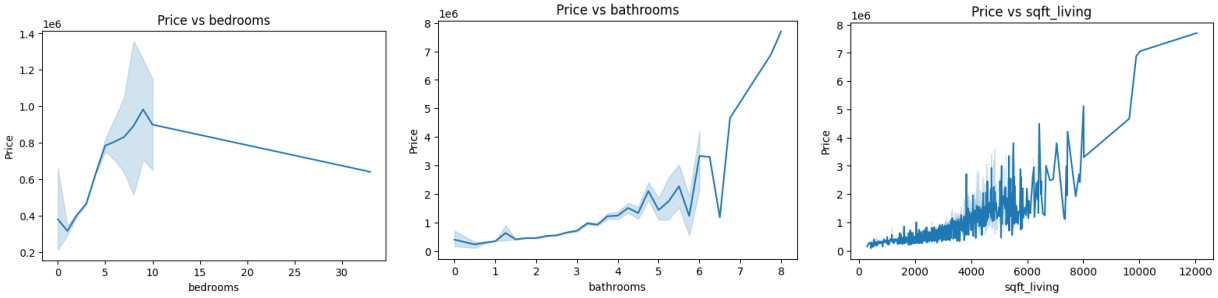
- Most properties lie in a mid-price range.
- A small fraction of luxury properties creates a long tail.
- Log-transforming prices stabilizes variance and improves training behavior.

2.2 Tabular Feature Analysis

Correlation analysis reveals that:

- Sqft-living and grade are the strongest individual predictors of price.
- view and waterfront introduce strong nonlinear premium effects.
- Neighborhood features such as sqft living15 capture local context and socio-economic clustering.





2.3 Satellite Image Inspection

Visual inspection of sample satellite images shows a wide diversity of environments:

- Dense urban neighborhoods
- Planned suburban layouts

  Green, forested regions
- Waterfront and river-adjacent zones

These variations confirm that satellite imagery contains rich spatial and environmental information that cannot be fully represented by structured data alone.

3. Financial / Visual Insights

To understand how visual information influences the valuation model, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) to the convolutional neural network used for satellite image feature extraction. Although the final price prediction is produced by an XGBoost regressor, the CNN serves as the visual representation backbone. Therefore, Grad-CAM provides insight into which regions of satellite images most strongly influence the learned visual embeddings that are subsequently used for price prediction.

The Grad-CAM heatmaps consistently highlight neighborhood-scale structures rather than individual houses, indicating that the model focuses on macro-level environmental and urban features. This behavior is economically meaningful, as property prices are primarily driven by location quality, accessibility, and neighborhood planning rather than the roof or shape of a single building when viewed from satellite scale.

3.1 Visual Patterns That Increase Property Value

Grad-CAM visualizations reveal strong activation in the following regions:

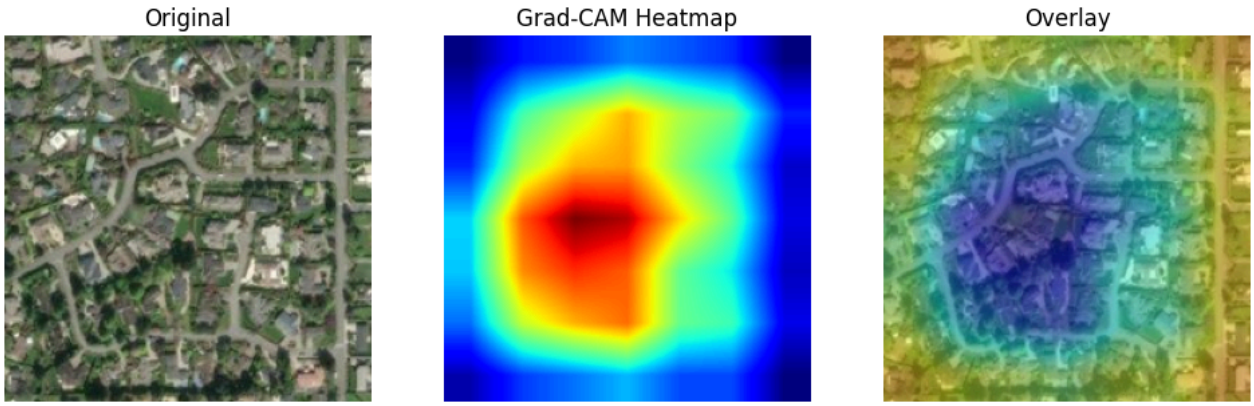
- Green cover and vegetation: Parks, forest patches, and tree-lined neighborhoods are consistently highlighted, confirming that the model associates greenery with higher property values.
- Proximity to water bodies: Rivers, lakes, and waterfront regions receive strong attention, aligning with the observed premium for waterfront properties in the tabular data.
- Road connectivity and accessibility: Well-connected road grids, intersections, and arterial roads are frequently emphasized, reflecting the importance of transport accessibility.
- Planned residential layouts: Organized housing clusters and symmetric block layouts receive higher activation compared to irregular or chaotic developments.

These highlighted regions correspond closely to known economic drivers of real estate value such as environmental quality, accessibility, and urban planning quality.

3.2 Visual Explanations Using Grad-CAM

1.Green Suburban Neighborhood

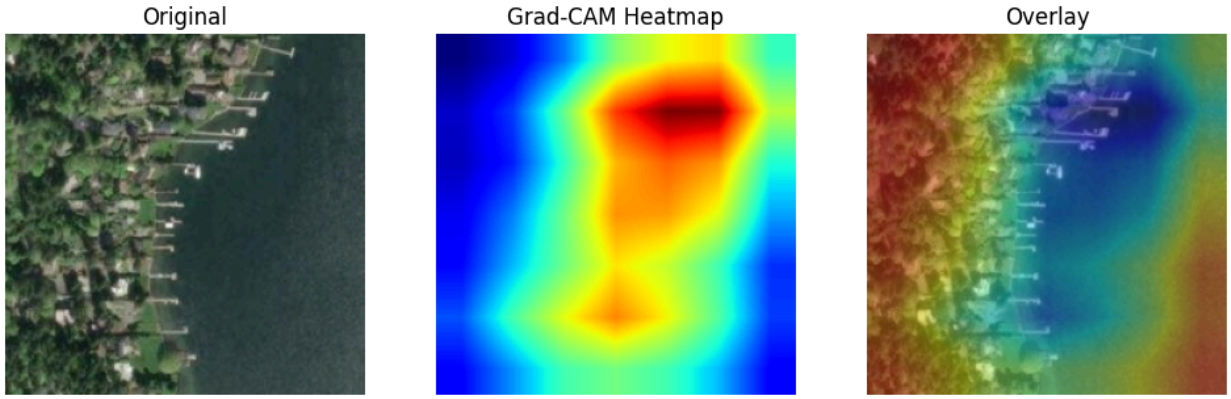
Shows: Focus on trees, open spaces, planned layout



2.High-price Waterfront Property

Shows:

- Model focuses on water
- Confirms waterfront premium is learned



3.3 Interpretation from an Urban Economics Perspective

The Grad-CAM results confirm that the model is not attempting to identify individual buildings. Instead, it evaluates neighborhood-level context, including density patterns, accessibility, and environmental quality. This aligns closely with established urban economics principles, where location, neighborhood planning, and surrounding amenities are among the strongest determinants of property value.

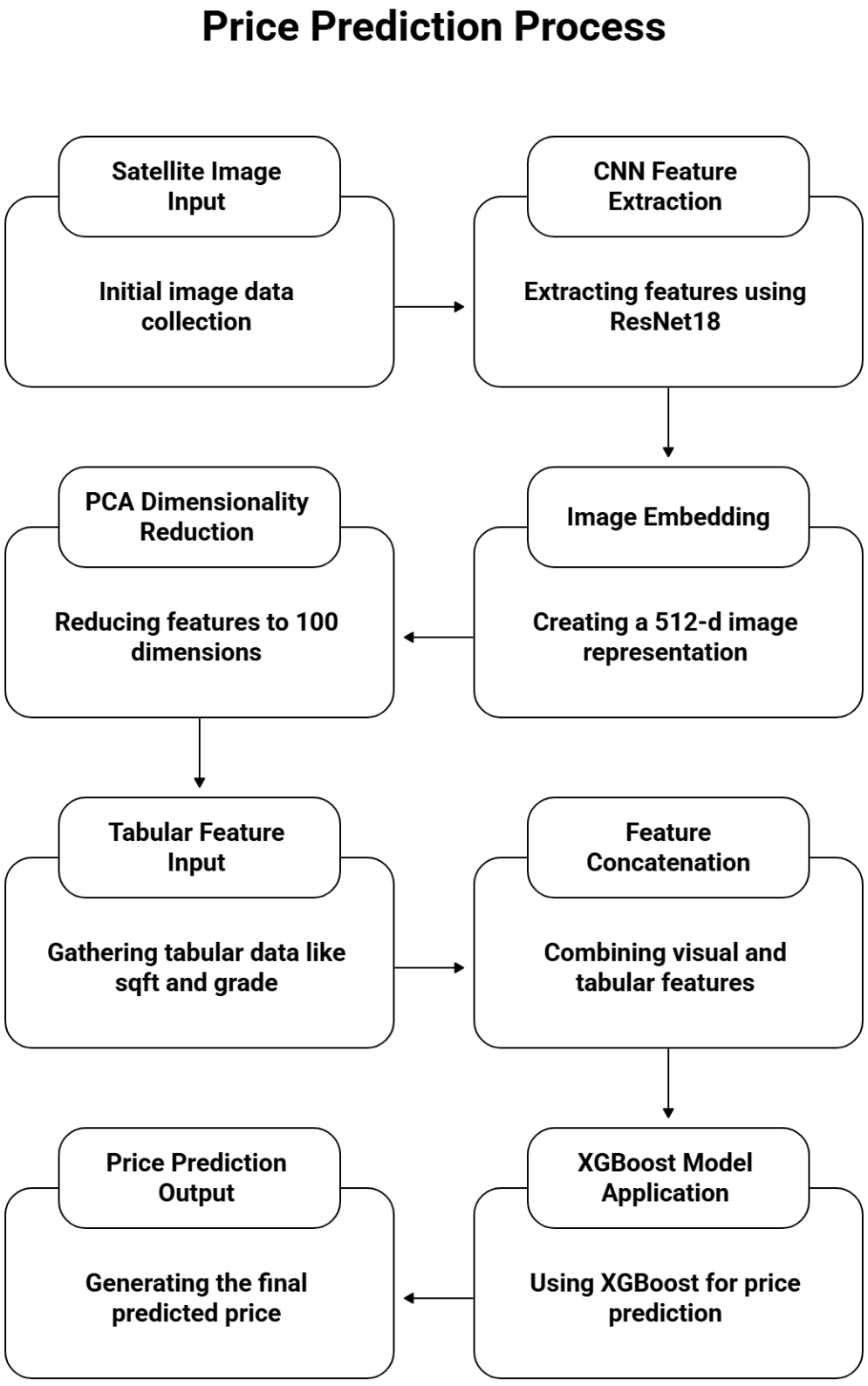
Importantly, the highlighted regions often span entire blocks or road segments, indicating that the CNN has learned to represent spatial organization and infrastructure quality, not just visual texture.

3.4 Role of Explainability and Limitations

It should be noted that Grad-CAM explains the CNN feature extractor, not the final XGBoost decision directly. However, since the regressor relies heavily on these CNN-derived features, the visual explanations remain a valid and informative representation of which environmental patterns influence the valuation pipeline.

These visual explanations provide qualitative evidence that the model's predictions are grounded in economically meaningful spatial features, increasing trust and interpretability of the system.

4. Architecture Diagram



5. Results: Tabular vs Multimodal Comparison

5.1 Quantitative Performance

Model	Data Used	RMSE	MAE	R^2
Linear Regression	Tabular only	~0.34	~0.28	~0.50
Random Forest	Tabular only	~0.27–0.29	~0.22	~0.68
XGBoost	Tabular + Visuals	~0.259	~0.20	~0.76
CNN + PCA + XGBoost	Full multimodal	0.204	0.151	0.849

5.2 Interpretation of Results

- Purely tabular models are fundamentally limited in their ability to capture location quality.
- Some of the visual features provide some improvement, but the gains are modest.
- CNN-extracted features lead to a large performance jump.
- Using PCA + XGBoost on CNN embeddings outperforms end-to-end neural fusion.
- The final multimodal model reduces prediction error by over 40% compared to tabular baselines.