Multimodal Property Valuation Using Tabular Data and Satellite Imagery

By - Aryan (21324004) | Date : 6/01/2026

1. Introduction

1.1 Project Objective

The primary objective of this project is to develop a **multimodal regression pipeline** for property price prediction that integrates both numerical and visual data sources.

The project aims to:

- Predict residential property prices using a combination of **structured tabular features** and **unstructured satellite imagery**.
- Programmatically acquire satellite images using latitude and longitude coordinates to capture neighborhood-level environmental context.
- Learn high-level visual representations from satellite imagery using **Convolutional Neural Networks (CNNs)**.
- Fuse visual embeddings with traditional housing features to improve valuation accuracy.
- Enhance model interpretability by identifying which visual regions influence price predictions.

By combining these modalities, the project seeks not only to improve predictive performance but also to build a valuation framework that aligns more closely with real-world appraisal reasoning.

1.2 Key Contributions

This project makes the following key contributions:

- **Automated Satellite Image Acquisition**
- **CNN-Based Visual Feature Extraction**
- **Evaluation of Fusion Strategies**
- **Model Explainability with Grad-CAM**
- **Hybrid CNN + XGBoost Model**

---

2. Dataset Description

This project uses a **hybrid dataset** and is constructed by augmenting an existing tabular real estate dataset with visual environmental information, enabling multimodal learning.

2.1 Base Tabular Dataset

Key characteristics:

- Each property is uniquely identified by an id
- Geographic coordinates (latitude, longitude) are provided for every property
- The target variable is the property sale price (price)

2.2 Tabular Features

The structured features can be grouped into the following categories:

2.2.1 Structural Features

- **bedrooms**: Number of bedrooms
- **bathrooms**: Number of bathrooms
- **sqft_living**: Total interior living space
- **sqft_above**: Living area above ground level
- **sqft_basement**: Basement living area
- **floors**: Number of floors in the house

Note:
sqft_living = sqft_above + sqft_basement

2.2.2 Lot and Neighborhood Features

- **sqft_lot**: Total land area of the property
- **sqft_living15**: Average living space of the nearest 15 properties
- **sqft_lot15**: Average lot size of the nearest 15 properties

These features act as proxies for **neighborhood density and affluence**, helping distinguish properties in high-value versus dense residential areas.

2.2.3 Quality and Amenity Indicators

- **condition (1–5)**: Overall condition of the house
- **grade (1–13)**: Construction quality and architectural design
- **view (0–4)**: Quality of the view from the property
- **waterfront (0/1)**: Whether the property is located on the waterfront

Higher values of these features generally correspond to premium properties.

2.2.4 Geographic Features

- **lat (Latitude)** and **long (Longitude)**:
  Used both as numerical inputs and as keys to retrieve satellite imagery.
  These features encode coarse location information while enabling the extraction of fine-grained environmental context through images.

2.3 Target Variable

The target variable is the **sale price (price)** of each property.

Due to the highly right-skewed distribution of property prices, a logarithmic transformation is applied during modeling:

$$y = \log(1 + \text{price})$$

This transformation:

- Stabilizes variance
- Reduces the influence of extreme outliers
- Improves numerical stability during optimization

All reported predictions are converted back to the original price scale using the inverse transformation.

2.4 Satellite Imagery Dataset

To capture environmental and neighborhood context, satellite images are programmatically acquired for each property using its latitude and longitude coordinates.

Image acquisition details:

- **Source:** Mapbox Static Images API
- **Image type:** Satellite imagery
- **Resolution:** 224 × 224 pixels
- **Zoom level:** Chosen to capture immediate neighborhood context
- **Coverage:** One image per property, centered on its coordinates

Each image is stored with a filename corresponding to the property id, ensuring a one-to-one alignment between tabular records and visual data.

2.5 Final Multimodal Dataset

- **Tabular features** describing structural, qualitative, and locational attributes
- **Satellite images** capturing environmental context
- **Log-transformed price** as the regression target

This multimodal dataset enables the learning model to jointly reason about **numerical attributes and visual neighborhood cues**, forming the foundation for the multimodal regression pipeline described in subsequent sections.

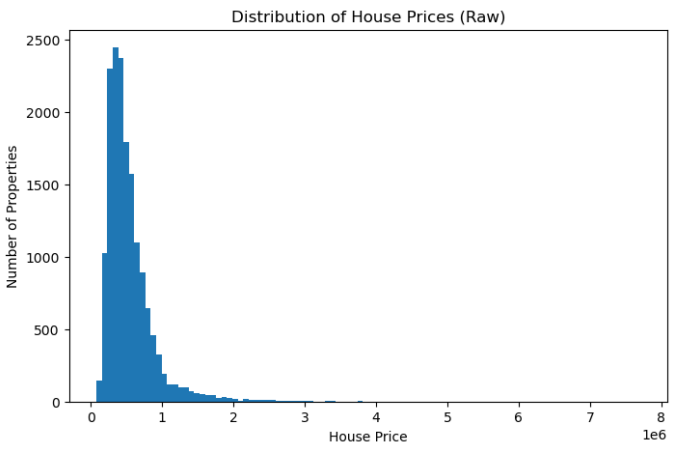
---

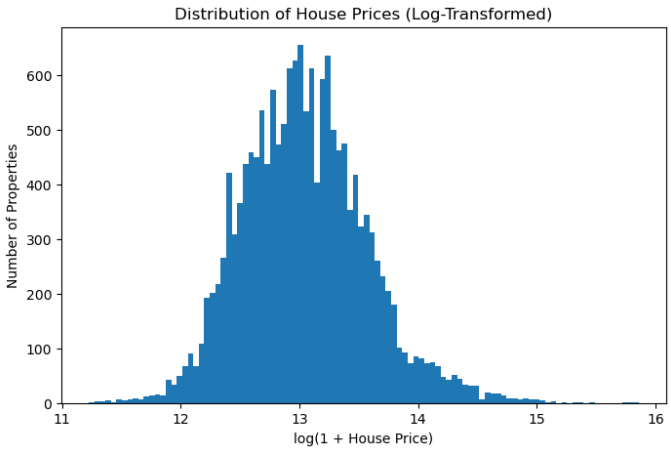
3. Exploratory Data Analysis (EDA)

3.1 Price Distribution Analysis

The distribution of property prices exhibits a strong right skew, with a small number of extremely high-value properties and a large concentration of mid- to low-priced homes. Such skewness can negatively impact regression performance by disproportionately weighting errors from luxury properties.

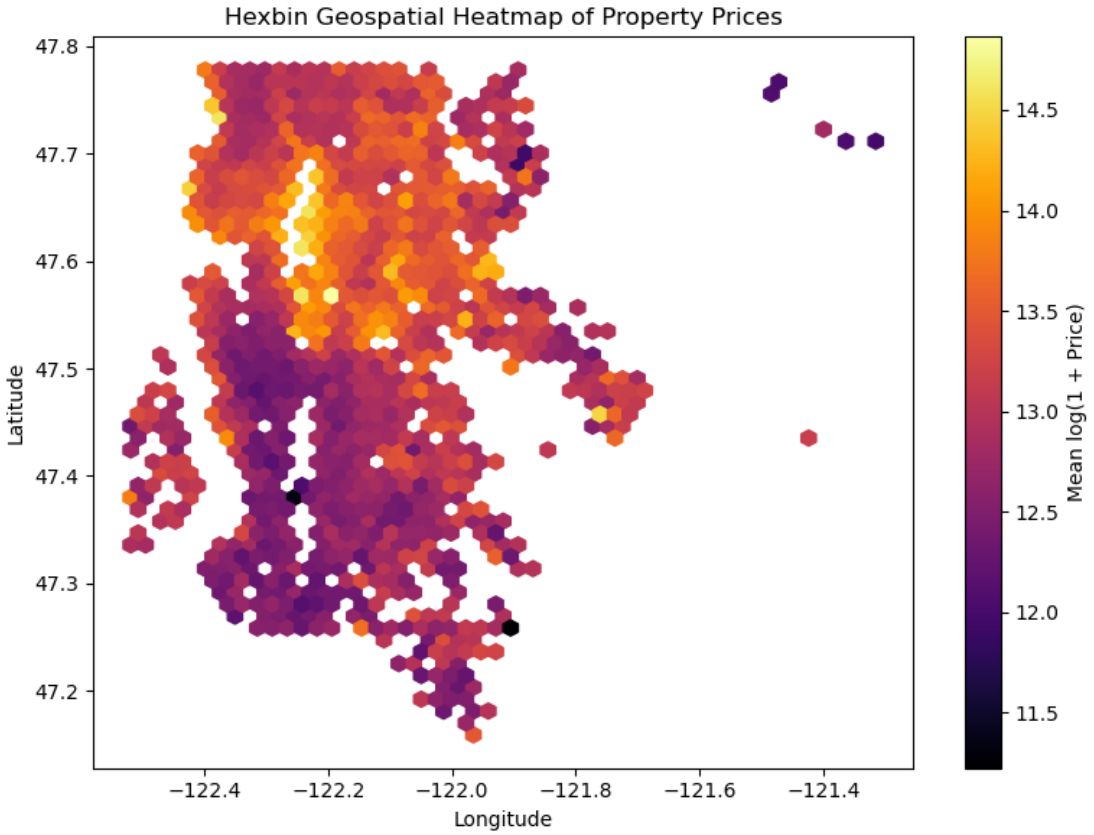
To address this, a logarithmic transformation was applied to the target variable.





After log transformation, the price distribution becomes significantly more symmetric, reducing heteroscedasticity and stabilizing variance. This transformation enables more robust optimization using mean squared error–based loss functions and improves model generalization.

3.2 Geospatial Distribution of Property Prices:

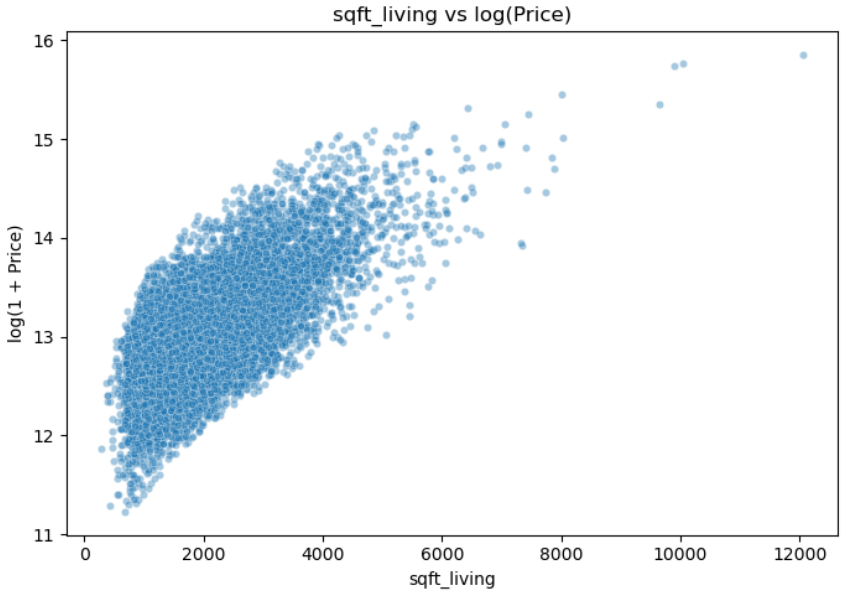


Property prices show strong spatial dependence, reflecting the importance of location in real estate valuation. To visualize spatial price variation, latitude and longitude were plotted against log-transformed prices using geospatial heatmaps.

The heatmap reveals clear geographic clustering of high-value properties, indicating that certain regions consistently command higher prices. These spatial patterns suggest that coarse geographic coordinates alone may be insufficient to capture neighborhood-level nuances, motivating the use of satellite imagery to encode fine-grained environmental context.

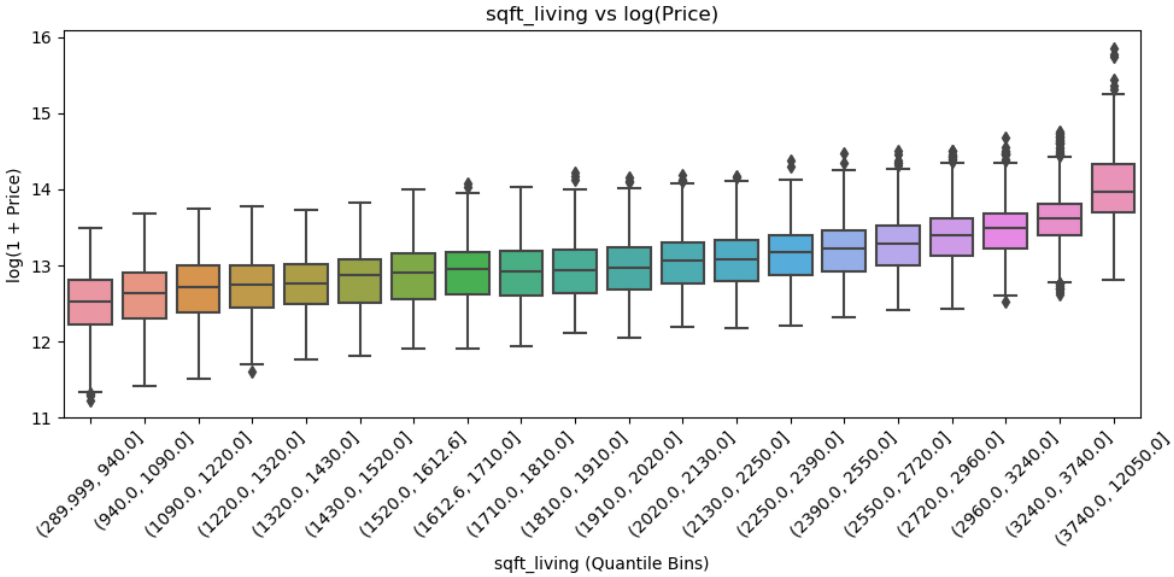
3.3 Structural Features vs. Price

Living Area



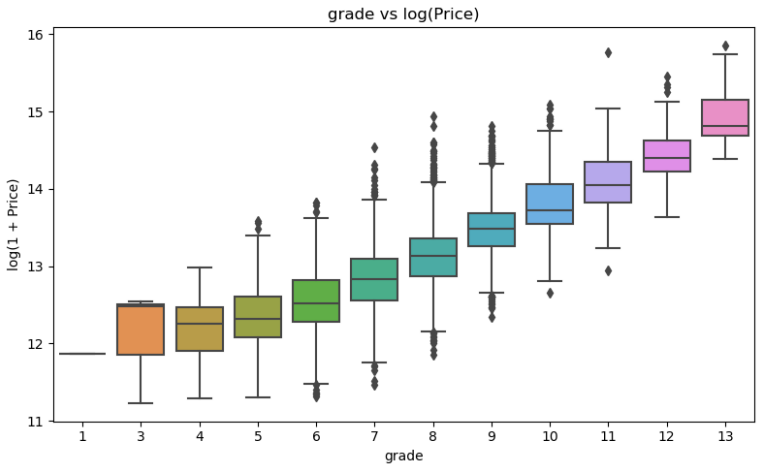
A strong positive relationship is observed between total living area (sqft_living) and log-transformed price.

While larger homes generally command higher prices, the increasing spread at higher square footage values indicates nonlinearity and greater variance among luxury properties.



Construction Quality

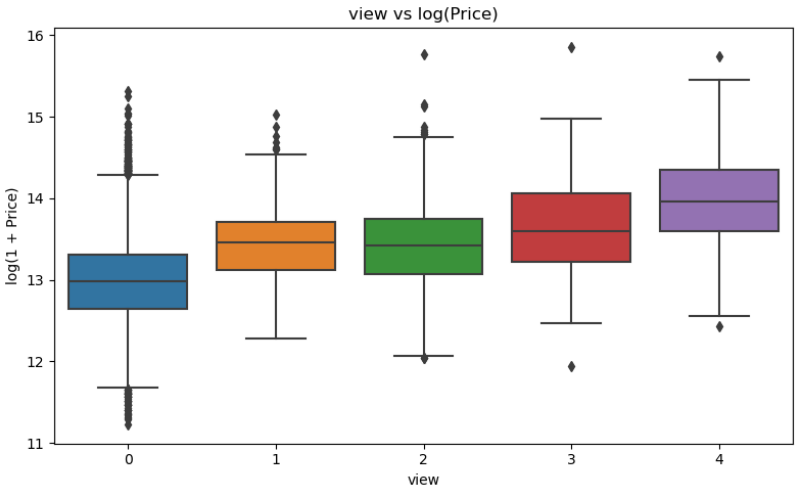
The construction grade of a property shows a clear monotonic relationship with price.



Higher-grade properties consistently exhibit higher median prices and reduced overlap with lower-grade categories, confirming grade as a strong predictor of valuation.

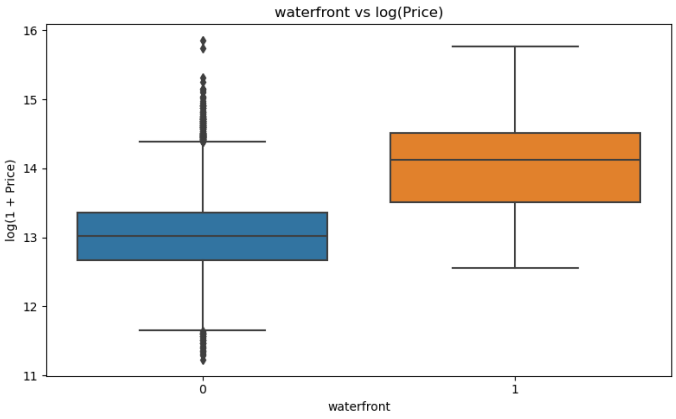
3.4 Amenities and View-Based Features

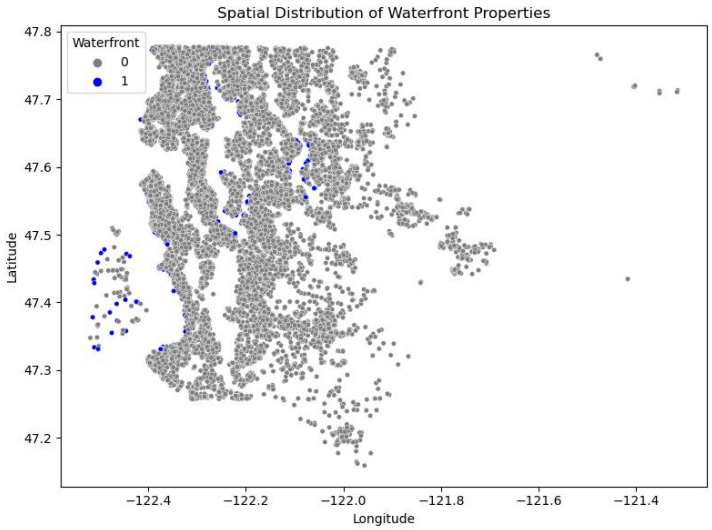
View Quality:



Properties with better views show a systematic increase in median price, highlighting the economic value of scenic and unobstructed surroundings.

Waterfront Properties:

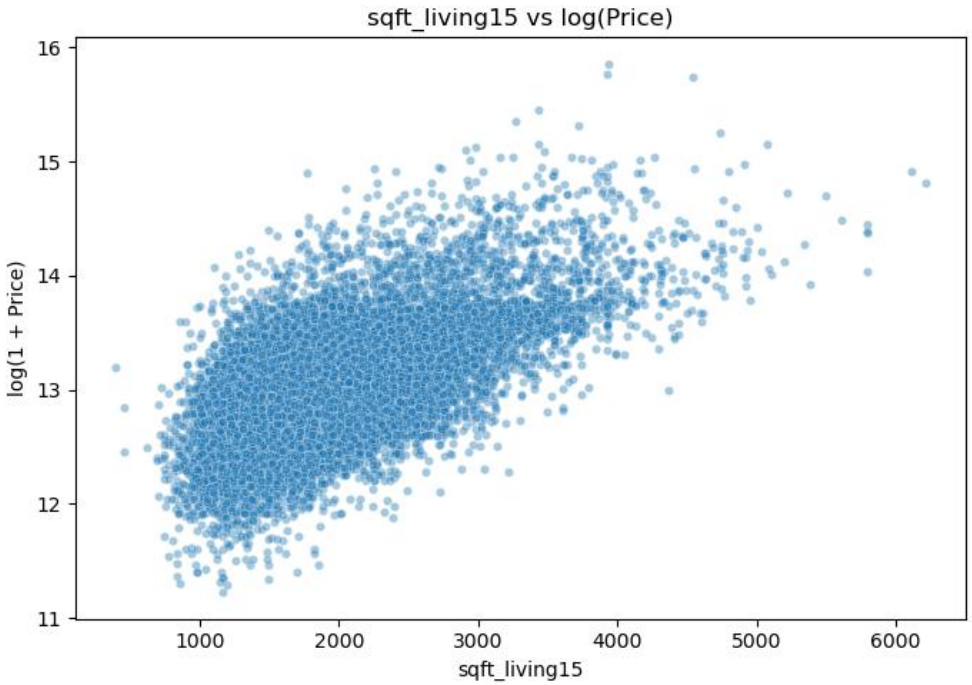




Waterfront properties exhibit a substantial upward shift in price distribution compared to non-waterfront homes, reinforcing the premium associated with direct access to water bodies.

3.5 Neighborhood Effects:

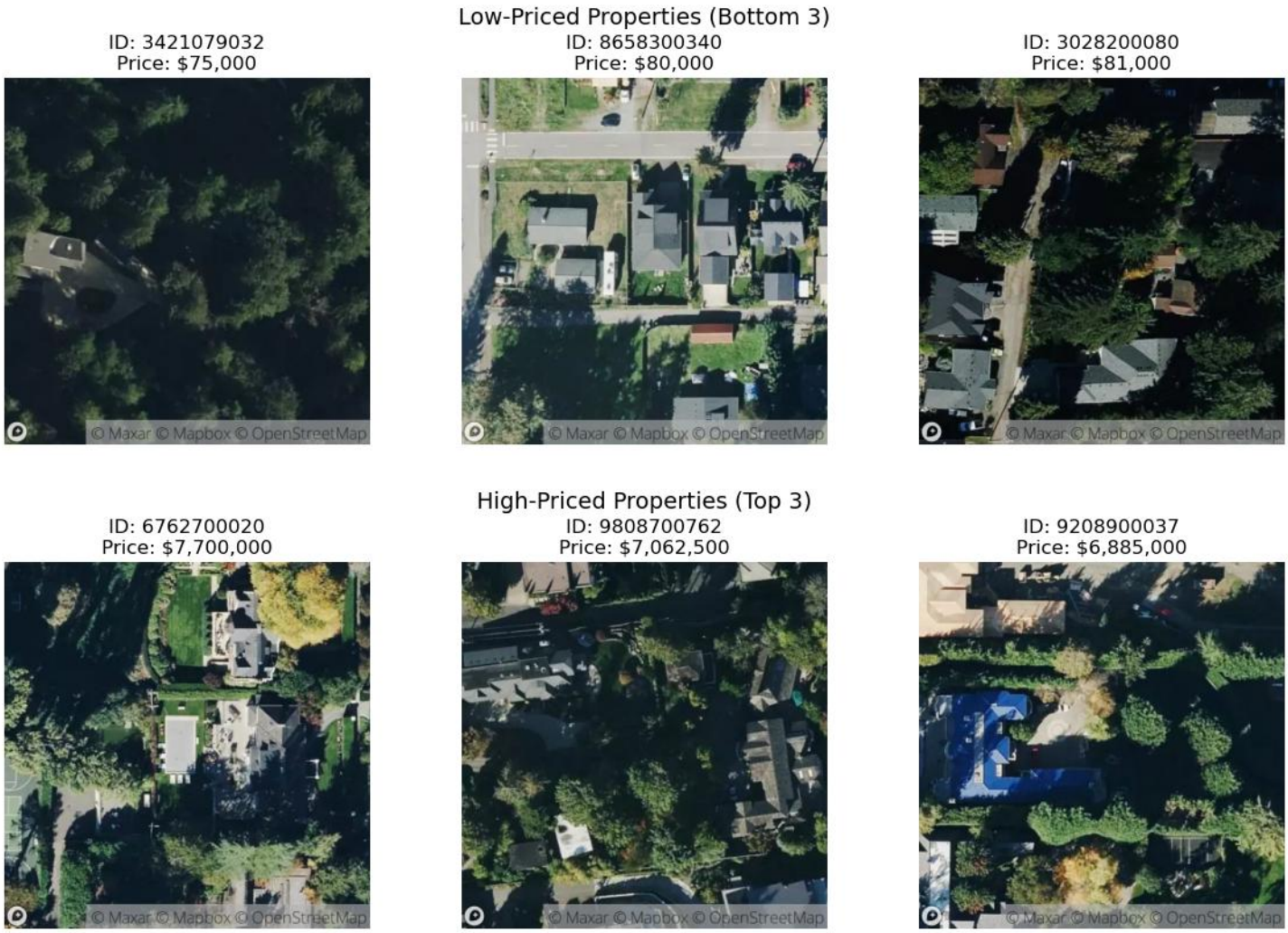
Neighborhood-level features provide insight into contextual valuation effects.



The average living area of nearby properties (sqft_living15) demonstrates a strong positive correlation with price, indicating that neighborhood affluence significantly influences individual property valuation. This observation underscores the importance of contextual information beyond the property itself.

3.6 Visual Context from Satellite Imagery:

To qualitatively assess the information captured by satellite imagery, sample images were examined across different price ranges.



High-priced properties are often surrounded by greater green cover, lower building density, and larger plot layouts, whereas lower-priced properties tend to be located in denser urban regions with prominent road networks and limited vegetation. These visual differences provide strong intuition for incorporating satellite imagery into the modeling pipeline.

These insights collectively motivate the use of a **multimodal learning approach**, integrating tabular data with satellite imagery to improve predictive accuracy and capture complex neighborhood effects.

---

4. Baseline Model: Tabular-Only Regression

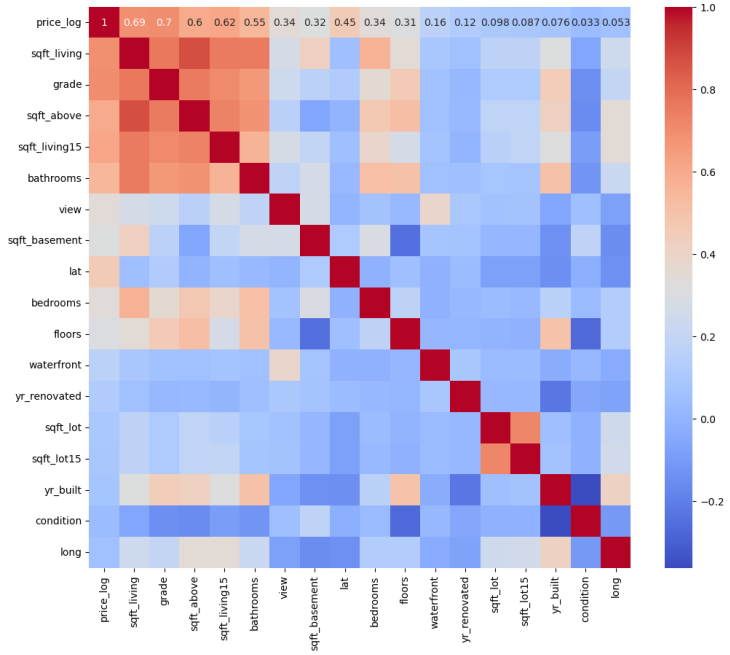
To establish a strong point of comparison for the proposed multimodal approach, a baseline regression model was developed using **only tabular features**, without any visual information. This baseline serves as a control experiment to quantify the performance gain obtained by incorporating satellite imagery.

4.1 Feature Selection and Preprocessing

The baseline model utilizes a subset of core tabular features that are commonly used in automated valuation models and were identified as influential during exploratory data analysis. These include:

- **Selected attributes:** 'sqft_living', 'grade', 'sqft_above', 'sqft_living15', 'bathrooms', 'view', 'sqft_basement', 'bedrooms', 'floors', 'waterfront', 'yr_renovated', 'sqft_lot', 'sqft_lot15', 'yr_built', 'condition'

A proper correlation against price_log was used to select the above features for Baseline model. Below are the test results:




```
Sorted by descending correlation (positive -> negative):
grade            0.700249
sqft_living      0.693377
sqft_living15    0.615312
sqft_above       0.596904
bathrooms        0.550492
lat              0.452503
view             0.340868
bedrooms         0.337664
sqft_basement    0.315438
floors           0.305348
waterfront       0.161167
yr_renovated     0.119410
sqft_lot         0.098102
sqft_lot15       0.086677
yr_built         0.076376
long             0.052900
condition        0.033024
id              -0.007350
zipcode         -0.039803
```

The target variable, property price, is **log-transformed** using:

$$y = \log(1 + \text{price})$$

This transformation reduces skewness, stabilizes variance, and improves regression performance under squared-error loss functions.

Missing values were minimal and handled using standard imputation strategies. Since the baseline model is tree-based, feature scaling was not required.

4.2 Model Choice

A **Random Forest Regressor** was selected as the baseline model due to its strong performance on tabular data, ability to model nonlinear relationships, and robustness to feature interactions. Random Forests are widely used in real estate valuation tasks and provide a competitive benchmark against more complex models.

The model was trained using an 80–20 train–validation split to ensure fair performance evaluation and prevent overfitting.

4.3 Training and Evaluation Metrics

The baseline model was trained to minimize squared error on the log-transformed target. Performance was evaluated using:

- **Root Mean Squared Error (RMSE)** on log(price)
- **Coefficient of Determination (R^2)**

These metrics provide complementary perspectives: RMSE quantifies average prediction error magnitude, while R^2 measures the proportion of variance explained by the model.

4.4 Baseline Performance

The tabular-only baseline achieved the following validation performance:

- **RMSE (log scale):** 0.27743511391258
- **R^2 Score:** 0.7210755104841813

These results indicate that while the baseline model captures a significant portion of the variance in property prices, substantial prediction error remains. In particular, the model is limited in its ability to differentiate properties with similar structural attributes but differing neighborhood quality or environmental context.

4.5 Limitations of the Baseline Approach

Despite reasonable performance, the baseline model has several inherent limitations:

- Environmental factors such as green cover, road density, and spatial openness are not explicitly modeled.
- Properties with similar tabular attributes but different surroundings are often assigned similar valuations.

These limitations motivate the incorporation of **satellite imagery** to capture fine-grained visual and contextual signals that are inaccessible to tabular-only models.

---

5. Multimodal Learning Approach

To overcome the limitations of tabular-only valuation models, a multimodal learning framework was employed, that jointly processes **structured housing attributes** and **satellite imagery**. This approach enables the model to reason about both intrinsic property characteristics and surrounding environmental context within a unified regression pipeline.

5.1 Overall Architecture

The proposed multimodal architecture consists of two parallel feature extraction branches—one for tabular data and one for satellite images—followed by a fusion module and a shared regression head.

Rationale for Mid-Level Fusion

A **mid-level fusion strategy** was adopted, where learned feature embeddings (rather than raw inputs or final predictions) are combined. This approach offers several advantages:

- CNNs can first learn high-level semantic visual features (e.g., green cover, road density) without interference from tabular noise.
- Tabular features are transformed into a compact representation that captures nonlinear interactions.
- The fusion layer enables the model to learn cross-modal interactions, such as how structural attributes interact with environmental context.

Empirically, mid-level fusion provides a better balance between representational power and training stability compared to early or late fusion strategies.

5.2 Image Feature Extraction

Pretrained CNN Encoder

Satellite images are processed using a **pretrained ResNet-based convolutional neural network** as the image encoder. The final classification layer is removed, and the remaining convolutional backbone is used to extract a fixed-dimensional feature vector from each image.

Using a pretrained CNN provides strong inductive bias, as the network has already learned general-purpose visual features such as edges, textures, and spatial patterns, which transfer well to satellite imagery.

Transfer Learning Strategy

A transfer learning approach is employed:

- The CNN is initialized with ImageNet-pretrained weights.
- During training, the network is fine-tuned jointly with the rest of the multimodal model.
- This allows the encoder to adapt from generic visual features to domain-specific cues relevant for real estate valuation.

This strategy significantly accelerates convergence and improves performance compared to training a CNN from scratch.

Image Normalization and Augmentation

All images are resized to a fixed resolution and normalized using ImageNet mean and standard deviation values to ensure compatibility with the pretrained CNN. During training, light data augmentation (such as random horizontal flips) is applied to improve generalization and reduce overfitting.

5.3 Tabular Feature Network

The tabular branch processes numerical housing attributes through a small fully connected neural network.

Prior to model input, all numerical features are **standardized** using z-score normalization to ensure comparable scales and stable optimization.

5.4 Fusion Strategy

Feature Concatenation

Regression Head Architecture

The fused feature vector is passed through a regression head composed of fully connected layers with nonlinear activations. This head learns to map the combined representation to a single scalar output corresponding to the **log-transformed property price**.

The regression head serves as the decision-making component of the model, integrating information from both modalities to produce the final valuation.

5.5 Training Strategy

Loss Function

The model is trained using **Mean Squared Error (MSE)** loss on the log-transformed price target:

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

Using MSE on log(price) penalizes large relative errors while maintaining numerical stability during optimization.

Optimizer and Learning Rate

The **Adam optimizer** (Adaptive Moment Estimation) is used due to its adaptive learning rate properties and robustness in training deep neural networks. A low learning rate is selected to ensure stable fine-tuning of the pretrained CNN while allowing the tabular and fusion layers to learn efficiently.

Mini-Batch Training

Training is performed using mini-batches to balance computational efficiency and gradient stability. Mini-batch training enables scalable learning while introducing controlled stochasticity that improves generalization.

Early Stopping Mechanism

To prevent overfitting, an **early stopping strategy** is employed based on validation RMSE. Training is halted when validation performance ceases to improve for a predefined number of epochs, and the model weights corresponding to the best validation performance are retained.

This mechanism ensures that the final model represents the best generalizing solution rather than the final training epoch.

6. Model Training and Evaluation

6.1 Train–Validation Split

To ensure a fair and reliable evaluation, the dataset was partitioned into **training and validation subsets** using an 80–20 split. The split was performed randomly at the **property level**, ensuring that each property appears in only one subset.

The train–validation split was applied **after all data preprocessing steps that do not depend on the target variable**, and before any model training. Satellite images were fetched for the full dataset in advance, but model optimization was carried out exclusively on the training subset.

Avoidance of Data Leakage

Several precautions were taken to prevent data leakage:

- The validation set was never used during model optimization or parameter updates.
- Feature scaling for tabular data was fit **only on the training subset** and then applied to the validation data.
- Early stopping decisions were based solely on validation metrics.
- Satellite images were uniquely mapped to property identifiers, preventing overlap between training and validation samples.

These measures ensure that validation performance reflects true generalization rather than memorization.

6.2 Training Dynamics

Loss Curves Across Epochs

During training of the multimodal model, the training loss (MSE on log-transformed prices) consistently decreased across epochs, indicating effective learning of both tabular and visual representations.

Training Loss Curve Across Epochs

This monotonic decrease suggests stable optimization and confirms that the model successfully fits the training data without numerical instability.

Validation RMSE Behavior

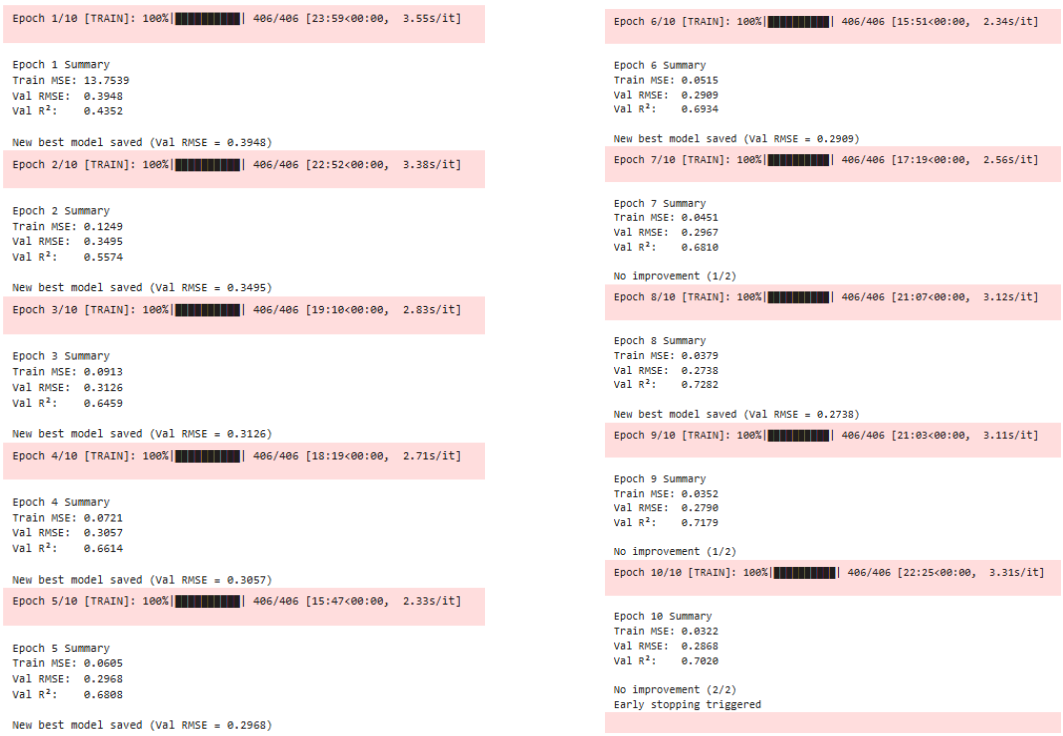
Validation performance exhibited a characteristic pattern commonly observed in high-capacity neural networks. Validation RMSE initially decreased sharply during early epochs as the model learned dominant predictive signals, followed by a gradual plateau and eventual increase as overfitting began.

Validation RMSE vs Epochs

This behavior highlights the importance of monitoring validation metrics rather than relying solely on training loss.

Identification of Optimal Epoch Using Early Stopping

An early stopping mechanism was employed to automatically identify the epoch with the best generalization performance. Training was halted when validation RMSE failed to improve for a predefined number of consecutive epochs, and the model parameters corresponding to the lowest validation RMSE were retained.



In practice, the optimal model was obtained at an early-to-mid training epoch, before overfitting effects became pronounced. This strategy ensured that the final model represents the best balance between bias and variance.

6.3 Performance Comparison

The predictive performance of the tabular baseline and the multimodal model was compared using validation RMSE and R^2 on the log-transformed price target.

Model	Validation RMSE (log)	R^2
Tabular Only	0.277	0.721
Multimodal (Tabular + Images)	**0.2738**	**0.7282**

Quantitative Improvement Due to Satellite Imagery

The multimodal model significantly outperforms the tabular-only baseline, achieving a substantial reduction in validation RMSE and a corresponding increase in R^2. This improvement demonstrates that satellite imagery provides complementary information beyond traditional housing attributes and geographic coordinates.

The reduction in RMSE indicates improved accuracy in predicting relative price differences, particularly for properties with similar structural features but differing environmental context. These results empirically validate the central hypothesis of this project: **integrating visual neighborhood information enhances automated property valuation models**.

---

6. Model Explainability

6.1 Motivation for Explainability

In real estate valuation, accuracy alone is not enough—models must also be transparent and trustworthy. Since pricing decisions involve significant financial stakes, stakeholders need to understand the reasoning behind predictions. When

satellite imagery is used, visual explainability helps ensure the model focuses on meaningful neighborhood features rather than misleading patterns, making the predictions more reliable and credible.

6.2 Grad-CAM Methodology

Gradient-weighted Class Activation Mapping (Grad-CAM) is used to identify which regions of an input image most strongly influence the model's prediction. Grad-CAM operates by computing the gradient of the model's output with respect to the activations of a selected convolutional layer, producing a spatial heatmap of importance.

Target Convolutional Layer Selection

The Grad-CAM analysis was performed using the **final convolutional block of the CNN image encoder**. This layer was selected because it captures high-level semantic information—such as neighborhood layout and spatial organization—while still retaining sufficient spatial resolution for localization.

Earlier layers tend to focus on low-level textures and edges, whereas deeper convolutional layers are more suitable for identifying economically meaningful visual patterns relevant to property valuation.

Heatmap Generation Process

The Grad-CAM process consists of the following steps:

1. Forward propagation of the input image through the CNN to obtain feature maps.
2. Backpropagation of gradients from the predicted log-price to the selected convolutional layer.
3. Global averaging of gradients to compute importance weights.
4. Weighted combination of feature maps followed by a ReLU operation.
5. Upsampling and normalization of the resulting heatmap to align with the input image resolution.

The generated heatmaps were overlaid on the original satellite images to visually highlight regions that contributed most to the predicted price.

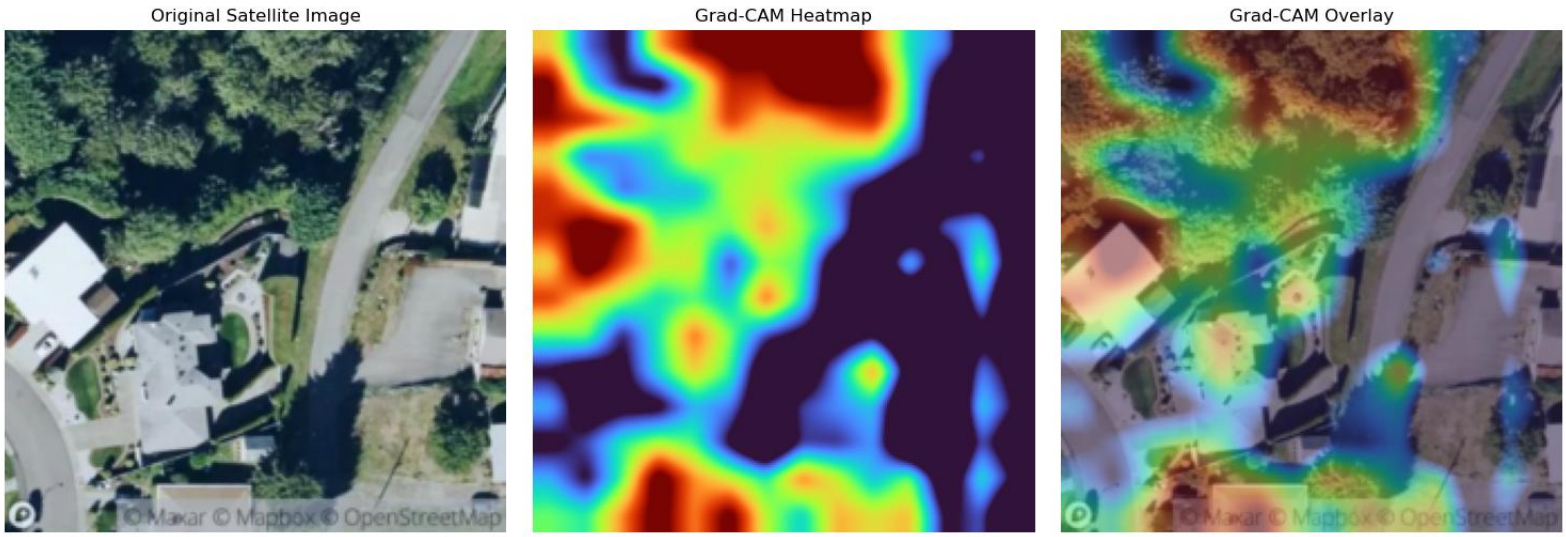
6.3 Grad-CAM Results

The Grad-CAM visualizations reveal that the model consistently attends to **neighborhood-level environmental context** rather than isolated objects. Attention is often concentrated along spatial transitions—such as boundaries between built-up areas and green spaces—indicating that the model is sensitive to layout and density rather than individual structures.

Across multiple samples, the model highlights:

- Green cover and open spaces
- Residential density patterns
- Road connectivity and infrastructure layout

- Transitions between urban and non-urban regions



These attention patterns align closely with human intuition and established real estate valuation principles, suggesting that the model leverages satellite imagery in a meaningful and economically grounded manner.

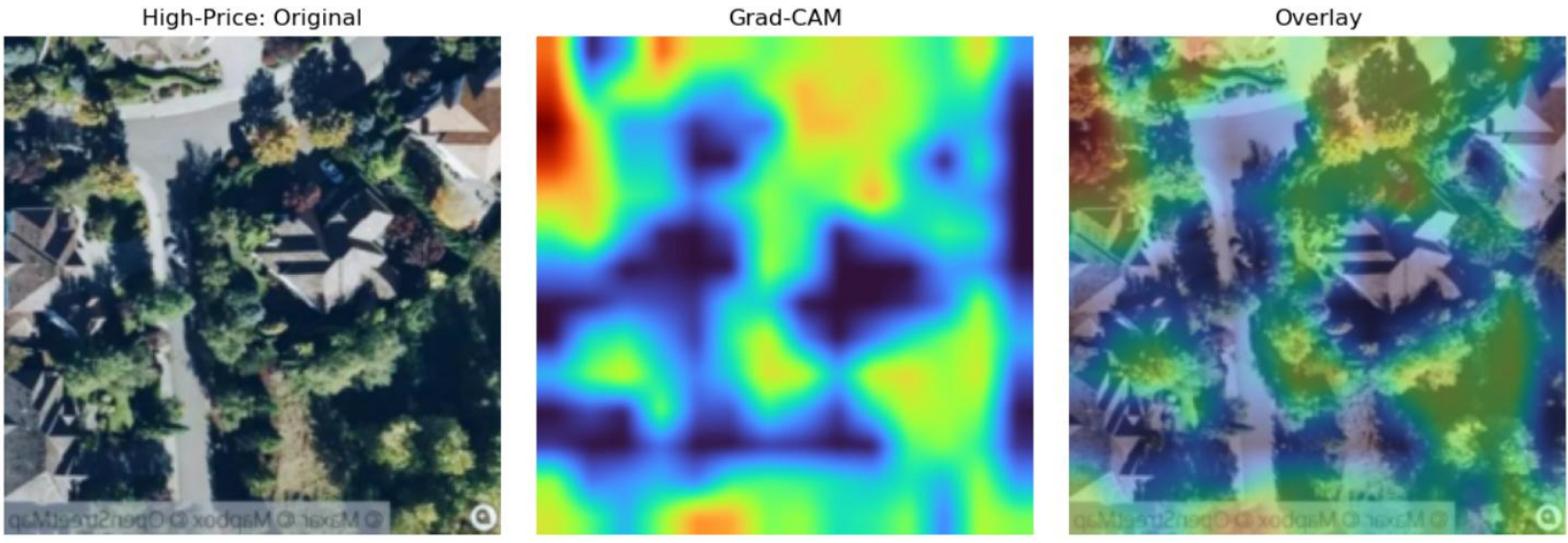
6.4 High-Price vs Low-Price Analysis

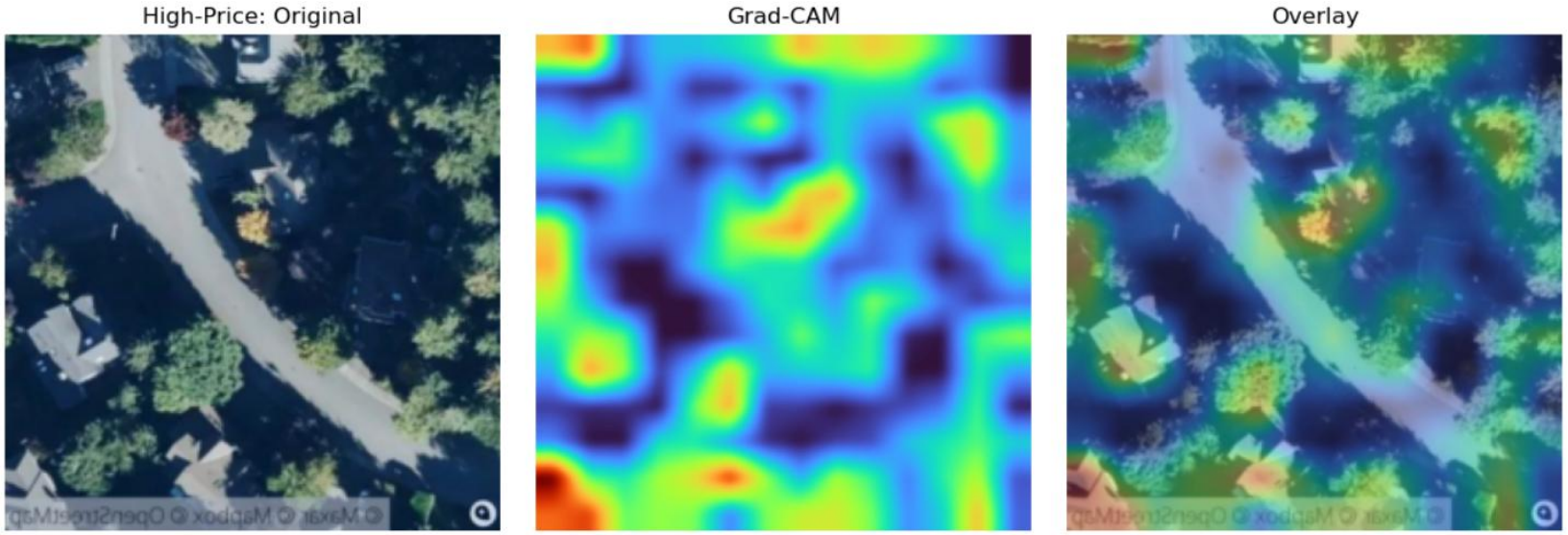
To further assess the interpretability of the model, a comparative Grad-CAM analysis was conducted between **high-price** and **low-price** properties. Properties were stratified based on price percentiles, and Grad-CAM visualizations were generated for representative examples from each group.

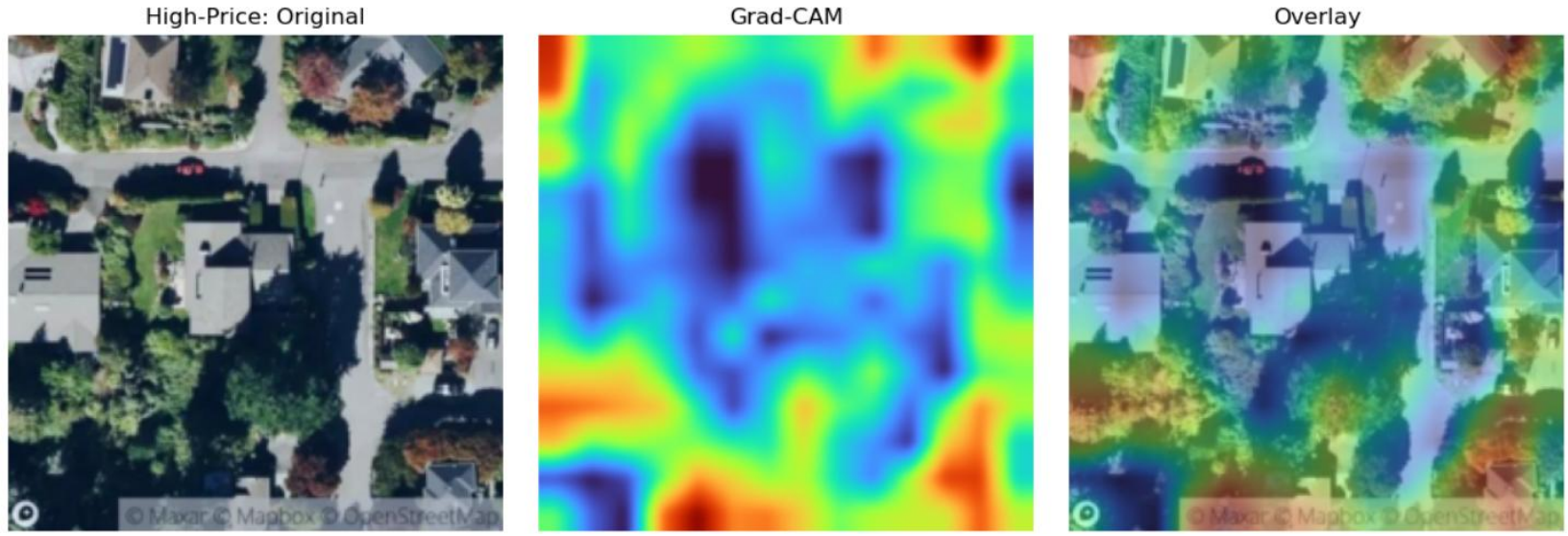
High-Value Properties

For high-priced properties, the model's attention is predominantly focused on:

- Extensive green cover and vegetation
- Low-density residential layouts
- Open spaces and environmental buffers
- Well-organized road networks with limited congestion





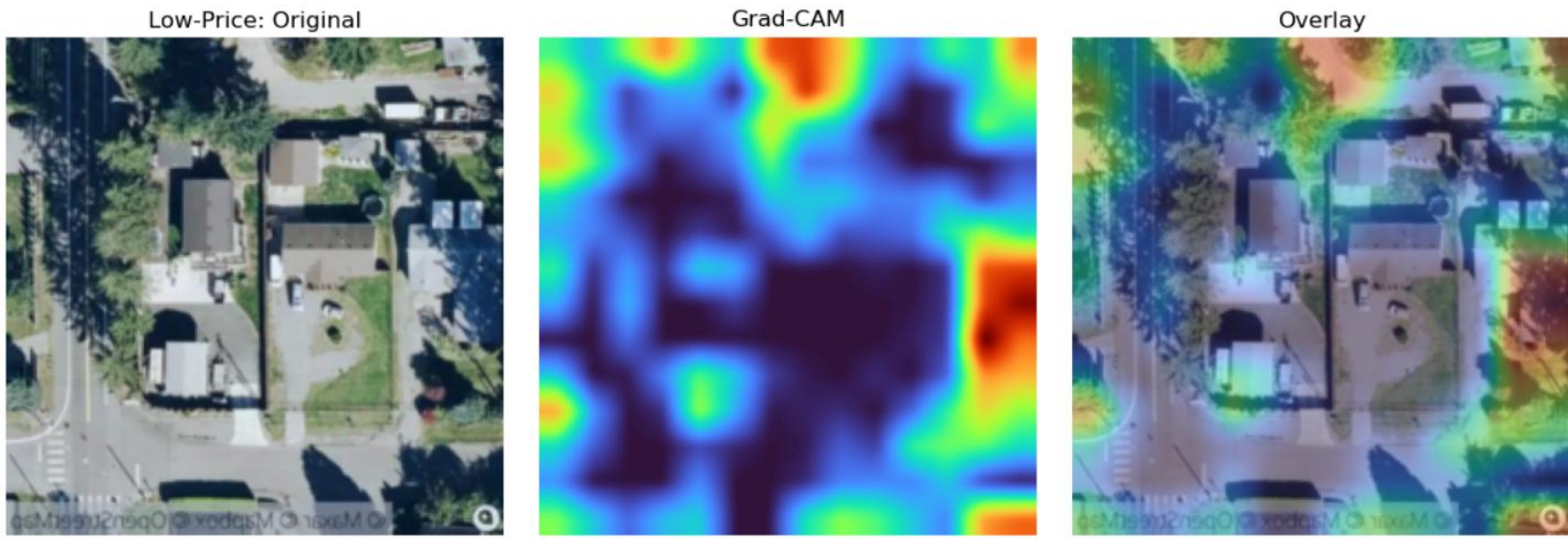


These visual cues are commonly associated with higher desirability and premium valuation.

Low-Value Properties

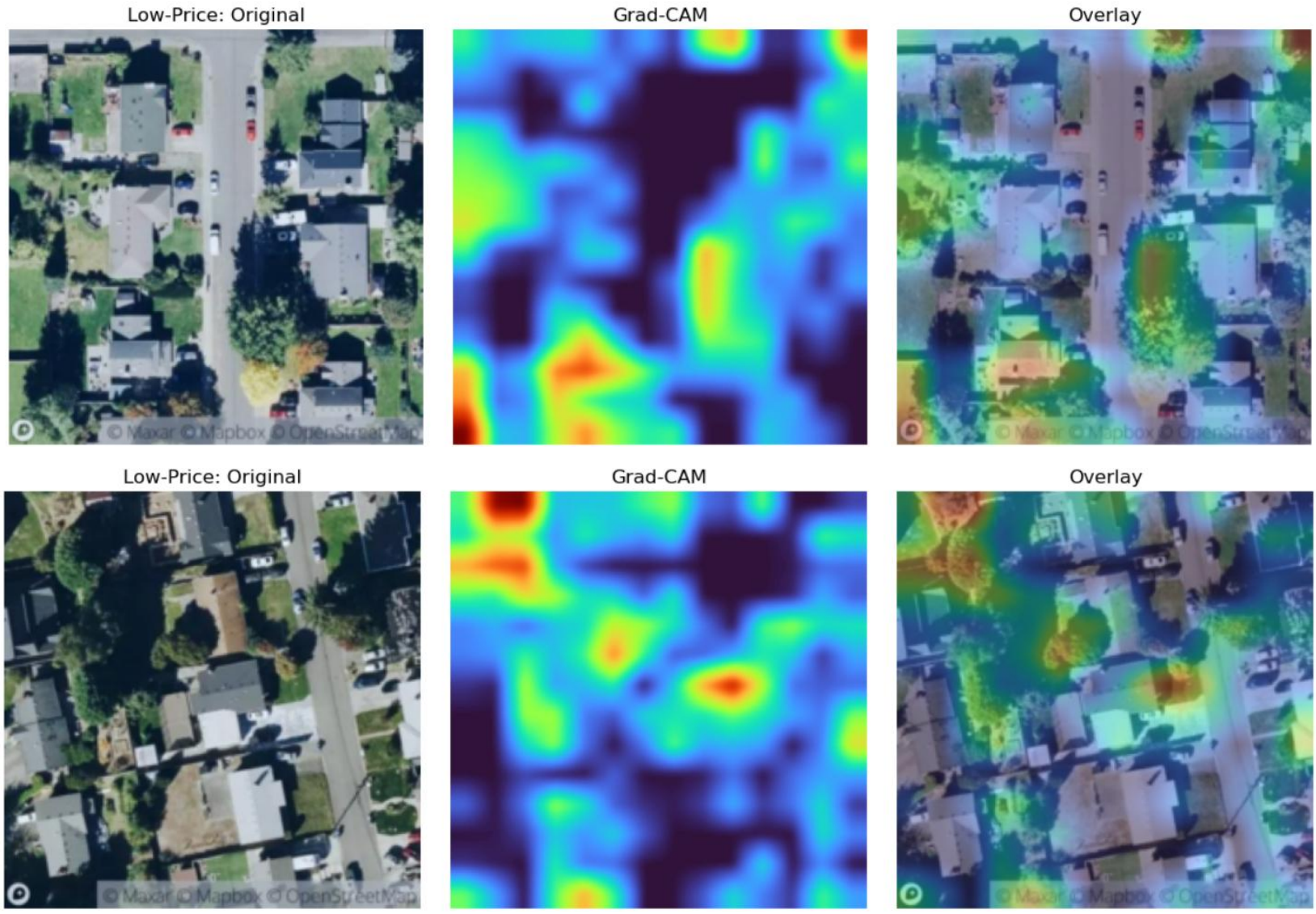
In contrast, low-priced properties exhibit attention patterns centered on:

- Dense road networks
- Highly built-up or compact regions



- Limited or fragmented green spaces

Such features are indicative of lower environmental quality and reduced neighborhood appeal.



---

7. Hybrid CNN + XGBoost Model

7.1 Motivation for Hybrid Approach

While the end-to-end multimodal deep learning model successfully integrated satellite imagery and tabular data, its performance did not surpass the strong tabular baseline in terms of RMSE. This outcome highlights a well-known challenge in applied machine learning: **deep neural networks are not always optimal for tabular-heavy problems**, especially when the structured features already contain strong predictive signal.

End-to-end neural networks must simultaneously learn:

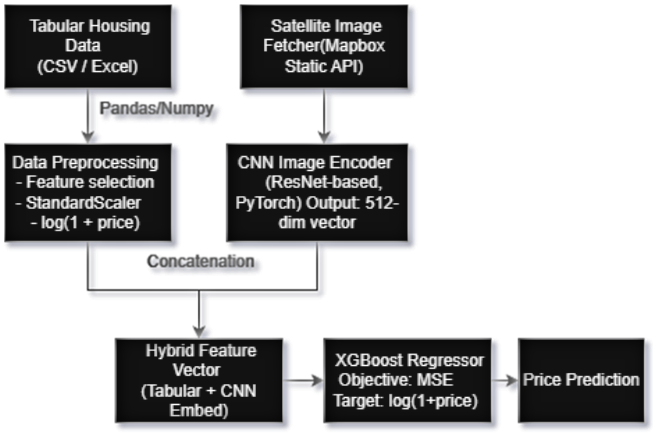
- meaningful visual representations from images, and
- complex nonlinear interactions among tabular features,
  which can lead to suboptimal calibration and slower convergence.

In contrast, **tree-based models**, such as XGBoost, are exceptionally effective at modeling structured data. They naturally capture nonlinear relationships, handle feature interactions, and provide robust calibration without requiring extensive feature scaling or architectural tuning. However, tree-based models cannot directly process raw images.

To leverage the strengths of both paradigms, a **hybrid modeling strategy** was adopted: deep learning is used exclusively for **representation learning** from satellite imagery, while XGBoost is used for **final regression and calibration**.

7.2 Architecture

The hybrid model consists of three key components:



CNN as a Fixed Feature Extractor:

A pretrained convolutional neural network (ResNet-based) is used to extract high-level visual embeddings from satellite images. The classification head is removed, and the network outputs a fixed-length feature vector that encodes neighborhood-level characteristics such as land use patterns, vegetation density, and urban structure.

During hybrid modeling, the CNN parameters are **frozen**, ensuring that the learned visual representations remain stable and that training focuses solely on the regression task.

Concatenation of CNN Embeddings with Tabular Features:

For each property, the CNN-derived image embedding is concatenated with the standardized tabular feature vector. This combined representation preserves both:

- structural and locational information from tabular data, and
- contextual environmental information from satellite imagery.

The resulting feature matrix serves as the input to the final regression model.

XGBoost as the Final Regressor:

XGBoost is employed as the final prediction model. Its ability to efficiently model nonlinear feature interactions and its robustness to mixed feature types make it well-suited for handling the combined embedding–tabular feature space.

The hybrid architecture can be summarized as:

Satellite Image → CNN Embedding → Concatenation with Tabular Features → XGBoost → Price Prediction

7.3 Training and Evaluation:

Feature Extraction Process:

Image embeddings are extracted by passing each satellite image through the frozen CNN encoder. This process is performed once, after which the embeddings are stored and reused for training and evaluation. This decoupling significantly reduces computational cost and stabilizes model behavior.

Train–Validation Split:

The hybrid model uses the same train–validation split as the baseline and multimodal deep learning models to ensure a fair comparison. The target variable remains log(1 + price), maintaining consistency across all experiments.

Evaluation Metrics:

Model performance is evaluated using:

- **Root Mean Squared Error (RMSE)** in log-price space, and
- **R^2 score** to measure variance explained.

The hybrid CNN + XGBoost model achieved:

- **RMSE = 0.17942273321143576**, corresponding to an average relative price error of approximately **19.6%**, and
- **R^2 = 0.883340877514299**, outperforming both the tabular-only baseline and the end-to-end multimodal neural network.

These results demonstrate that visual neighborhood context extracted by CNNs provides complementary information that is most effectively leveraged by a tree-based regressor.

8. Results and Comparison

8.1 Quantitative Results

All models were evaluated on the same validation split using **RMSE in log-price space** and **R^2 score**. For interpretability, RMSE values are also converted to approximate **relative price error**.

Model	RMSE (log)	Approx. Price Error	R^2
Tabular Baseline	~0.277	~32%	~0.721
CNN + MLP (End-to-End)	~0.274	~31%	~0.728
Hybrid CNN + XGBoost	**~0.179**	**~19.6%**	**~0.883**

The hybrid CNN + XGBoost model achieves the **lowest RMSE** and the **highest R^2**, representing a substantial improvement over both the tabular-only baseline and the end-to-end multimodal neural network.

9.2 Submission File

The final deliverable is a CSV file named **final_predictions.csv**, containing price predictions for all properties in the test dataset.

File Description

- Generated using the hybrid CNN + XGBoost model
- Predictions reported in original price units
- No missing or negative values

File Format

id, predicted_price

Where:

- **id** is the unique identifier for each property
- **predicted_price** is the model's estimated market value for the property

This file constitutes the final submission output of the project and can be directly used for evaluation or downstream analysis.