# Satellite Imagery-Based Property Valuation – Report

## Multimodal Regression Pipeline

**Student Name :** Sahil Yuvraj Kamble

**Enrollment No:** 23112087

**Branch** : Chemical Engineering, 3rd Year
**Repository:** https://github.com/sahil2448/PropertyValuationMultimodal
**Date:** January 2026

---

## 1. Overview: Approach & Modeling Strategy

### Solution Approach

**Data Pipeline:** 1. Loaded housing dataset (16,209 training samples) 2. Programmatically fetched Sentinel-2 satellite images for each property using lat/long coordinates 3. Engineered 27 tabular features from 15 raw features 4. Extracted 512-dimensional visual embeddings using pretrained ResNet18

**Modeling Strategy:** - **Baseline:** HistGradientBoostingRegressor on tabular features only - **Multimodal:** HistGradientBoostingRegressor on concatenated [tabular + image embeddings] - **Fusion Type:** Intermediate fusion (feature-level concatenation before final regressor)

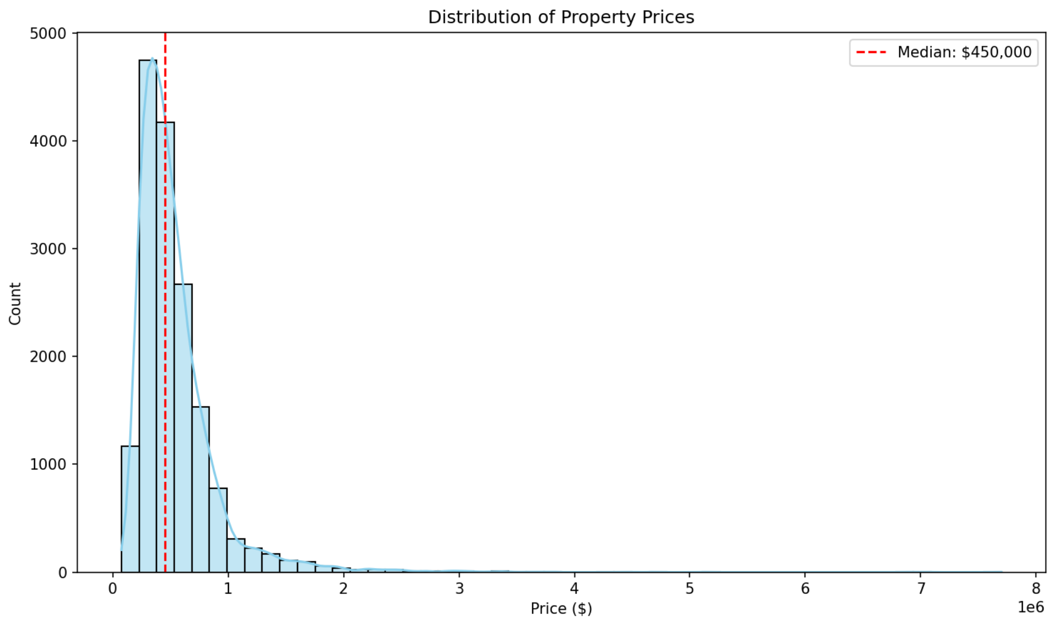
### Key Technical Decisions

| Component | Choice | Rationale |
|---|---|---|
| Image API | Sentinel Hub (Copernicus) | Free, high-quality Sentinel-2 imagery |
| CNN Backbone | ResNet18 (pretrained, frozen) | Fast inference, strong transfer learning |
| Fusion Method | Feature concatenation | Simple, interpretable, effective |
| Final Regressor | HistGradientBoosting Regressor | Handles mixed features well, robust |

---

## 2. EDA: Exploratory Data Analysis

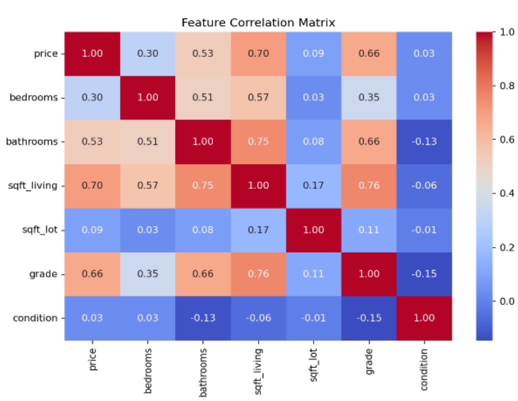
### 2.1 Price Distribution




*Price Distribution*

| Statistic | Value |
|---|---|
| Median Price | $450,000 |
| Min Price | $75,000 |
| Max Price | $7,700,000 |
| Distribution | Right-skewed (long tail of luxury properties) |

**Insight:** Most properties cluster below $1M, with sparse luxury outliers requiring special handling.




### 2.2 Feature Correlations

*Correlation Matrix*

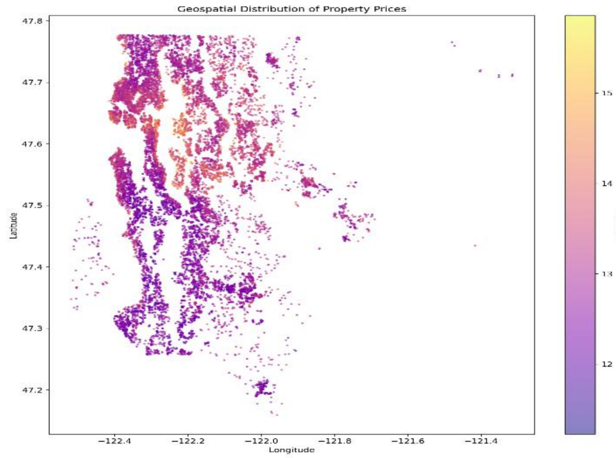
***Strongest Predictors of Price:***

1. *sqft_living → **0.70** correlation*
2. *grade → **0.66** correlation*
3. *bathrooms → **0.53** correlation*

**Weak Predictors:** `sqft_lot` (0.09), `condition` (0.03)

### 2.3 Geospatial Analysis

*Geospatial Map*




**Geographic Price Patterns:** - **High-value core:** Seattle metro area (lat 47.5–47.7) - **Waterfront premium:** Properties along Lake Washington command higher prices - **Price gradient:** Values decrease moving south and east from urban center

### 2.4 Sample Satellite Images

Sample images - Low-price properties: Smaller lots, dense neighborhoods, less vegetation - High-price properties: Larger lots, waterfront access, more greenery

---

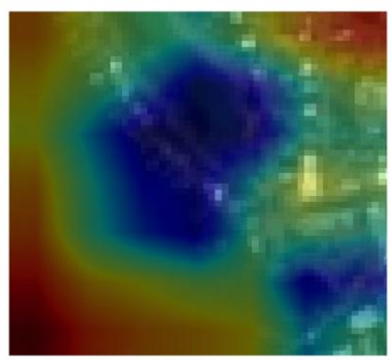
## 3. Financial/Visual Insights: What Drives Value?

### Grad-CAM Explainability Analysis

Used Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which image regions influence the model's predictions.

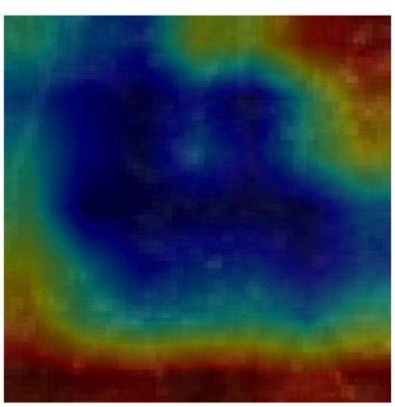
### High-Value Property Analysis ($5.1M)

*GradCAM High Value*



**Model Focus Areas (Red = High Influence):** - Lot boundaries and property size - Surrounding vegetation/greenery - Corner positioning and open space

**Key Finding:** Model values **lot size and environmental context** over building footprint.



### Low-Value Property Analysis ($75K)

*GradCAM Low Value*

**Model Focus Areas:** - Peripheral infrastructure (roads, curbs) - Lack of landscaping in central region - Dense surrounding development

### Visual Features Summary

| Feature | High-Value Signal | Low-Value Signal |
|---|---|---|
| **Vegetation** | Abundant trees, landscaping | Sparse, barren lots |
| **Lot Size** | Large, visible boundaries | Small, cramped |
| **Open Space** | Waterfront, parks nearby | Urban density |
| **Infrastructure** | Well-maintained surroundings | Industrial/bare roads |

---

## 4. Architecture Diagram

### Data Flow Pipeline

**Stage 1: Data Acquisition**

- Input: train.csv containing latitude, longitude, and housing features
- Process: Sentinel Hub API fetches satellite imagery for each property location
- Output: 224×224 RGB satellite images

**Stage 2: Feature Extraction**

*Tabular Branch:*

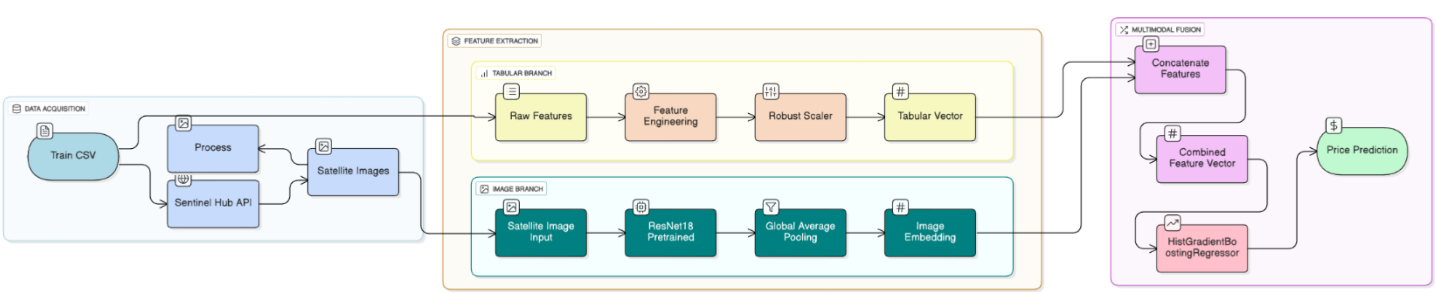
- 15 raw features → Feature engineering (12 derived features) → RobustScaler → 27-dimensional vector

*Image Branch:*

- Satellite image → ResNet18 (pretrained, frozen backbone) → Global Average Pooling → 512-dimensional embedding

**Stage 3: Multimodal Fusion**

- Concatenate tabular (27-dim) + image (512-dim) → 539-dimensional feature vector
- Feed into HistGradientBoostingRegressor (lr=0.05, max_depth=8, max_iter=800)
- Output: Property price prediction



### Model Architecture Details

| Layer | Input Dim | Output Dim | Notes |
|---|---|---|---|
| ResNet18 Backbone | 224×224×3 | 512 | Frozen pretrained weights |
| Tabular Scaling | 27 | 27 | RobustScaler normalization |
| Concatenation | 27 + 512 | 539 | Feature-level fusion |
| HGBR Regressor | 539 | 1 | Final price prediction |

---

## 5. Results: Model Comparison

### Performance Metrics (Validation Set: 3,242 samples)

| Model | Modalities | RMSE ($) ↓ | $R^2$ Score ↑ |
|---|---|---|---|
| **Tabular Baseline** | Tabular only | 134,809 | 0.8552 |
| **Multimodal Fusion** | Tabular + Satellite | **132,247** | **0.8606** |

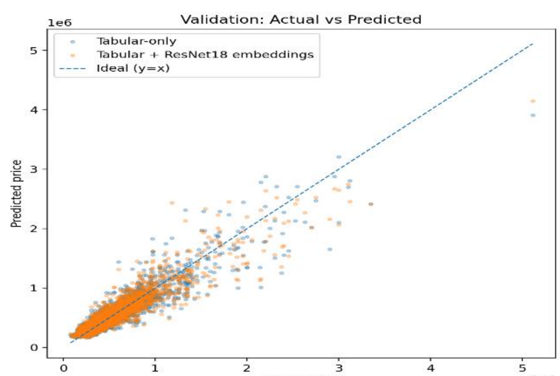
### Improvement Analysis

| Metric | Improvement |
|---|---|
| **RMSE Reduction** | $2,562 (1.9% decrease) |
| **$R^2$ Increase** | +0.0054 (0.54 percentage points) |

### Validation Scatter Plot

*Predicted vs Actual*

**Observations:** 1. **Strong correlation** for properties under $1.5M (majority of data) 2. **Fusion model (orange)** shows tighter clustering than baseline (blue) 3. **Underprediction** for luxury properties above $3M (model limitation)




### Key Findings

1. **Satellite imagery improves predictions** — consistent 1.9% RMSE reduction
2. **Visual context captured:** lot size, vegetation, neighborhood density
3. **Fusion architecture works:** simple concatenation is effective
4. **Limitation:** Both models struggle with luxury outliers (>$3M)