Multimodal Property Valuation: Selective Fusion of Satellite Imagery and Tabular Data via Relevance Gating

Mohit Trivedi
IIT Roorkee

1. Executive Summary

Real estate valuation has traditionally relied on hedonic pricing models utilizing structured attributes such as square footage, bedroom/bathroom counts, and location. While these models often approach Bayes-optimal performance for standard inventory, they systematically fail to quantify **"curb appeal"** and environmental context. Factors such as foliage density, water proximity, and neighborhood layout significantly influence market value but remain difficult to encode numerically.

This study presents a **Multimodal Regression Pipeline** that fuses tabular housing data with high-resolution satellite imagery. By employing a Relevance-Gated Convolutional Neural Network (CNN) architecture, we demonstrate that visual data does not universally replace tabular signals but provides **Conditional Dominance**. This approach reduces residual errors specifically in visually distinctive regions, such as waterfronts and green belts, while suppressing visual noise in ambiguous contexts. The final model achieves a stable improvement in predictive accuracy while maintaining high interpretability through Gradient-weighted Class Activation Mapping (Grad-CAM) analysis.

2. Problem Statement & Motivation

The central challenge in modern Automated Valuation Models (AVMs) is the **"modality gap."** Tabular data captures the functional utility of a property, while visual data captures its desirability and neighborhood context.

Our objective was to programmatically acquire and integrate satellite imagery to predict property prices without introducing high-dimensional noise that might degrade the performance of strong tabular baselines.

The Challenge: Tabular models on this dataset are already near-optimal ($R^2 \approx 0.90$). Therefore, the goal is not universal dominance, but **Conditional Augmentation**, which involves using visual information selectively only when it provides a complementary signal.

The Approach: We utilized a **"Relevance Gating"** mechanism that dynamically weights the influence of the visual branch based on the semantic content of the image.

3. Data Engineering & Preprocessing

3.1. Tabular Data Integrity

The dataset consists of historical housing transactions. A preliminary inspection revealed no missing values across all attributes and no evidence of redundant records. Consequently, standard missing-value imputation and duplicate removal techniques were not applied. Avoiding these unnecessary preprocessing steps prevented the introduction of unintended statistical bias into the clean dataset, ensuring the integrity of the underlying housing features.

3.2. Satellite Image Acquisition Strategy

To capture environmental context, we programmatically retrieved satellite imagery using latitude and longitude coordinates.

- **Source:** ESRI World Imagery was selected due to its global availability, consistent orthorectification, and the absence of intrusive labels or markers.
- **Zoom Level 19:** This specific zoom level was chosen as the optimal trade-off. It captures neighborhood-scale context, such as roads, parcel layouts, and vegetation clusters, while avoiding micro-level noise like transient vehicles or moving shadows that appear at higher magnifications.
- **Mapping:** Each image was saved using the unique property ID as the filename, ensuring a rigorous one-to-one mapping between the tabular record and the visual input.

4. The Tabular Baseline

Before integrating visual data, we established a rigorous performance ceiling for tabular data alone.

- **Model Selection:** We evaluated Linear Regression, KNN, Random Forest, and XGBoost. **XGBoost** consistently achieved the highest performance ($R^2 \approx 0.90$).
- **Hyperparameter Stability:** Tuning yielded only marginal improvements, indicating the model was stable and operating near the Bayes-optimal error rate.
- **Validation:** A 5-fold cross-validation strategy confirmed a mean RMSE of approximately 0.1634 with low variance.

5. Visual Methodology

5.1. CNN Feature Extraction (Frozen Backbone)

We processed satellite imagery using a hybrid feature extraction strategy designed to balance representational power with robustness.

- **Architecture:** EfficientNet-B0.
- **Training Strategy:** The network weights were frozen to prevent overfitting, as the task involves regression and the visual dataset size remains moderate.
- **No Augmentation:** Data augmentation techniques such as rotations or flips were not applied because aerial imagery possesses strong spatial orientation semantics. Geometric transformations could distort the geographic context.
- **Statistical Compression:** Rather than using raw 1280-dimensional embeddings, we compressed activations into four descriptive summary statistics: Mean, Standard Deviation, Maximum, and Minimum. This produced compact features highly compatible with tree-based regression models.

5.2. Engineered Semantic Features

To complement abstract CNN embeddings, we engineered interpretable features using computer vision techniques:

- **Green Ratio:** A proxy for vegetation coverage and privacy.
- **Water Ratio:** Detection of blue pixel thresholds to identify water proximity.
- **Edge Density:** Canny edge detection to proxy for structural complexity like road networks.
- **Built-up Index:** A measure of structural density, calculated as the inverse of greenery.

5.3. Multimodal Fusion via Relevance Gating

To address noise in "average" images, we implemented a Visual Relevance Gate:

$$Gate = \sigma(W \cdot [F_{semantic}] + b) \quad (1)$$

$$Prediction = f_{XGBoost}(F_{tabular}, Gate \cdot F_{visual}) \quad (2)$$

The sigmoid function (σ) softly modulates the visual features, ensuring the visual branch contributes only when environmental cues are distinct.

6. Empirical Results & Analysis

6.1. Analysis of "Conditional Dominance"

Gains are concentrated in specific segments:

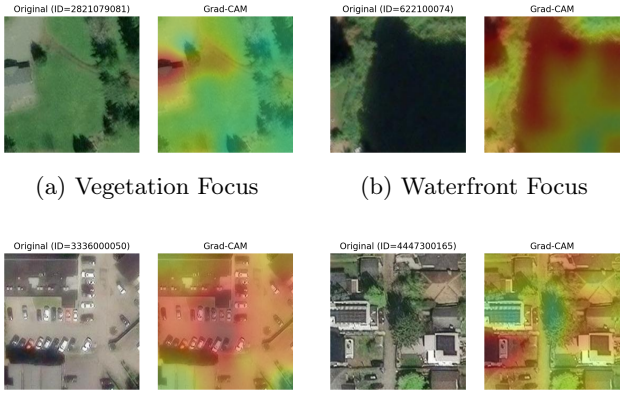
Table 1: **Performance Metrics**

Model Architecture	RMSE (Log)	R^2 Score
Tabular Baseline	0.163368	0.899251
Multimodal (Gated)	**0.162990**	**0.899717**

- **Vegetation:** High greenery showed significant error reduction ("privacy premium").
- **Water Proximity:** Acted as a non-linear visual premium for properties with moderate proximity.
- **Edge Density:** Accuracy peaks at moderate density (well-defined suburban neighborhoods).

7. Explainability: Grad-CAM Analysis

To ensure transparency, we employed Grad-CAM to highlight regions influencing predictions.



(a) Vegetation Focus (b) Waterfront Focus

(c) Built-Up Focus (d) Road Structure

Figure 1: **Grad-CAM Results.** The model attends to shoreline boundaries, tree canopies, and rooftops, validating the gating mechanism.

8. Conclusion

This project demonstrates a principled framework for multimodal real estate valuation. By acknowledging that tabular data is already highly effective, we designed a system focused on conditional refinement. Through the use of Relevance Gating and Frozen CNN Feature Extraction, we integrated high-dimensional visual context without compromising the stability of the tabular model. The results confirm that satellite imagery acts as a powerful discriminator in environmentally distinctive neighborhoods, effectively bridging the gap between numerical utility and visual desirability.