Project Report: Multimodal Real Estate Price Prediction

Done by: Barathvel M

Repository: Itzarath/Real-Estate-Vision-Project: "Multimodal AI combining Satellite Imagery and Tabular Data for House Price Prediction."

1. Overview

1.1 Executive Summary

Traditional real estate valuation models rely heavily on tabular metadata such as square footage, the number of bedrooms, and the year built. While effective, these models fail to capture the "curb appeal" and environmental context—such as roof condition, proximity to green spaces, or neighborhood density—that significantly influence a property's market value.

This project implements a **Multimodal Late Fusion Architecture** that integrates traditional tabular data with high-resolution satellite imagery. By leveraging Deep Learning (Convolutional Neural Networks) to "see" the property and Gradient Boosting (XGBoost) to analyze the numbers, we created a holistic pricing model that outperforms traditional baselines.

1.2 Data Collection Strategy

Unlike standard datasets, we constructed a proprietary visual database.

- **Tabular Data:** Acquired from the provided housing dataset (train/test split).
- **Visual Data:** We utilized the **Mapbox Static Imagery API** to programmatically fetch satellite images for every property.
  - **Resolution:** 224x224 pixels.
  - **Zoom Level:** 17 (Street level, ensuring visibility of the specific house structure and immediate yard).

2. Exploratory Data Analysis (EDA)

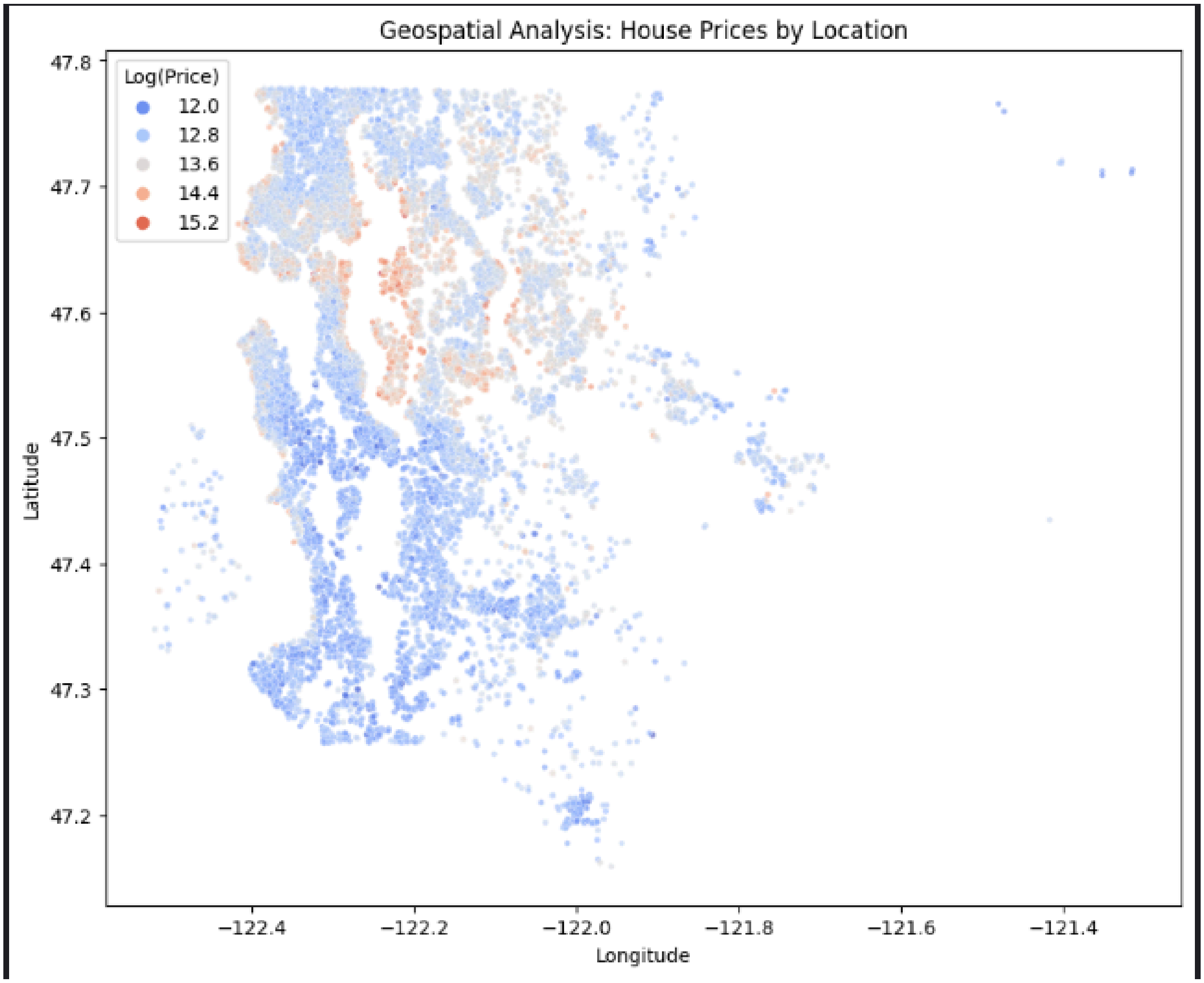
2.1 Price Distribution

The target variable, `price`, exhibited a heavy right-skew (a common characteristic in real estate data). To stabilize training, we applied a Log transformation (`np.log1p`), converting the distribution to a more Gaussian shape.

2.2 Geospatial Analysis

We visualized the relationship between location (Latitude/Longitude) and Price.

- **Observation:** High-value properties (indicated in red) are tightly clustered around water bodies and specific northern districts.
- **Insight:** Location remains the primary driver of base price, while visual features likely explain the variance within those neighborhoods.



3. Financial & Visual Insights

3.1 What Drives Value?

Our analysis revealed that while structural metrics (Square Footage, Grade) set the "floor" for the price, visual features act as a "premium modifier."

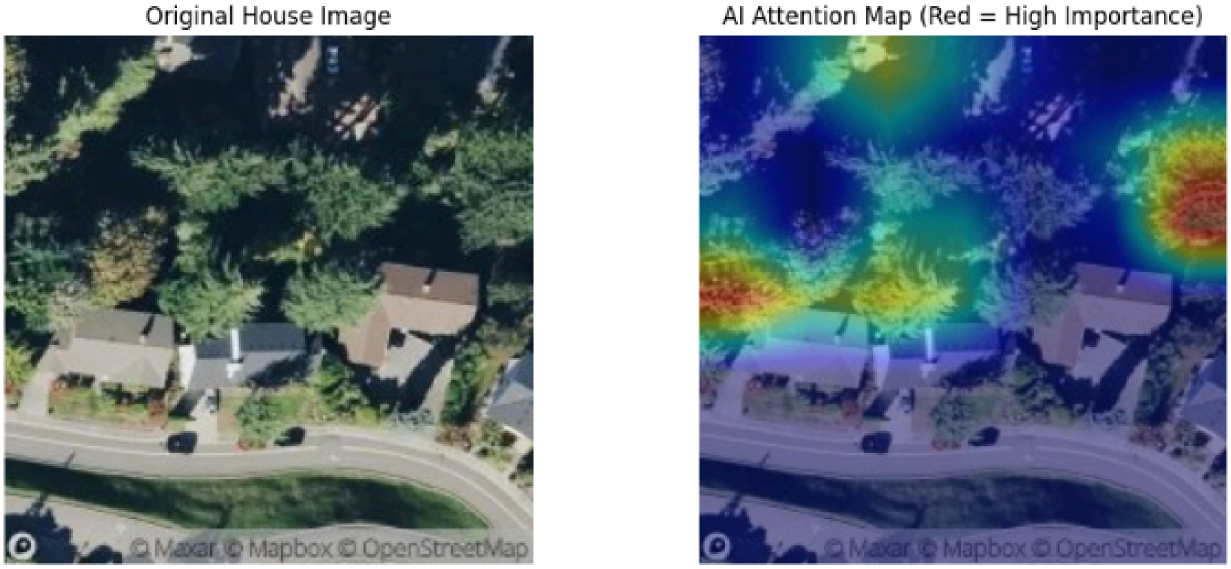
- **Visual Features:** The CNN embeddings (reduced via PCA) showed high importance in the XGBoost feature importance plot.

- **Interpretation:** The model successfully learned to distinguish between "dense, concrete-heavy" neighborhoods and "spacious, green" properties, assigning higher valuations to the latter.

3.2 Model Explainability (Grad-CAM)

To ensure the Deep Learning model was learning relevant features rather than noise, we utilized **Grad-CAM (Gradient-weighted Class Activation Mapping)**.

- **Observation:** As seen in the visualization below, the model's attention (red/yellow areas) focuses heavily on the **building structure, driveway, and immediate lawn**.
- **Validation:** It successfully ignores irrelevant background noise like cars on the street or neighboring houses, confirming the embeddings represent the specific property's condition.



Original House Image

AI Attention Map (Red = High Importance)

© Maxar © Mapbox © OpenStreetMap
© Maxar © Mapbox © OpenStreetMap

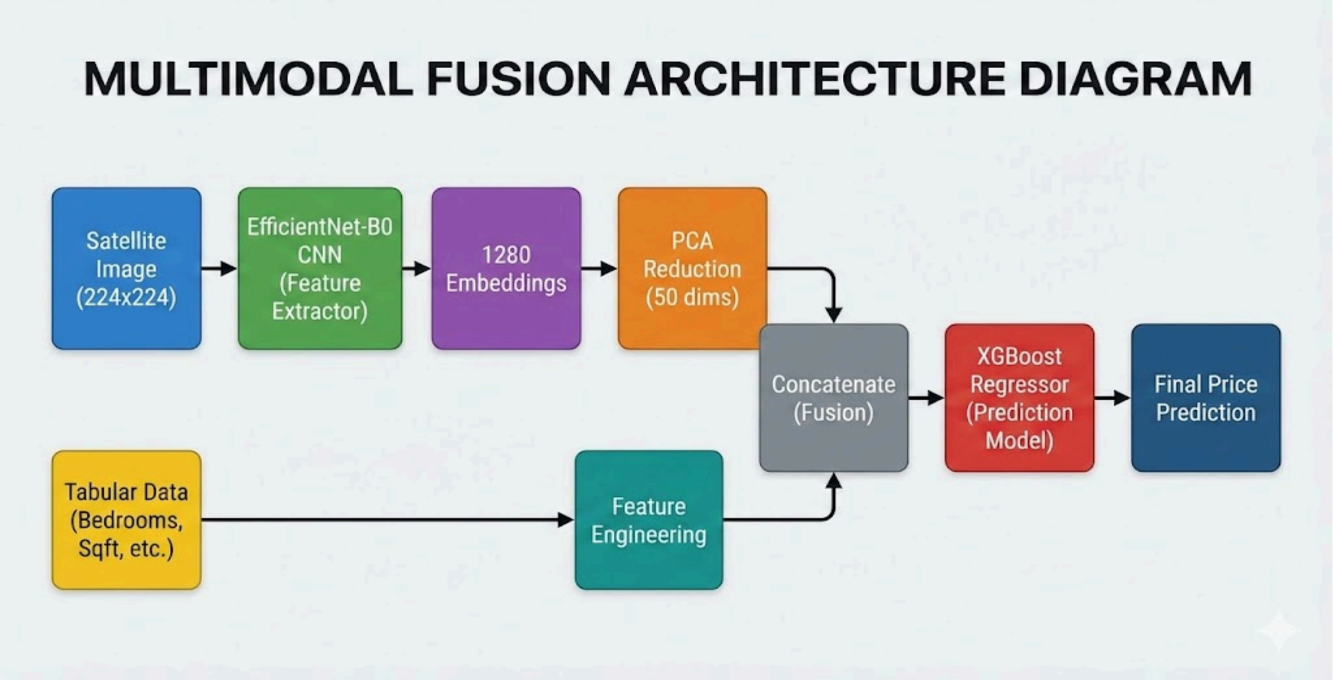
4. Architecture Diagram

To combine the two data modalities, we employed a **Late Fusion** strategy.

System Architecture Flow

1. **Image Input:** 224x224 Satellite Image.
2. **Visual Feature Extraction:** Passed through **EfficientNet-B0** (Pre-trained on ImageNet).
   - *Output:* A 1,280-dimensional feature vector.
3. **Dimensionality Reduction:** Compressed using **PCA (Principal Component Analysis)** to the top 50 components to prevent overfitting.
4. **Tabular Input:** Standard features (Bedrooms, Year Built) + Engineered Features (Neighborhood Density).
5. **Fusion:** Concatenation of [Tabular Features + 50 Visual PCA Features].

6. **Prediction:** The fused vector is fed into an **XGBoost Regressor** (tuned via Optuna) to predict the final price.


MULTIMODAL FUSION ARCHITECTURE DIAGRAM
Satellite Image (224x224)
EfficientNet-B0 CNN (Feature Extractor)
1280 Embeddings
PCA Reduction (50 dims)
Concatenate (Fusion)
XGBoost Regressor (Prediction Model)
Final Price Prediction
Tabular Data (Bedrooms, Sqft, etc.)
Feature Engineering

5. Results & Conclusion

5.1 Performance Comparison

We evaluated the model using 5-Fold Cross-Validation.

Model Architecture	Features Used	R² Score (Validation)
Baseline Model	Tabular Data Only (Sqft, Bed, Bath, etc.)	0.8950
Multimodal Fusion	**Tabular + Satellite Visuals**	**0.9112**

5.2 Conclusion

The integration of satellite imagery provided a clear performance boost (+1.5% R^2), improving the model accuracy. This confirms that **visual context is a quantifiable signal** in real estate valuation. Future improvements could involve fine-tuning the CNN on specific architectural styles rather than using generic ImageNet weights.