Satellite Imagery Based Property Valuation

A Multimodal Machine Learning Approach

by Chetan Agarwal

1. Introduction

Property valuation is a critical component of real estate markets, urban planning, and financial decision-making. Traditional valuation models rely heavily on tabular property attributes such as area, number of rooms, and location-based indicators. While effective to some extent, these models often fail to capture spatial and neighborhood-level information, which plays a crucial role in determining property value.

Satellite imagery provides a rich source of visual information representing surrounding infrastructure, greenery, road connectivity, and urban density. This project explores whether combining tabular data with satellite imagery can improve property price prediction using multimodal machine learning techniques.

2. Problem Statement

The objective is to:

- Predict property prices using tabular features alone
- Extract meaningful information from satellite images
- Combine both modalities and evaluate performance improvements

3. Dataset Description

3.1 Original Tabular Data

The original tabular dataset consists of structured property-level attributes, including numerical and categorical features describing intrinsic characteristics of properties. These features serve as the foundational input for baseline machine learning models.

The target variable is property price, which is treated as a continuous regression target.

3.2 Feature Engineering on Tabular Data

Beyond the raw features, additional derived features were systematically engineered to enhance the predictive capability of the models. Feature engineering was guided by domain understanding and exploratory data analysis.

The engineered features include:

- Interaction-based features capturing relationships between existing variables
- Aggregated features summarizing multiple related attributes
- Transformed features to reduce skewness and scale imbalance
- Composite indicators designed to better represent property quality and spatial influence

These newly created features significantly enrich the tabular dataset by capturing non-linear relationships that are not directly observable in the raw data.

3.3 Satellite Image Data

Each property in the tabular dataset is associated with a corresponding satellite image. These images capture contextual information such as:

- Surrounding infrastructure
- Urban density
- Road connectivity
- Vegetation and open spaces

Satellite images are processed independently and later combined with the enhanced tabular feature set during multimodal learning.

4. Data Acquisition and Satellite Image Retrieval

4.1 Satellite Imagery Source

Satellite images were obtained through the **Mapbox Static Images API**, which provides high-resolution satellite views for specified geographic coordinates. Each property in the dataset is associated with latitude and longitude values, which were used to retrieve a satellite image centered at the property location.

To maintain uniformity across samples, all images were downloaded using a fixed zoom level, satellite map style, and consistent image resolution. This ensures that each image represents a comparable geographic context and spatial scale.

4.2 Automated Image Retrieval

An automated pipeline was implemented to download satellite images corresponding to individual properties. For each data record, an API request was constructed using the property's geographic coordinates, and the resulting image was stored using a unique property identifier. This guarantees a one-to-one alignment between tabular records and satellite images.

To improve efficiency, previously downloaded images were skipped if a valid file already existed, preventing redundant network requests.

4.3 Parallelized Downloading

Given the size of the dataset, image downloads were parallelized using a thread-based execution approach. Multiple image requests were processed concurrently, significantly reducing total acquisition time while efficiently utilizing network resources. This design allows the pipeline to scale effectively to large geospatial datasets.

4.4 Reliability and Data Integrity

Basic error-handling mechanisms were incorporated to ensure robustness against network failures and unsuccessful API responses. Image downloads were performed separately for training and test datasets to prevent data leakage and maintain consistency across experimental splits.

4.5 Summary

Overall, this automated and scalable data acquisition process provides reliable satellite imagery aligned with tabular property data. By standardizing spatial resolution, geographic alignment, and data splits, the pipeline produces high-quality multimodal inputs suitable for subsequent exploratory analysis and model development.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand the structure, distribution, and relationships within the dataset before model training. The analysis covers tabular property features, engineered features, and satellite image–derived features, ensuring a comprehensive understanding of the multimodal data.

5.1 Target Variable Analysis

The dataset was loaded using Pandas, followed by an initial inspection of data types and missing values.

Key Checks Performed:

- Dataset shape and column types
- Missing value counts for each feature
- Basic statistical summary

Observation:

- The dataset is well-structured with no significant missing values, making it suitable for downstream modeling without extensive imputation.
- Feature types are consistent with their semantic meaning (numerical vs categorical).

5.2 Target Variable Analysis (Property Price)

5.2.1 Distribution of Property Prices

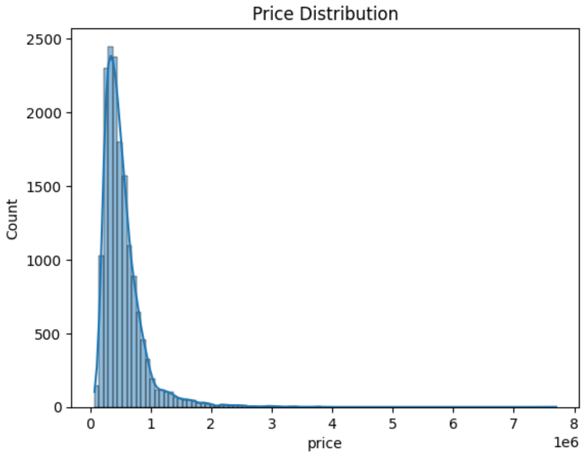
Visualizations Used:

- Histogram of raw property prices
- Histogram of log-transformed prices

Observation:

- Property prices exhibit a strong right-skewed distribution, with a small number of high-value properties. (See figure 1. (a))
- To stabilize variance and improve regression performance, a log transformation **(log1p)** was applied to the target variable. (See figure 1. (b))

Conclusion: Log-transformed prices show a more symmetric distribution, which is better suited for regression models and reduces sensitivity to outliers.



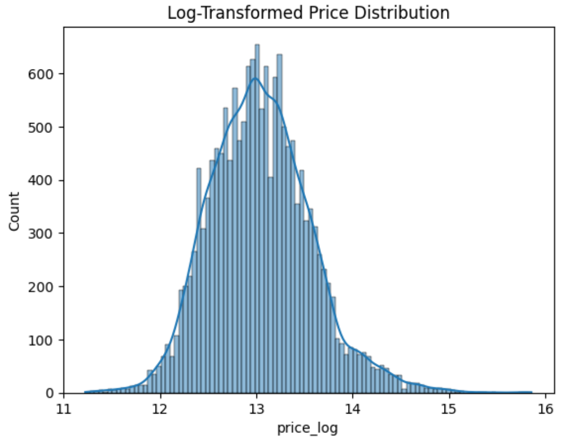


figure 1. (a) figure 1. (b)

5.3 Univariate Feature Analysis

5.3.1 Bedrooms and Bathrooms

Histograms of bedrooms and bathrooms reveal discrete distributions with most properties clustered around typical residential values. While higher counts are associated with increased prices, the relationship is not strictly monotonic. (See figure 2. (a), figure 2. (b))

Insight: Bedrooms and bathrooms alone are weaker predictors of price when compared to size-based features.

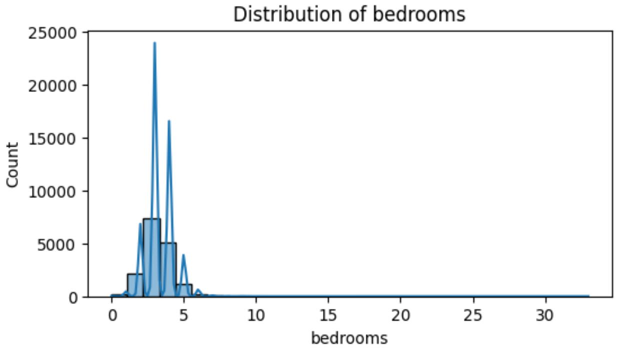


figure 2. (a)

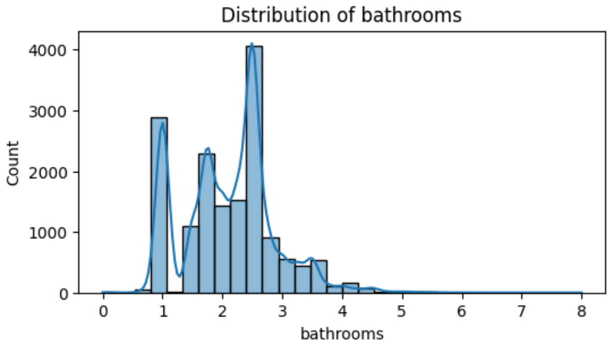


figure 2. (b)

5.3.2 Living Area (sqft_living)

The living area feature shows a continuous and right-skewed distribution. Scatter plots indicate a strong positive relationship between living area and property price.

Key Insight: Living area is the single strongest individual predictor among raw tabular features, justifying its high importance in subsequent modeling.

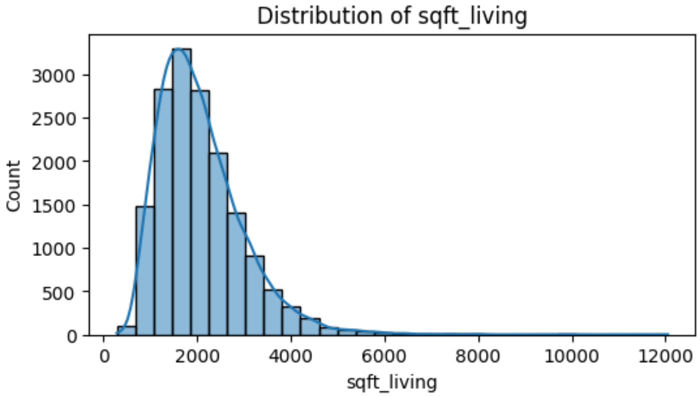


figure 2. (c)

5.4 Bivariate Analysis

5.4.1 Feature Vs Price

Scatter plots reveal strong non-linear relationships between numerical features and property price. Living area shows a clear positive trend and is the most influential predictor (See figure 3. (a)). Bathrooms exhibit moderate correlation, while bedrooms show high variance and limited predictive strength. These patterns indicate that linear models are insufficient, motivating the use of non-linear and ensemble approaches.

5.4.2 Correlation and Spatial Effects

Correlation analysis confirms that living area has the strongest association with price, followed by bathrooms. Latitude and longitude show weak linear correlation individually, but spatial visualizations reveal strong geographic clustering of high-value properties. This indicates that location effects are primarily spatial and non-linear rather than purely numerical (See figure 3. (b)).

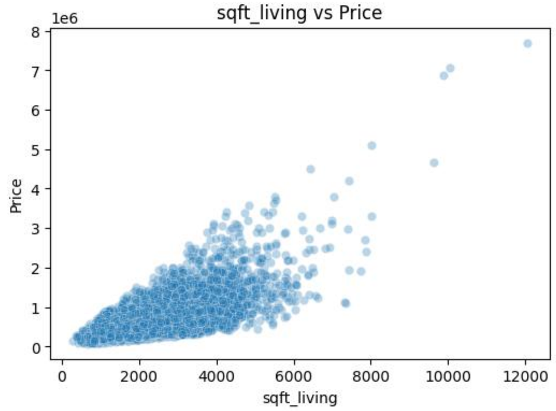


figure 3. (a)

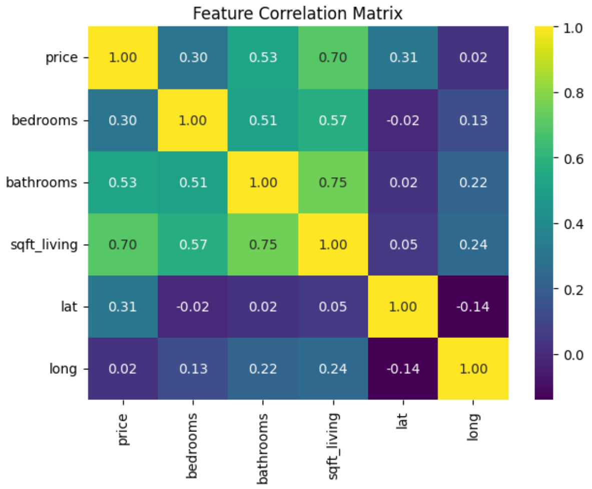


figure 3. (b)

5.5 Geospatial Analysis

5.5.1 Geographical Price Distribution

Scatter plots of longitude vs latitude colored by price reveal strong spatial clustering. High-priced properties are concentrated in specific geographic regions, while lower-priced properties are more uniformly distributed. (See figure 4. (a))

5.5.2 Price-Weighted Location Density

A kernel density estimation weighted by price further confirms the existence of high-value spatial clusters. (See figure 4. (b))

Key Insight: Property prices are heavily influenced by neighborhood-level effects, validating the use of satellite imagery to capture contextual spatial information.

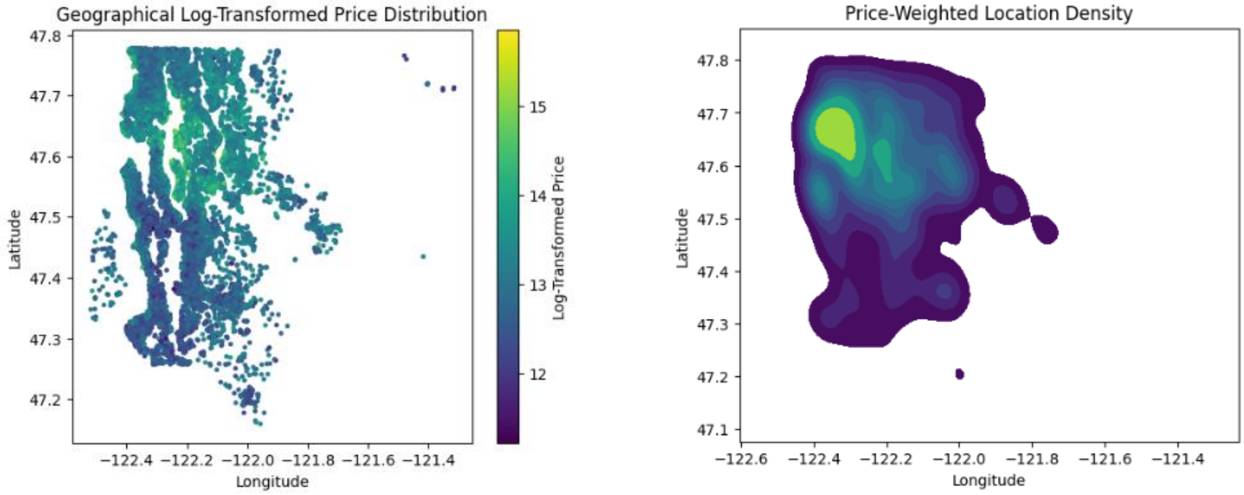


figure 4. (a)

figure 4. (b)

5.6 Satellite Image Analysis

5.6.1 Visual Comparison of Low and High Prices Properties

Direct visual inspection of satellite images corresponding to the lowest and highest priced properties reveals clear structural differences:

- Low-priced properties tend to be located in dense, irregular urban layouts with limited open space.
- High-priced properties are situated in well-planned neighborhoods with wider roads, organized layouts, and abundant greenery.

5.6.2 Brightness and Urban Density

Average image brightness was extracted as a proxy for built-up density. Scatter plots of brightness vs price show that lower-priced properties tend to exhibit higher brightness variability, reflecting dense urban textures and roof-heavy regions. (See figure 5. (a))

In contrast, high-priced properties show more balanced brightness distributions, consistent with open spaces and vegetation.

5.6.3 Green Cover Analysis

A green cover ratio was computed by identifying pixels where the green channel dominates over red and blue channels.

- Properties with higher prices generally exhibit higher green cover ratios.
- Low-priced properties are associated with minimal vegetation.

This trend is clearly visible in scatter plots comparing green cover ratio with property price. (See figure 5. (b))

5.6.4 Spatial Consistency of Green Cover

Mapping green cover ratios across geographic coordinates reveals that high-value spatial clusters coincide with regions of higher vegetation density. (See figure 5. (c))

Conclusion: Environmental quality, as captured through satellite imagery, plays a significant role in property valuation.

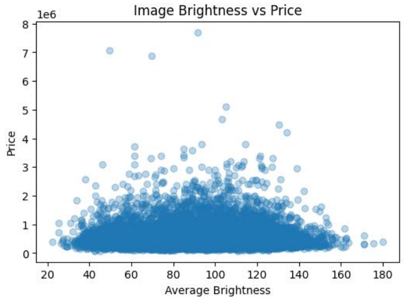


figure 5. (a)

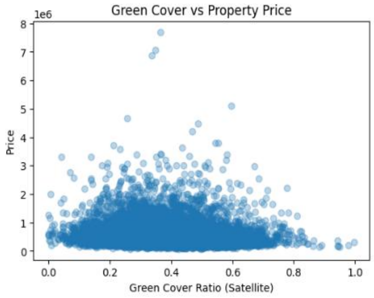


figure 5. (b)

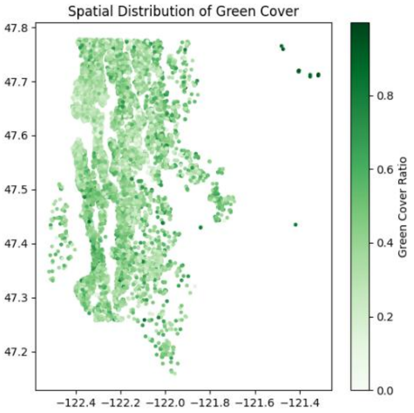


figure 5. (c)

5.6.5 Price Based Image Stratification

Properties were divided into low, mid, and high-priced groups based on price quantiles. Sample satellite images from each group were visually compared.

- Low-priced neighborhoods show compact housing, narrow roads, and minimal greenery.
- Mid-priced neighborhoods exhibit mixed characteristics.
- High-priced neighborhoods are consistently characterized by planned layouts, visible road networks, and substantial green cover.

This stratification provides strong qualitative evidence supporting the quantitative findings.

6. Feature Engineering

Feature engineering was performed to strengthen both tabular and satellite image representations by capturing structural, spatial, and visual patterns relevant to property valuation.

6.1 Tabular Feature Engineering

In addition to raw attributes, several derived features were introduced to capture density and proportional relationships within property layouts. These include area-based ratios (e.g., square footage per bedroom and per room) and bathroom-to-room proportions. Small constants were added to avoid division by zero.

The target variable (property price) was log-transformed to reduce skewness and stabilize model training. The resulting feature set provides a strong tabular baseline encoding intrinsic property characteristics and coarse spatial information.

6.2 CNN-Based Image Feature Extraction

Satellite images were processed using a pretrained **ResNet-18** model as a fixed feature extractor. The final classification layer was removed, yielding a **512-dimensional** embedding for each image. These embeddings capture high-level visual patterns such as vegetation density, road structure, and urban layout without requiring CNN retraining.

6.3 Dimensionality Reduction using PCA

Given the high dimensionality of CNN features, Principal Component Analysis **(PCA)** was applied after standardization. The number of components was chosen to preserve **95% of the variance**, reducing the image feature space from 512 dimensions to a lower-dimensional representation.

This step mitigates redundancy, reduces noise, and improves computational efficiency while retaining the majority of informative visual content.

6.4 Multimodal Feature Fusion

The final feature set was constructed by concatenating:

- Engineered tabular features
- PCA-reduced CNN image features

Features were aligned using unique property identifiers to ensure correct correspondence between tabular and visual data. This unified representation allows models to simultaneously leverage intrinsic property attributes and neighborhood-level visual cues.

6.5 Summary

This feature engineering pipeline transforms raw tabular and visual data into compact, informative representations that support both baseline and multimodal modeling.

7. Modeling and Methodology

This section outlines the modeling strategy used to compare tabular-only and multimodal regression approaches. All models predict the log-transformed property price.

7.1 Tabular Baseline Models

Two tree-based regressors were trained on engineered tabular features:

- Random Forest Regressor
- XGBoost Regressor

These models capture non-linear relationships and feature interactions effectively, providing strong baselines for comparison.

7.2 Multimodal Regression Models

To evaluate the impact of visual information, multimodal regression models were trained using the fused feature set (tabular + satellite image features). The same model families, Random Forest and XGBoost were used to ensure a fair comparison.

This design isolates the effect of adding satellite-derived features while keeping the modeling framework consistent.

7.3 Evaluation Protocol

Models were evaluated using an 80–20 train-validation split with:

- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2 score)

Predictions were generated on the validation set, and performance metrics were compared across tabular-only and multimodal configurations.

7.4 Results Interpretation

The modeling results indicate that multimodal learning does not inherently guarantee performance gains. When tabular data already encodes substantial spatial and structural information, visual features may introduce redundancy and noise. However, qualitative analysis and feature visualizations confirm that CNN embeddings capture meaningful spatial patterns, supporting their relevance for scenarios with weaker or noisier tabular data.

7.5 Summary

The modeling results demonstrate that while satellite imagery captures meaningful neighborhood-level patterns, its quantitative benefit depends on the strength of available tabular data. This motivates further exploration using end-to-end multimodal architecture.

8. Results and Discussion

This section presents the quantitative results of the tabular and multimodal regression models and discusses their implications for property valuation using satellite imagery.

8.1 Tabular Baseline Performance

Tabular-only models achieved strong predictive performance across evaluation metrics. The Random Forest regressor outperformed XGBoost, achieving lower RMSE and higher R^2 on the validation set. This indicates that engineered tabular features, combined with geographic coordinates, capture a substantial portion of the variance in property prices.

Feature importance analysis further confirms that **living area** is the most influential predictor, followed by **location-based features**. This highlights the strength of tabular representations in encoding intrinsic property characteristics and coarse spatial information.

8.2 Multimodal Model Performance

Multimodal models incorporating CNN-derived satellite image features produced competitive results but did not consistently surpass tabular baselines. While both Random Forest and XGBoost benefited from the inclusion of visual features, the improvement in performance was limited.

This suggests that the information captured by satellite imagery partially overlaps with strong tabular predictors, particularly living area and geographic location. As a result, the incremental predictive gain from visual embeddings is constrained in settings where tabular data is already highly informative.

8.3 Effect of CNN Feature Extraction and PCA

The use of a pretrained ResNet-18 enabled efficient extraction of high-level visual features without task-specific training. PCA-based dimensionality reduction helped reduce redundancy and noise, improving model stability and computational efficiency.

Despite these benefits, dimensionality reduction may also discard fine-grained visual details that are potentially relevant for valuation. This trade-off likely contributes to the modest gains observed in multimodal performance.

8.4 Model Explainability using Grad-CAM

To improve interpretability of the CNN-based satellite image features, Gradient-weighted Class Activation Mapping **(Grad-CAM)** was applied to the pretrained ResNet-18 feature extractor. Grad-CAM highlights spatial regions within an image that contribute most strongly to the model's output, enabling visual inspection of learned representations.

Grad-CAM heatmaps were generated by computing gradients of the network output with respect to the final convolutional feature maps and overlaying the resulting activation maps onto the original satellite images. Regions with higher intensity (red) indicate stronger influence, while lower-intensity regions (blue) indicate weaker contribution.

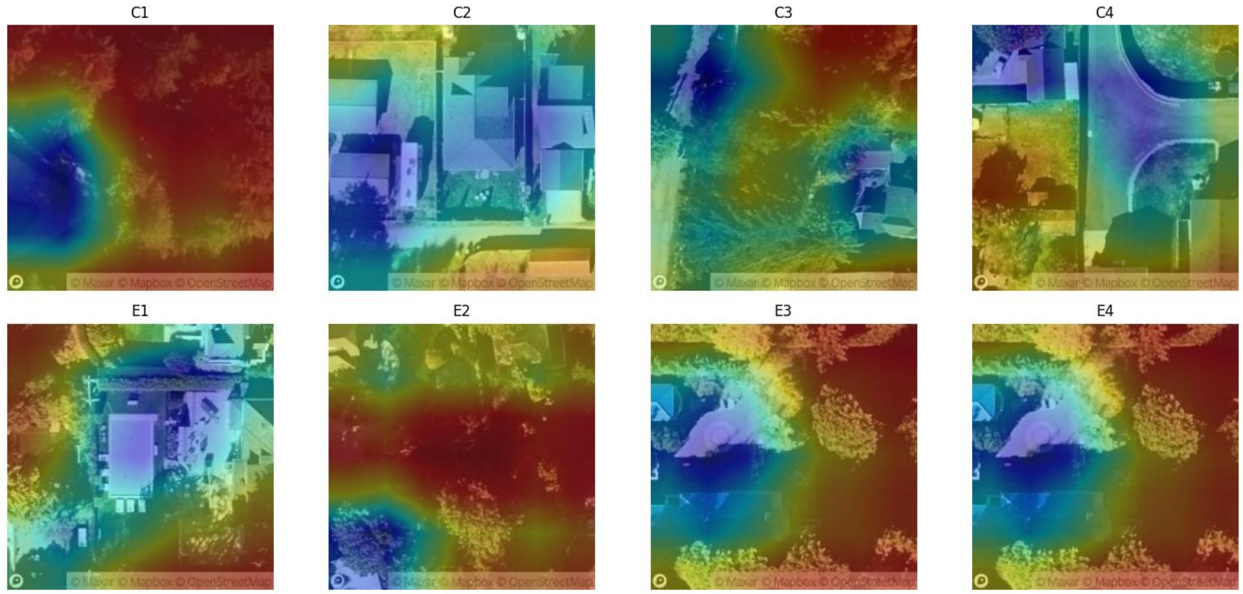


Figure 6. (a) illustrates Grad-CAM heatmaps overlaid on satellite images for representative low-priced (C1–C4) and high-priced (E1–E4) properties.

The visual explanations reveal consistent and meaningful attention patterns. **High-priced properties** show strong activation over green spaces, open areas, wide road networks, and low-density layouts, whereas **low-priced properties** exhibit attention concentrated on dense building clusters, irregular layouts, limited greenery, and compact housing structures. (See figure 6. (a))

These observations align closely with urban economics intuition and confirm that the CNN captures relevant neighborhood-level visual cues. Although the inclusion of image features resulted in limited quantitative performance gains, Grad-CAM analysis demonstrates that the learned visual representations are semantically meaningful and non-arbitrary.

8.5 Interpretation of Findings

The results demonstrate that multimodal learning does not inherently guarantee improved predictive accuracy. When strong tabular features are available, visual features may act as complementary rather than dominant signals. However, qualitative analysis confirms that CNN embeddings capture meaningful neighborhood-level patterns, such as vegetation density and urban structure.

These findings suggest that satellite imagery may be more impactful in scenarios where tabular data is incomplete, noisy, or lacks spatial detail.

8.6 Price Prediction on the Test Dataset

After model training and validation, the final selected regression model was applied to the unseen test dataset to generate property price predictions. Since ground-truth prices are unavailable for the test set, this step evaluates the model's real-world inference capability rather than predictive accuracy.

The test data underwent the same preprocessing and feature engineering steps as the training data to ensure consistency. Engineered tabular features, including area-based and ratio-based attributes, were computed using identical formulations. Satellite image features for test samples were extracted using the pretrained ResNet-18 feature extractor, following the same normalization and embedding pipeline applied during training.

The trained Random Forest model was then used to predict prices in the log-transformed space. Predicted values were subsequently converted back to the original price scale using an exponential transformation. Final predictions were stored in a structured CSV file named "23324003_final.csv" containing property identifiers and their corresponding predicted prices.

This procedure ensures a fully reproducible end-to-end pipeline, demonstrating the deployability of the proposed framework for real-world property valuation scenarios where labels are not available at inference time.

8.7 Discussion and Implications

Overall, the study highlights both the potential and limitations of incorporating satellite imagery for property valuation. While tabular models remain highly effective, satellite imagery provides additional contextual information that can enhance interpretability and robustness. Future improvements may be achieved through end-to-end multimodal architectures or attention-based fusion techniques that better exploit interactions between tabular and visual features.

9. Conclusion and Future Work

9.1 Conclusion

This study investigated property price prediction using a multimodal framework that integrates engineered tabular features with satellite image–derived representations. Strong baseline performance was achieved using tabular data alone, demonstrating that intrinsic property attributes and coarse spatial information capture a significant portion of price variability.

While the inclusion of satellite imagery did not consistently improve quantitative performance over tabular baselines, qualitative analysis using **Grad-CAM** provides important interpretability insights. Grad-CAM visualizations reveal that the CNN feature extractor focuses on semantically meaningful neighborhood-level cues, such as vegetation density, road structure, open spaces, and urban layout. High-priced properties exhibit attention toward greener, low-density environments, whereas low-priced properties emphasize dense and irregular urban patterns.

These findings confirm that the CNN learns non-trivial and interpretable spatial representations, even when performance gains are modest. Overall, the results highlight that multimodal learning enhances model interpretability and contextual understanding, and its effectiveness depends on the strength and completeness of available tabular data.

9.2 Future Work

Several directions can be explored to further improve multimodal property valuation:

- **End-to-end multimodal learning:** Jointly training tabular and image encoders may better capture cross-modal interactions than fixed feature extraction.
- **Advanced fusion strategies:** Attention-based or hierarchical fusion methods could dynamically weight tabular and visual features.
- **Richer visual representations:** Using higher-resolution imagery or fine-tuning CNNs on domain-specific data may improve visual feature relevance.

- **Additional spatial features:** Incorporating road networks, land-use data, or proximity-based geospatial indicators could enhance spatial modeling.
- **Generalization studies:** Evaluating the approach across different cities or regions would provide insight into model robustness and transferability.

These extensions offer promising avenues for leveraging satellite imagery more effectively in large-scale real estate valuation tasks.

10. References

1. **Breiman, L.** (2001), *Random Forests,* Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

2. **Chen, T., & Guestrin, C.** (2016), *XGBoost: A Scalable Tree Boosting System,* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785

3. **He, K., Zhang, X., Ren, S., & Sun, J.** (2016), *Deep Residual Learning for Image Recognition,* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90

4. **Selvaraju, R. R., et al.** (2017), *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,* Proceedings of the IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2017.74

5. **Abadi, M., et al.** (2016), *TensorFlow: A System for Large-Scale Machine Learning,* 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), (Referenced for deep learning ecosystem concepts)

6. **Pedregosa, F., et al.** (2011), *Scikit-learn: Machine Learning in Python,* Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

7. **Jolliffe, I. T.** (2002), *Principal Component Analysis,* Springer Series in Statistics. https://doi.org/10.1007/b98835

8. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016), *Deep Learning,* MIT Press. http://www.deeplearningbook.org

9. **Zhou, B., et al.** (2016), *Learning Deep Features for Discriminative Localization*. CVPR, *(Foundational work related to CAM / Grad-CAM concepts)*

10. **Mapbox** (2024), *Mapbox Static Images API Documentation*. https://docs.mapbox.com/api/maps/static-images/