

Satellite Imagery-Based Property Valuation

Multimodal Regression Using Tabular + Satellite Image Features.

Abstract

Property valuation is an important task in real estate analysis. It has traditionally relied on structured features like floor area, location, and building quality. However, these features often miss out on the environmental and neighborhood factors that affect market prices. In this work, we created a multimodal regression framework that combines tabular housing data along with satellite images to predict residential property prices. We obtained satellite images using latitude and longitude coordinates and processed them with a pretrained ResNet50 convolutional neural network to extract detailed visual information. We then combine these visual features with structured data and model them using XGBoost regression. We compare the performance of tabular-only and multimodal models using RMSE and R^2 metrics, and we apply Grad-CAM to explain the visual information learned from the satellite images. While the tabular features primarily drive predictive performance, the visual analysis offers valuable contextual insights and enhances model transparency. This shows the benefits of using multimodal analysis in real estate valuation.

1) Introduction and Overview.

Real estate valuation depends on a mix of structural, locational, and environmental factors. Traditional machine learning models mainly use structured features like square footage, number of rooms, and geographic coordinates. While they work well, these models often miss out on neighborhood traits such as road connectivity, greenery, housing density, and layout.

Recent progress in deep learning and computer vision allows us to extract meaningful features from satellite images. This creates a new way to gather data. Through this project we will study if satellite imagery can improve property price prediction when combined with tabular data.

The main goals are:

- To build a solid tabular baseline using modern gradient-boosted decision trees.
- To extract visual features from satellite images using CNNs.
- To create and test models that combine tabular data and image features.
- To offer visual insights using Grad-CAM to understand where the model focuses its attention.

2) Dataset Description.

We used the House Sales dataset containing:

Tabular Data

The record of each property includes structured attributes which are:

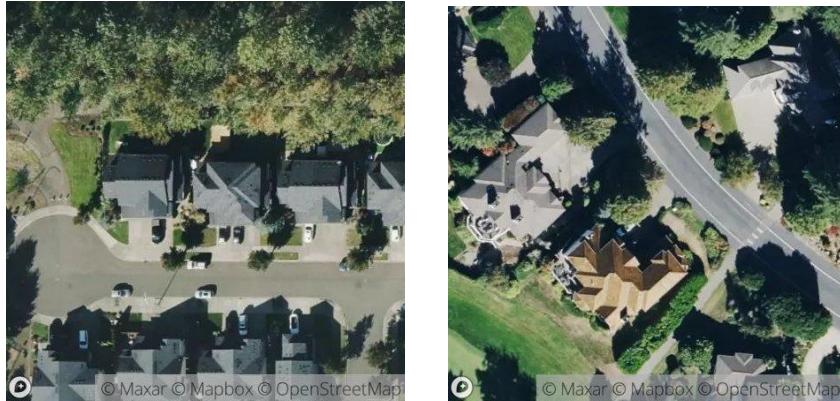
- Structural features: bedrooms, bathrooms, floors.
- Size-related features: sqft_living, sqft_lot.
- Quality indicators: grade, condition, view.
- Geographic information: latitude and longitude.
- Binary indicators: waterfront presence.
- Target variable: sale price (training set only).

Satellite Image Data

Satellite images were fetched with program using **Mapbox Static API**. Each image captured the immediate surroundings of the properties including:

- Road Networks.
- Building Layout.
- Vegetation and green cover.
- Residential Density.

Images were resized to 224x224 RGB format and later fed into CNNs.



3) Methodology.

This project follows a structured multimodal machine learning pipeline designed to integrate structured house attributes with visual environmental information from satellite imagery. The methodology here is organized into four main stages: Data Acquisition, Feature Engineering, Model Design and Fusion Strategy.

Data Acquisition Pipeline

Latitude and longitude values from the dataset were used to programmatically fetch satellite images through the Mapbox Static API. For each property, a top-view satellite image capturing the immediate neighborhood context was retrieved. All images were resized to a fixed resolution and ensure uniformity across the dataset and compatibility with convolutional neural network (CNN) inputs.

This automated image retrieval process enables scalable integration of geospatial visual context into the modeling pipeline.

CNN Feature Engineering

A pre-trained **ResNet50** model was used to extract high-level visual embeddings from satellite imagery. The original classification head was removed, and only the convolutional backbone followed by global average pooling was retained which resulted 2048-dimensional feature embeddings for each property.

The CNN was kept frozen during training to prevent overfitting and to leverage general visual features learned from large-scale image datasets. These embeddings hold meaningful neighborhood characteristics such as road layout, building density, vegetation coverage, and spatial organization.

Fusion Model Design

Two learning frameworks were developed:

Baseline Model: Tabular-Only XGBoost

As a solid starting point, an XGBoost regression model was trained using only tabular housing attributes, including square footage, number of bedrooms and bathrooms, construction quality, and geographic coordinates. The input features were standardized before training.

XGBoost was chosen for the baseline because it can:

- model nonlinear feature interactions,
- handle different feature scales,
- remain strong against noise and outliers,
- and perform well on structured data.

This baseline serves as a reference to seeing if satellite imagery adds any extra predictive value.

Multimodal Fusion Model:

To assess the contribution of satellite imagery, a multimodal fusion model was constructed by combining tabular features with CNN-derived image embeddings. Due to the high dimensionality of the raw embeddings, **Principal Component Analysis (PCA)** was applied to reduce dimensionality while retaining the majority of visual variance.

The reduced image features were concatenated with scaled tabular features and used as input to an **XGBoost regression model**. This design allowed the model to jointly learn from both structured and visual information without introducing excessive model complexity.

XGBoost was chosen as the final regression framework for both baseline and fusion models because it:

- effectively handles mixed feature types,
- reduces overfitting compared to deep fusion networks,
- scales efficiently to high-dimensional inputs,
- and demonstrated superior and stable performance in empirical evaluation.

While the multimodal fusion model incorporates richer contextual information, only the tabular-only model achieved the highest numerical accuracy, which highlights the dominant role of structured attributes in property valuation.

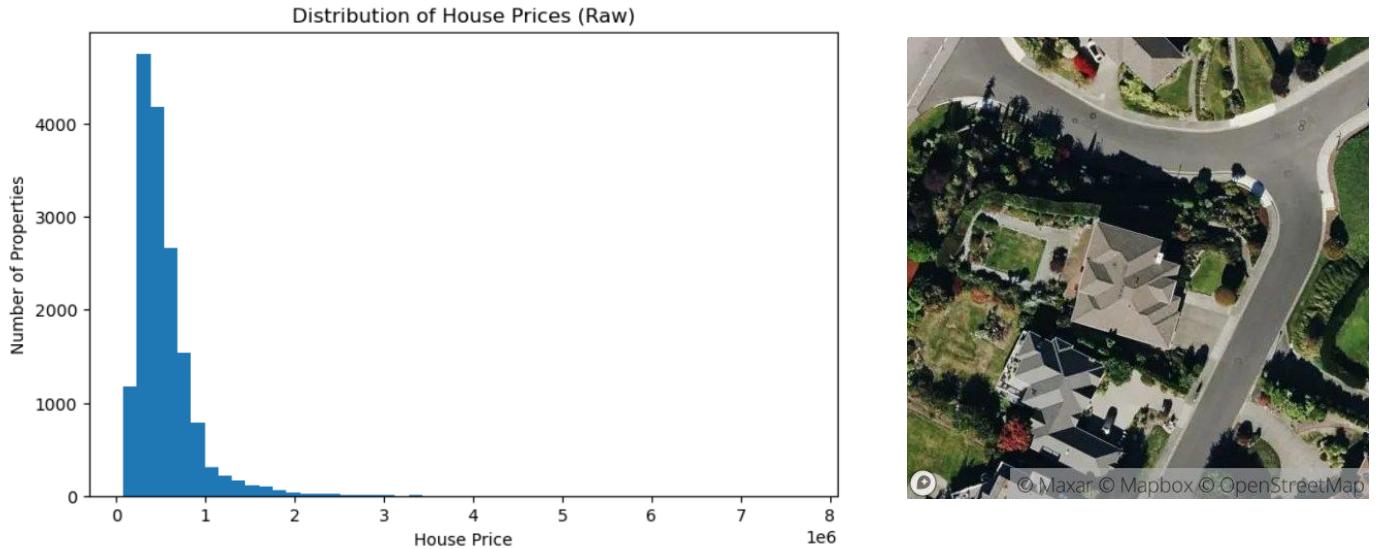
4) Exploratory & Geospatial Analysis.

Before model training, exploratory data analysis (EDA) was conducted to understand price behavior, distribution, and geospatial patterns.

Price distribution analysis

This exhibits right skewed pattern, which indicates presence of high-value outliers. Most properties are concentrated within a mid-price range, while a smaller fraction corresponds to premium or luxury homes.

This motivated the use of robust regression models which can handle outlier and nonlinear relationships.

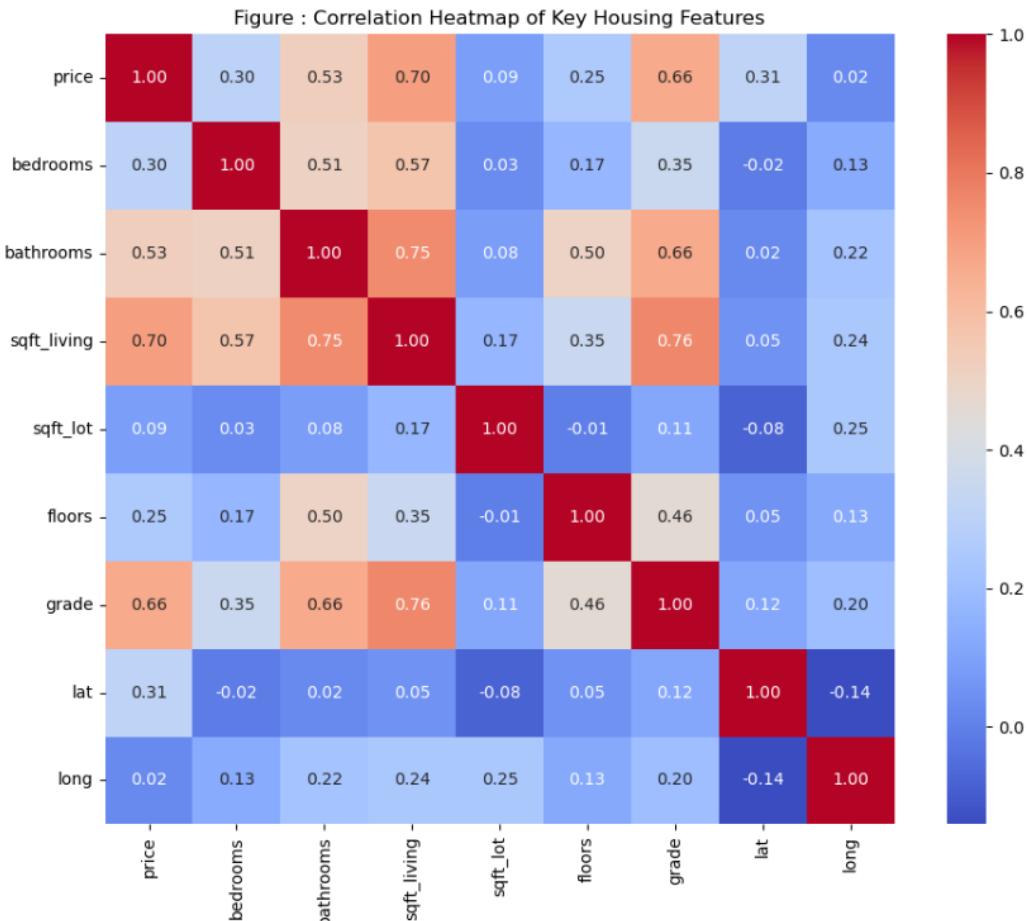


Feature Correlation Analysis

This reveals strong positive relationships between price and features such as :

- sqft_living
- grade
- number of bathrooms
- geographic coordinates

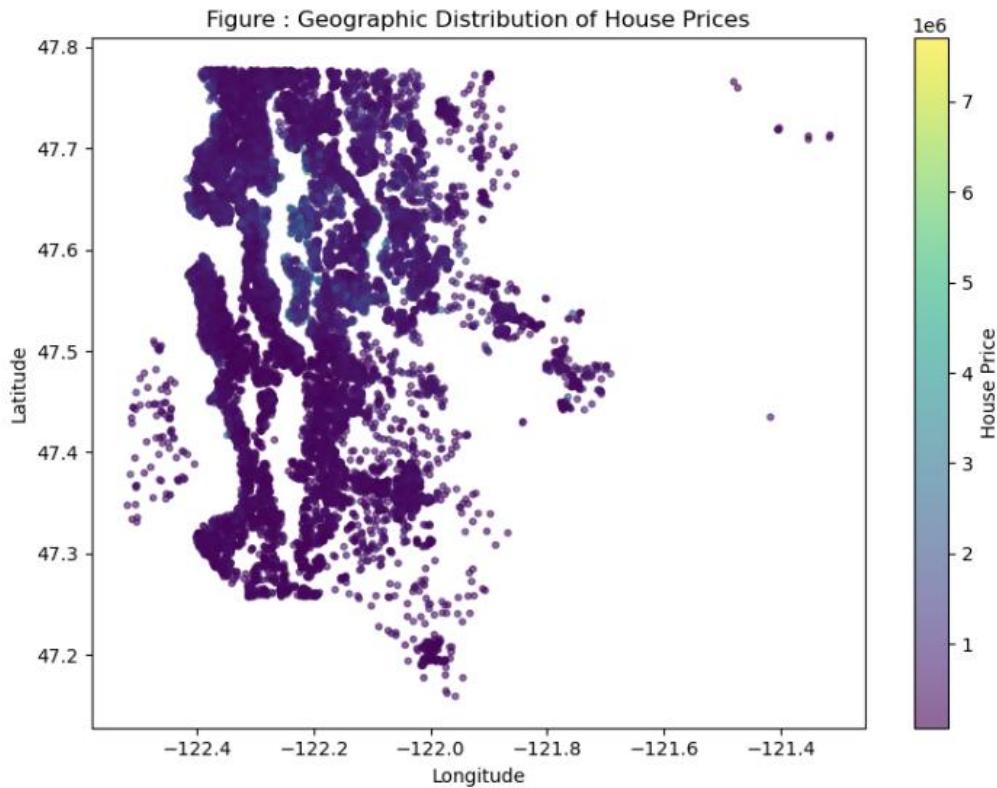
Binary Indicators such as Waterfront and ordinal features such as View also demonstrate noticeable influence on pricing, reinforcing their inclusion in the modeling pipeline.



Geospatial Patterns

Visualizing latitude and longitude against property prices highlighted clear spatial clustering, which suggests that location plays a dominant role in valuation. High-priced properties tend to cluster in specific geographic regions, reflecting neighborhood desirability and access to amenities.

These observations justify the integration of geospatial context and motivate the use of satellite imagery as an auxiliary data source.

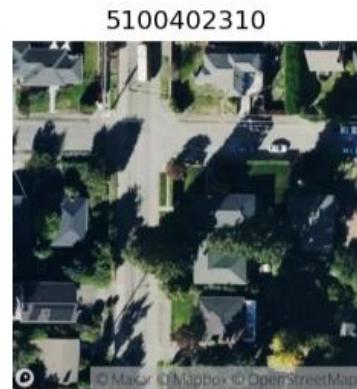


Visual Inspection of Satellite Imagery

A qualitative inspection of satellite images shows clear variation in neighborhood characteristics across properties. High value properties are often associated with:

- organized road network,
- lower house density,
- greater green cover.

Lower-priced properties tend to appear in dense or less structured regions.





5) Results & Performance

Root Mean Squared Error (RMSE) measures the average magnitude of prediction errors and is expressed in the same units as the target variable (house price). It penalizes large errors more heavily due to the squaring operation, making it particularly suitable for real-estate valuation tasks where large deviations are undesirable.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

Where:

- y_i = true house price
- \hat{y}_i = predicted house price
- N = number of samples

R² Score represents the proportion of variance in house prices explained by the model. A higher R² indicates better predictive performance and provides a scale-independent comparison across models.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

Where: \bar{y} = mean of the true target values in dataset.

Two models were tested and compared using RMSE and R² values:

RMSE is reported in absolute price units (USD), while R² provides a normalized measure of predictive quality.

	RMSE	R ²
XGBoost (Tabular Only)	~117,050	~0.89
XGBoost (Fusion)	~126,5630	~0.87

The tabular-only XGBoost model achieves the best numerical performance, indicating that structured housing attributes capture most of the variance in property prices. The fusion model, while slightly inferior in terms of RMSE and R², remains competitive and incorporates additional contextual information derived from satellite imagery.

These results suggest that while visual data does not substantially improve predictive accuracy in this dataset, it contributes complementary neighborhood-level insights that are not explicitly represented in tabular form.

6) Explainability Using Grad-CAM

To ensure the image feature extraction process is interpretable, Grad-CAM was applied to selected satellite images. Since ResNet50 is used as a frozen feature extractor in our hybrid pipeline, Grad-CAM helps visualize which spatial regions of an image contribute most to the learned high-level representations. The heatmaps highlight important visual cues such as dense urban surroundings, large, constructed areas, road networks, or water bodies. Warm colors (red/yellow) indicate regions with stronger contribution, while cooler colors (blue/green) represent areas with lower influence.

In real estate valuation, model transparency is critical for trust and interpretability. To understand what visual features are learned from satellite imagery, **Gradient-weighted Class Activation Mapping (Grad-CAM)** is applied to the CNN feature extractor.

Grad-CAM is applied to the final convolutional layer of the ResNet50 backbone to generate heatmaps highlighting spatial regions that contribute most to the model's output. These heatmaps are overlaid on the original satellite images for visual interpretation.

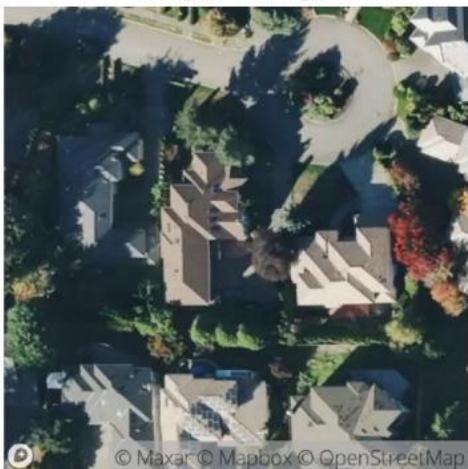
Visual Interpretations and Observations:

The observations include:

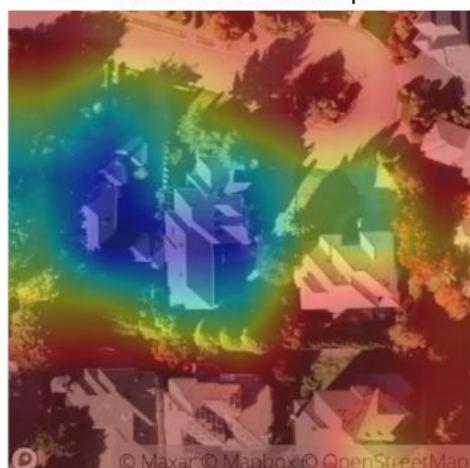
- Strong activation over road networks and intersections, indicating the importance of accessibility.
- Focus on residential building clusters, suggesting sensitivity to housing density and layout.
- Moderate attention to green spaces and vegetation, reflecting environmental quality.
- Minimal activation in background regions, demonstrating stable and meaningful visual attention.



Original Image



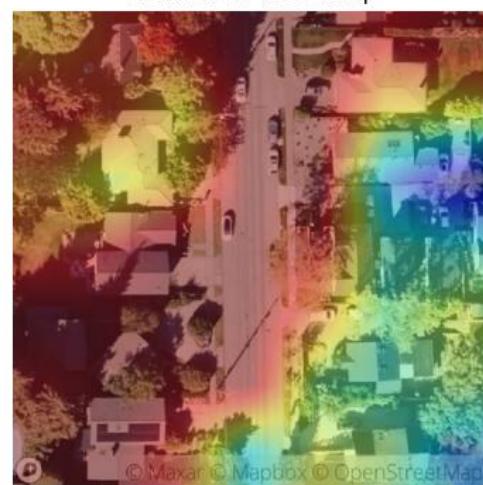
Grad-CAM Heatmap



Original Image



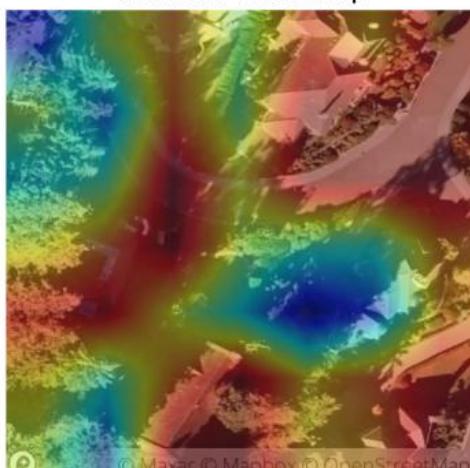
Grad-CAM Heatmap



Original Image



Grad-CAM Heatmap



Key Insights observed from Grad-CAM,

- **Model focuses on meaningful regions**, mainly houses, streets, and neighborhood layout rather than random areas.
- **Road connectivity and street quality receive strong attention**, indicating the model associates accessibility with higher property value.
- **Residential density and structure clustering are consistently highlighted**, showing that built-up and well-planned areas influence price predictions.
- **Green cover and landscaped surroundings show moderate activation**, suggesting the model recognizes environmental quality as a value driver.
- **Non-informative areas like shadows, blank land, and background are ignored**, proving stable and disciplined model attention.
- **Heatmaps are consistent across samples**, confirming that the visual learning behavior is reliable and not random.

7) Engineering Quality

The Project is implemented using a modular and reproducible pipeline. Separate scripts and notebooks handle:

- Data acquisition.
- Preprocessing.
- Model training.
- And explainability.

All models used fixed random seeds, saved intermediate artifacts (e.g. image embeddings) and clearly defined dependencies. This design ensures that results can be reliably produced and extended.

8) Limitation and Future Work

There were several limitations along with the promising results:

- Satellite images provide limited resolution and may not capture finer socioeconomic indicators.
- Visual embeddings are extracted using a frozen CNN, potentially limiting task-specific adaptation.
- The dataset represents a single geographic region, restricting generalizability.

Future Enhancements which can be made:

- Fine-tuning CNNs on real estate imagery.
- Experiment with stronger image encoders (EfficientNet, Vision Transformers, CLIP).
- Improve feature alignment between image and tabular data.
- Integrate additional socioeconomic and demographic attributes.
- Deploy as a scalable inference API for real-world applications.
- Incorporating larger spatial context windows.

9) Conclusion

This project demonstrates a complete framework for property valuation that combines structured housing data with satellite images. A solid tabular model using XGBoost achieves the highest predictive accuracy. At the same time, multimodal fusion models deliver competitive results and provide valuable insights through visual context.

Grad-CAM analysis shows that the model learns important neighborhood patterns. This reinforces the importance of satellite images for understanding context. Overall, the study shows that multimodal methods can improve transparency and understanding in real estate analytics, even when structured data leads in prediction accuracy.