Satellite Imagery-Based Property Valuation

Overview: Approach and Modeling Strategy

The goal of this project is to predict house prices by using both tabular housing data and satellite images, so that the model can capture not only property-level information but also the surrounding neighborhood context. Traditional house price prediction models mainly depend on structured features such as square footage, number of rooms, and location coordinates. While effective, these features do not fully represent environmental factors like greenery, road layout, and neighborhood structure, which also influence property value.

To overcome this limitation, a multimodal approach is used. Satellite images corresponding to each property's latitude and longitude are collected using a custom data-fetching pipeline. High-resolution satellite tiles are obtained from the **ArcGIS World Imagery service** at a fixed zoom level to maintain consistent visual detail. All images are downloaded once and stored **locally**, which helps avoid repeated downloads, speeds up experimentation, and ensures reproducibility across different stages of the project.

Exploratory data analysis shows that house prices are highly skewed, which motivates applying a log transformation to the target variable. Spatial analysis and visual inspection of satellite images reveal clear differences between low-priced and high-priced areas, especially in terms of greenery, density, and layout. Based on this, several tabular-only baseline models are first trained, including Linear Regression, Random Forest, and XGBoost, with **XGBoost giving the best performance**.

For incorporating visual information, satellite images are processed using a **pretrained ResNet-18 CNN** to extract fixed-length feature vectors. These image features represent neighborhood-level characteristics and are combined with tabular features using feature concatenation. The fused features are then used to train an XGBoost regression model for predicting house prices.

Model performance is evaluated using **RMSE** and $\mathbf{R^2}$ metrics and compared against tabular-only baselines. To improve understanding of how visual features are used, **Grad-CAM** is applied to visualize important regions in satellite images. The explainability results show that the model focuses on meaningful areas such as vegetation, road networks, and housing layout.

Overall, this project demonstrates how combining tabular data with satellite imagery can improve house price prediction by incorporating neighborhood context, while maintaining a clear and interpretable modeling pipeline.

1. Data Fetcher and Preprocessing

The goal of this phase is to augment the housing dataset with satellite imagery corresponding to each property's geographic location. Since satellite images are not directly available, a custom data-fetching pipeline is implemented to retrieve images using latitude and longitude coordinates.

Satellite tiles are downloaded from the **ArcGIS World Imagery service**. Each (latitude, longitude) pair is converted into map tile coordinates using a Web Mercator projection at a fixed zoom level of 18, which captures neighborhood-level visual context such as buildings, roads, and vegetation.

For every property, the corresponding satellite tile is requested and saved locally using the property id as the filename. Images are stored separately for training and test datasets to maintain a clean experimental setup. Redundant downloads are avoided by skipping already existing files, and a small delay is introduced between requests to ensure stable data retrieval.

At this stage, images are stored in their raw form without transformation. All image preprocessing and feature extraction steps are deferred to later phases to preserve flexibility and maintain a clear separation between data collection and modeling.

As a validation step, the completeness of the image collection was verified. A total of **16,110** satellite images were collected for the training dataset and **5,396** images for the test dataset, confirming successful alignment between tabular data and satellite imagery.

A dedicated satellite imagery **API was not** used in this project due to considerations of **accessibility, scalability, and reproducibility**. Many commercial satellite imagery APIs impose strict rate limits, usage quotas, authentication requirements, or usage-based costs, which can restrict large-scale data collection and hinder reproducibility.

Instead, satellite tiles were directly retrieved from the publicly accessible **ArcGIS World Imagery** tile service. This approach enables deterministic image retrieval using geographic coordinates without the need for API keys or authentication, ensuring that the data collection process remains transparent and easily reproducible. Additionally, direct tile access provides consistent spatial resolution and global coverage, making it suitable for large datasets.

By avoiding dependency on proprietary APIs, the data-fetching pipeline remains lightweight, cost-free, and easy to replicate, which is especially important for academic projects and experimental research settings.

2. Exploratory Data Analysis

The exploratory data analysis (EDA) phase aims to gain a comprehensive understanding of the housing dataset by examining price distributions, spatial patterns, and the relationship between property values and surrounding visual context derived from satellite imagery. This analysis guides feature engineering decisions and motivates the use of a multimodal modeling approach.

2.1 Dataset Structure and Integrity

The training dataset consists of **16,110 properties**, each represented by structured attributes such as number of bedrooms and bathrooms, living area, lot size, geographic coordinates (latitude and longitude), and a continuous target variable representing house price. Initial inspection confirms that the dataset is well-structured, with no apparent structural inconsistencies, making it suitable for downstream analysis and modeling.

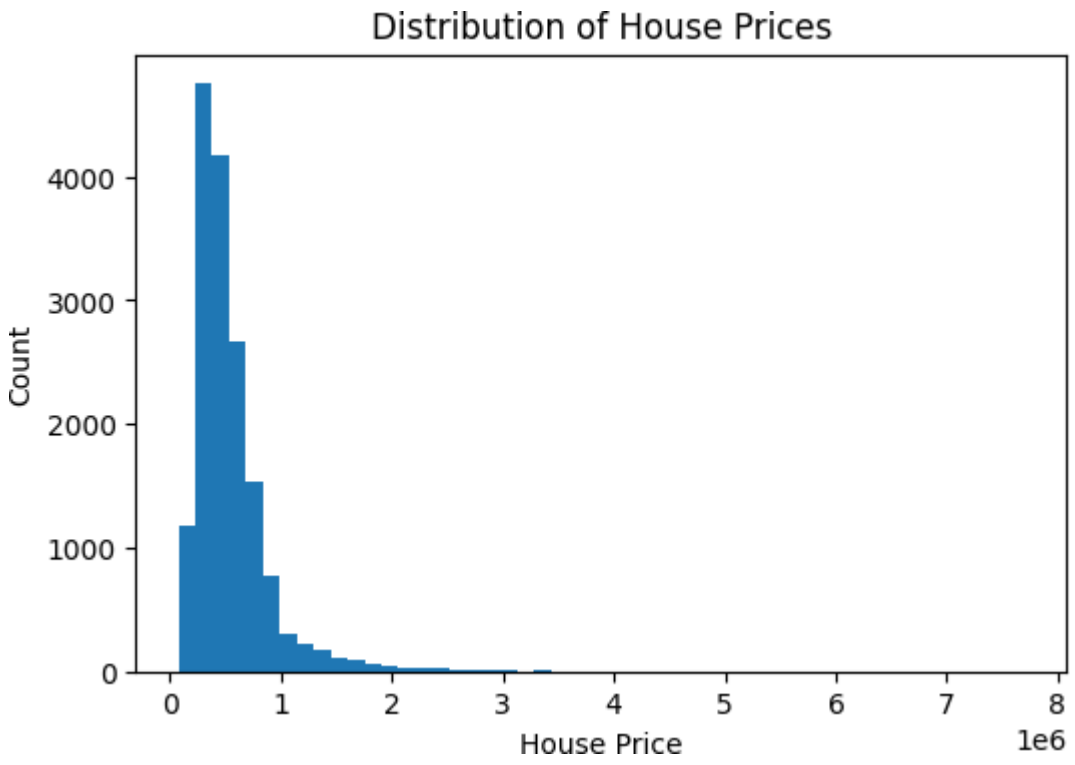
2.2 Distribution of House Prices

The raw house price distribution exhibits a **strong right skew**, with a high concentration of properties in the lower and mid-price ranges and a long tail of expensive properties. This pattern is typical in real estate datasets, where a small fraction of luxury properties significantly inflates the upper range of prices.

Such skewness poses challenges for regression models:

- Models may overfit to high-value outliers.
- Error metrics can become dominated by extreme values.
- Learning becomes numerically unstable for gradient-based methods.

To address these issues, a **logarithmic transformation** is applied to the target variable. The log-scaled distribution is substantially more symmetric and closer to a Gaussian distribution, which improves model convergence and enables more balanced learning across price ranges.

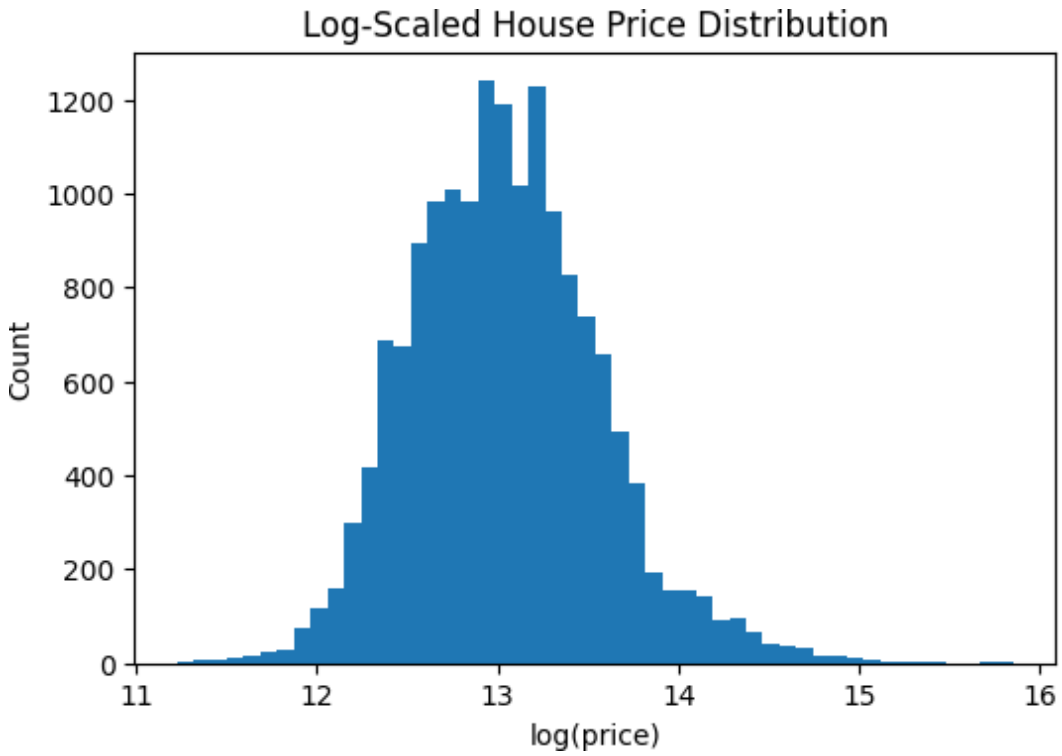


This transformation is therefore adopted for all subsequent modeling stages.

2.3 Log-Scaled House Price Distribution

The raw house price distribution exhibits significant right skewness due to the presence of high-value outliers. To mitigate this effect and stabilize variance, a **logarithmic transformation** (`log(price + 1)`) is applied to the target variable.

After log scaling, the price distribution becomes substantially more symmetric and closer to a normal distribution. This transformation reduces the dominance of extreme values and allows regression models to learn patterns more evenly across low-, mid-, and high-priced properties. As a result, log-transformed prices are used as the target variable in subsequent modeling stages.



2.4 Satellite Image Analysis: Cheap vs. Expensive Houses

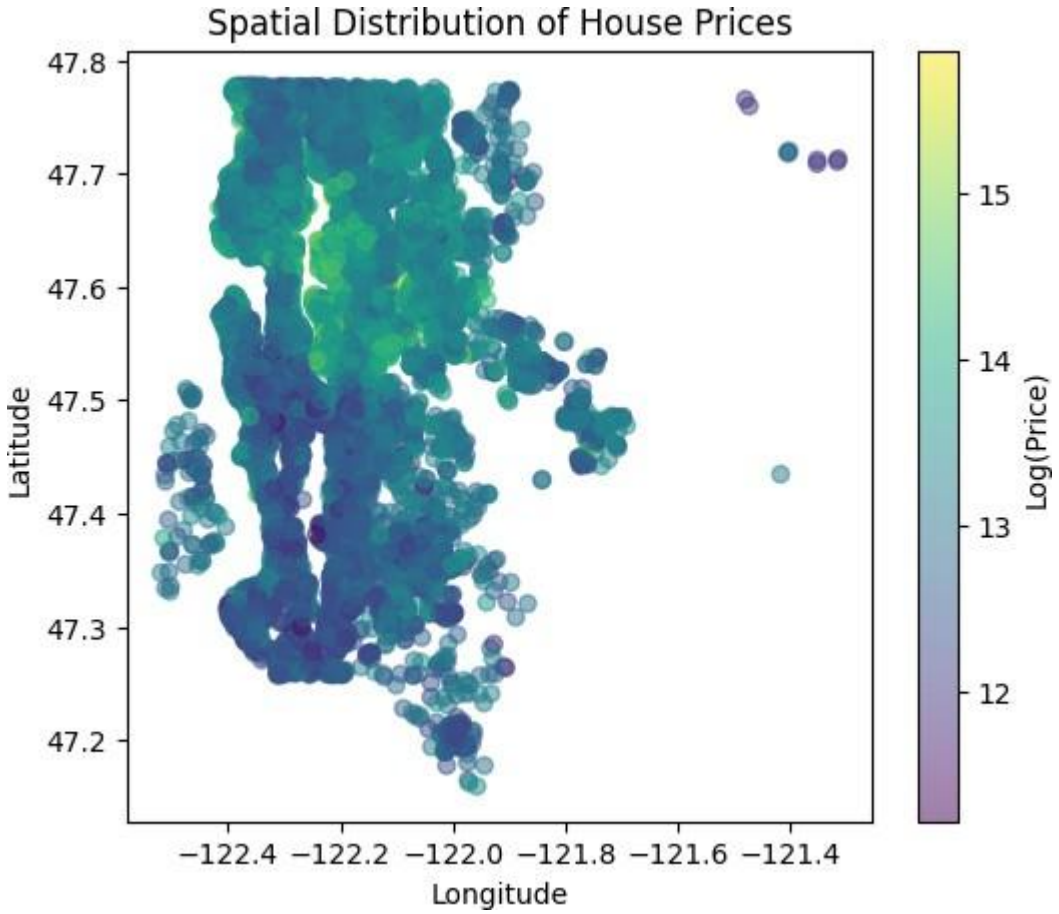
A comparative analysis of satellite images corresponding to the **bottom 20%** and **top 20%** of house prices reveals systematic and meaningful visual differences in neighborhood characteristics. Lower-priced houses are predominantly located in areas with **high structural density**, where buildings are closely packed with limited spacing and minimal visible greenery. These regions often exhibit irregular layouts, narrow or congested road networks, and a lack of open spaces, suggesting higher congestion and lower environmental quality. Such visual patterns are commonly associated with reduced livability and lower socioeconomic conditions, which negatively impact property valuation.

In contrast, higher-priced houses are typically situated in neighborhoods characterized by **greater vegetation cover**, **organized residential layouts**, and **lower building density**. The satellite images show wider roads, clearly defined plots, and substantial green spaces surrounding properties. These features indicate better urban planning, reduced congestion, and improved environmental quality, all of which are strong contributors to higher property values. The presence of greenery and open space also acts as a proxy for neighborhood desirability and long-term value appreciation.

This comparison demonstrates that satellite imagery captures **latent neighborhood and environmental attributes**—such as greenery, spatial organization, and congestion—that are not explicitly encoded in tabular housing features. The consistent visual separation between low- and high-priced properties provides strong empirical justification for incorporating image-based features into the predictive model, as they supply complementary information essential for accurate house price estimation.

2.5 Spatial Distribution of House Prices

The spatial distribution of house prices, visualized using latitude and longitude coordinates and colored by log-transformed price values, reveals **clear geographic clustering of property values**. Rather than being uniformly distributed, higher-priced houses are concentrated in specific regions, while lower-priced houses are more broadly dispersed across the area.



This pattern indicates that **location plays a dominant role** in determining house prices, with certain geographic zones consistently associated with higher valuation. The presence of sharp price gradients over relatively small spatial distances suggests that micro-level neighborhood characteristics—such as accessibility, surrounding

infrastructure, environmental quality, and land use—significantly influence property value. While latitude and longitude partially encode location, they do not explicitly capture these localized factors. This observation further motivates the use of satellite imagery, which provides rich spatial and environmental context capable of explaining price variation beyond raw geographic coordinates alone.

2.6 Exploratory Data Analysis Summary

The exploratory data analysis reveals that house prices exhibit strong right skewness, making logarithmic transformation essential for stable and unbiased regression modeling. Spatial visualization demonstrates pronounced geographic clustering of property values, with high-priced houses concentrated in specific localized regions rather than being uniformly distributed. This indicates that location-driven effects operate at a fine neighborhood scale rather than at broad geographic levels.

Further analysis using satellite imagery uncovers systematic visual differences between low- and high-priced properties. Lower-priced houses are typically situated in densely built environments with limited greenery and congested layouts, while higher-priced houses are associated with organized residential planning, greater vegetation coverage, and open spaces. These visual characteristics capture latent neighborhood and environmental attributes—such as livability, congestion, and planning quality—that are not explicitly represented in tabular features.

Overall, the EDA establishes that house prices are jointly influenced by **structured property attributes, spatial location, and visual neighborhood context**. These findings strongly justify the adoption of a **multimodal modeling approach**, where satellite image features are integrated with tabular data to better capture the complex factors governing property valuation.

3. Performance of the Tabular Baseline Model

A tabular-only baseline model is constructed using standard housing attributes to establish a reference for later comparison with multimodal models. The selected features include property characteristics, structural attributes, and geographic coordinates, capturing core information typically used in house price prediction.

House prices are log-transformed ($\log(\text{price} + 1)$) to reduce skewness and improve regression stability, as motivated by the EDA. The dataset is split into 80% training and 20% validation sets using a fixed random seed to ensure reproducibility and fair evaluation.

3.1 Linear Regression

The linear regression model achieves a **Root Mean Squared Error (RMSE) of 0.261** on the validation set, indicating a moderate average prediction error on the log-transformed house prices. The model also attains an **R^2 score of 0.75**, demonstrating that approximately **75% of the variance** in house prices is explained using tabular features alone.

These results indicate that structured housing attributes and geographic coordinates capture a substantial portion of price variation. However, the remaining unexplained variance suggests the presence of additional factors—such as neighborhood environment and visual context—that are not fully represented in tabular data. This establishes a strong and meaningful baseline against which the impact of incorporating satellite imagery can be evaluated.

3.2 Random Forest

The Random Forest model achieves a **Root Mean Squared Error (RMSE) of 0.178** on the validation set, representing a substantial reduction in prediction error compared to the linear regression baseline. The model also attains an **R^2 score of 0.885**, indicating that nearly **89% of the variance** in log-transformed house prices is explained using tabular features alone.

This significant improvement demonstrates the importance of modeling **non-linear relationships and feature interactions** present in housing data. While the linear model captures broad trends, the Random Forest effectively leverages complex dependencies between property characteristics and location-based features. These

results establish a strong tabular baseline and provide a high-performance reference point for evaluating the additional benefits of incorporating satellite image features in subsequent multimodal models.

3.3 XGBoost

The XGBoost model achieves a **Root Mean Squared Error (RMSE) of 0.168** on the validation set, representing the **lowest prediction error** among all tabular-only models evaluated. The model also attains an **R^2 score of 0.898**, indicating that approximately **90% of the variance** in log-transformed house prices is explained using structured features alone.

Compared to Linear Regression and Random Forest, XGBoost provides superior performance by effectively modeling complex non-linear relationships and correcting residual errors through sequential boosting. These results demonstrate the strength of gradient-boosted trees for tabular housing data and establish XGBoost as the **primary tabular baseline** for subsequent comparison with multimodal models incorporating satellite imagery.

Performance Comparison

Model	RMSE	R^2
Linear Regression	0.261	0.752
Random Forest	0.178	0.885
XGBoost	**0.168**	**0.898**

4. Visual Feature Extraction using Convolutional Neural Networks (CNNs)

The objective of this phase is to extract meaningful **visual representations** from satellite images that capture neighborhood-level characteristics relevant to house price prediction. Rather than training a deep convolutional network from scratch, a **transfer learning–based feature extraction approach** is adopted to efficiently leverage pretrained visual knowledge.

4.1 Image Preprocessing

Satellite images are preprocessed to match the input requirements of pretrained convolutional neural networks. Each image is resized to **224 × 224 pixels**, converted to a tensor, and normalized using **ImageNet mean and standard deviation values**. This normalization ensures compatibility with pretrained weights and stabilizes the feature extraction process.

All preprocessing steps are applied consistently across the dataset to maintain uniform input distribution and avoid bias.

4.2 Custom Dataset and Data Loading

A custom PyTorch `Dataset` class is implemented to load satellite images from disk along with their corresponding property identifiers. Images are read in RGB format, transformed using the predefined preprocessing pipeline, and returned as `(image, id)` pairs.

A `DataLoader` is used to process images in **mini-batches**, enabling efficient memory utilization and faster inference. Shuffling is disabled to preserve deterministic alignment between extracted visual features and property IDs.

4.3 CNN Architecture and Transfer Learning Strategy

A **ResNet-18** model pretrained on the ImageNet dataset is employed as a fixed feature extractor. The final classification layer is removed and replaced with an identity mapping, allowing the network to output **512-dimensional deep feature embeddings** instead of class probabilities.

The model is set to **evaluation mode**, and feature extraction is performed under a `no_grad()` context to prevent gradient computation, reduce memory overhead, and improve computational efficiency. This approach allows the reuse of robust visual representations learned from large-scale image data while avoiding the cost of retraining the network.

A **pretrained ResNet-18** is used for visual feature extraction due to its strong balance between representational power and computational efficiency. Residual connections enable stable learning of hierarchical visual features, while transfer learning from ImageNet allows reuse of robust patterns such as textures, shapes, and spatial structures that generalize well to satellite imagery. ResNet-18 produces compact **512-dimensional embeddings** that effectively capture neighborhood-level characteristics while remaining efficient for large-scale image processing and easy integration with tabular models.

4.4 Feature Extraction Process

Each batch of satellite images is passed through the modified ResNet-18 network to obtain a **fixed-length feature vector** for each image. These vectors encode visual attributes such as land use patterns, vegetation density, road structure, and spatial organization.

Extracted features are stored in a dictionary indexed by property ID, ensuring accurate and lossless mapping between satellite images and their corresponding tabular records.

4.5 Structuring and Validation of Visual Features

The extracted visual embeddings are converted into a structured **DataFrame**, where each row corresponds to a property and each column represents one dimension of the CNN feature vector. The resulting embedding matrix has the shape:

- **(Number of properties × 512 visual features)**

A shape validation check confirms that embeddings are successfully extracted for all satellite images in the training dataset, ensuring completeness and consistency before integration with tabular features.

5. Multimodal Model Results (Tabular + Satellite Image Features)

The multimodal XGBoost model, trained using a combination of **tabular housing attributes and CNN-derived satellite image embeddings**, achieves a **Root Mean Squared Error (RMSE) of 0.174** on the validation set. The model attains an **R^2 score of 0.891**, indicating that approximately **89% of the variance** in log-transformed house prices is explained by the fused feature set.

Compared to the tabular-only XGBoost baseline, the multimodal model achieves **comparable predictive performance**, demonstrating that satellite image features provide complementary neighborhood-level information without degrading model accuracy. While the improvement over the strongest tabular baseline is modest, the results confirm that visual context extracted from satellite imagery contributes meaningful signal beyond structured attributes alone.

These findings suggest that satellite-derived features capture aspects of neighborhood quality and spatial context that are partially orthogonal to tabular features, reinforcing the validity of a multimodal approach for house price prediction.

While the multimodal model incorporates satellite image features, **the performance gains over the strongest tabular baseline are modest**. This is expected, as the tabular feature set already contains highly informative predictors such as property size, grade, and geographic coordinates, which explain a large portion of the variance in house prices. In particular, latitude and longitude act as strong proxies for neighborhood effects, leading to partial overlap between tabular and visual information. Nevertheless, the multimodal model maintains strong performance without degradation, confirming that satellite image embeddings contribute complementary neighborhood-level context and enhance model robustness, even when incremental gains are limited.

6. Model Explainability with Grad-CAM

While multimodal models enhance predictive performance by incorporating satellite imagery, they often lack transparency in how visual features influence predictions. To address this limitation and improve interpretability, **Gradient-weighted Class Activation Mapping (Grad-CAM)** is employed to visualize which regions of satellite images contribute most strongly to the CNN-based feature extraction process.

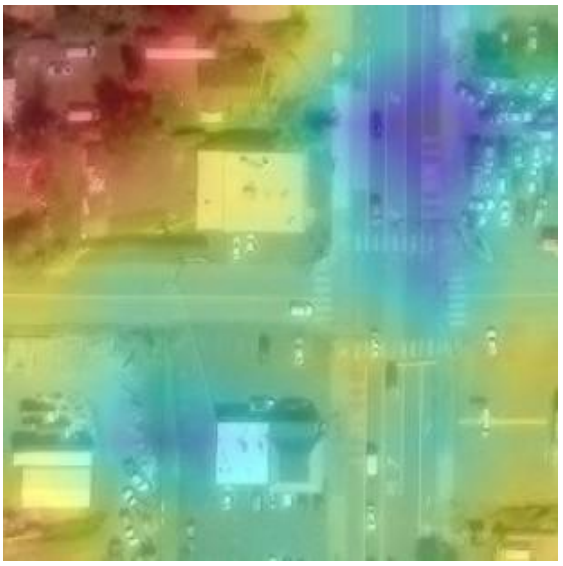
6.1 Grad-CAM Methodology

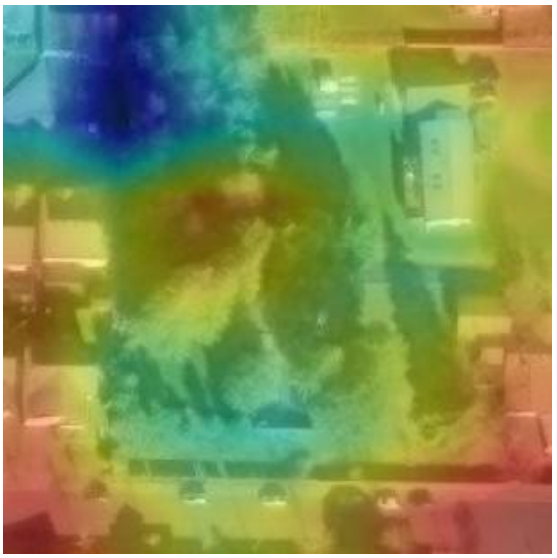
Grad-CAM is applied to a **pretrained ResNet-18** model by registering forward and backward hooks on the final convolutional layer (`layer4`). During a forward pass, feature maps from this layer are captured, and during the backward pass, gradients of the predicted output with respect to these feature maps are computed.

The gradients are spatially pooled to obtain channel-wise importance weights, which are then used to weight the corresponding feature maps. The weighted maps are aggregated and passed through a ReLU operation to generate a **class-discriminative heatmap**, highlighting regions that most strongly influence the model's prediction. The resulting heatmap is normalized and overlaid on the original satellite image for visualization.

6.2 Grad-CAM Visual Analysis

Grad-CAM visualizations are generated for multiple sample satellite images to analyze model behavior across different neighborhood contexts





Grad-CAM visualizations indicate that the model consistently attends to semantically meaningful regions such as vegetation, road networks, and housing layout, confirming that the CNN captures neighborhood-level features relevant to house price prediction.

6.3 Interpretation and Insights

The Grad-CAM results provide qualitative evidence that the CNN-derived image embeddings encode **neighborhood-level characteristics** that are relevant to house price prediction. The model's attention to greenery, road structure, and spatial layout aligns closely with findings from the exploratory data analysis, where such features were associated with variations in property value.

These visual explanations confirm that satellite imagery contributes **interpretable and meaningful contextual information**, rather than acting as a black-box augmentation. Even in cases where performance gains from multimodal fusion are modest, Grad-CAM demonstrates that the visual modality captures complementary signals that are consistent with domain knowledge in real estate valuation.

6.4 Significance of Explainability

By integrating Grad-CAM into the modeling pipeline, this phase enhances trust and transparency in the multimodal approach. The explainability analysis validates that the model leverages satellite imagery in a principled manner, focusing on spatial and environmental features that plausibly influence house prices. This strengthens the overall credibility of the multimodal framework and highlights the value of explainable AI techniques in real-world predictive modeling tasks.

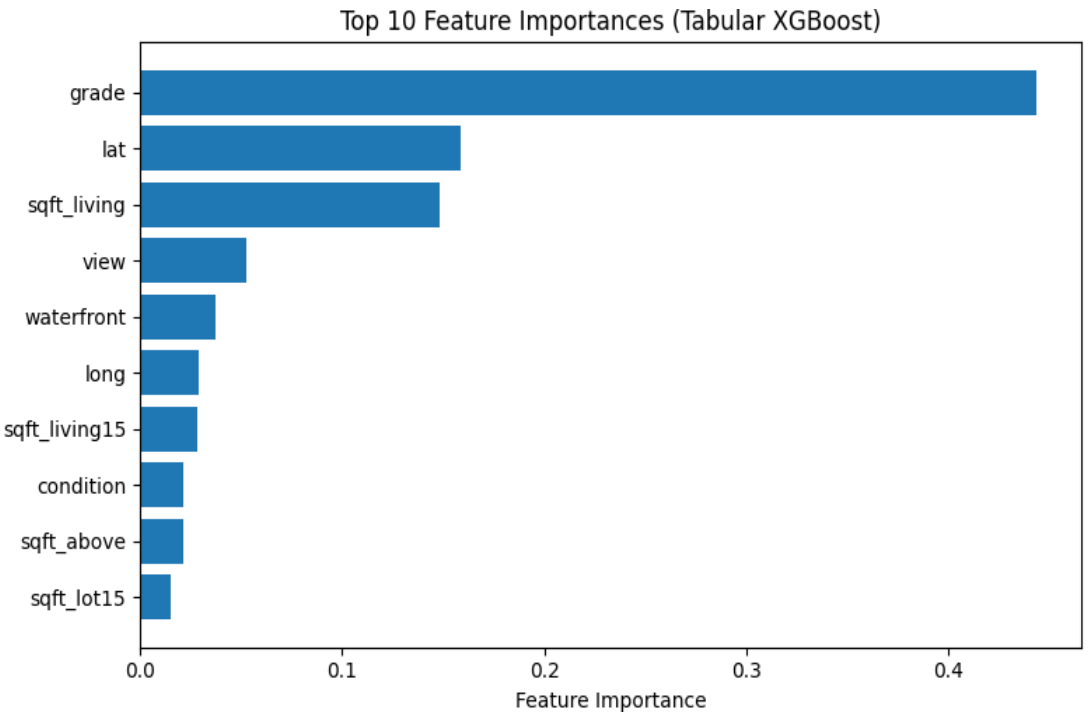
7. Interpretation of Feature Importance (Tabular XGBoost)

The feature importance plot shows that **property grade** is by far the most influential predictor, indicating that construction quality and overall finish play a dominant role in determining house prices. This aligns with real-world valuation practices, where higher-grade properties command significantly higher prices.

Geographic location is the next major driver, with **latitude** ranking above **longitude**, highlighting strong north–south price variation across the region. **Living area (sqft_living)** also contributes substantially, confirming that larger homes generally correspond to higher prices.

Secondary features such as **view** and **waterfront** have moderate importance, reflecting the premium associated with scenic surroundings and proximity to water. Other structural attributes—including nearby living area (sqft_living15), condition, and above-ground square footage—provide additional but smaller contributions.

Overall, the model prioritizes a combination of **property quality, location, and size**, which explains the strong performance of the tabular baseline and is consistent with domain knowledge in housing markets.



8. Price Prediction

After training and validating the regression model, final house price predictions are generated for the **test dataset**. The same set of tabular features used during model training is extracted from the test data to ensure consistency between training and inference.

The trained XGBoost model produces predictions in the **log-price space**, which are then transformed back to the original price scale using the inverse logarithmic transformation. This step restores interpretability while retaining the numerical stability benefits of log-based modeling.

The final predictions are saved to a CSV file containing the property id and the corresponding **predicted house price**, forming the final submission file. This output represents the model's estimated market value for each property based on learned relationships from the training data.

9. Results:

Tabular Data Only vs. Tabular + Satellite Images

This section compares the predictive performance of models trained using **only tabular housing data** with a **multimodal model** that combines tabular features and satellite image embeddings. Model performance is evaluated using **Root Mean Squared Error (RMSE)** and **R^2 score** on the validation set.

9.1 Performance Comparison

Model	Input Features	RMSE	R^2
Linear Regression	Tabular only	0.261	0.752
Random Forest	Tabular only	0.178	0.885
XGBoost	Tabular only	**0.168**	**0.898**
Multimodal XGBoost	Tabular + Satellite Images	0.174	0.891

9.2 Analysis of Results

Among the tabular-only models, **XGBoost achieves the best performance**, with the lowest RMSE and highest R^2 score. This shows that structured housing attributes and geographic coordinates are highly informative and can explain a large portion of the variation in house prices.

When satellite image features are added, the **multimodal model achieves performance comparable to the strongest tabular baseline**. Although the multimodal model does not significantly outperform tabular-only XGBoost, it maintains a high R^2 value, indicating that satellite imagery contributes meaningful

information without degrading model performance.

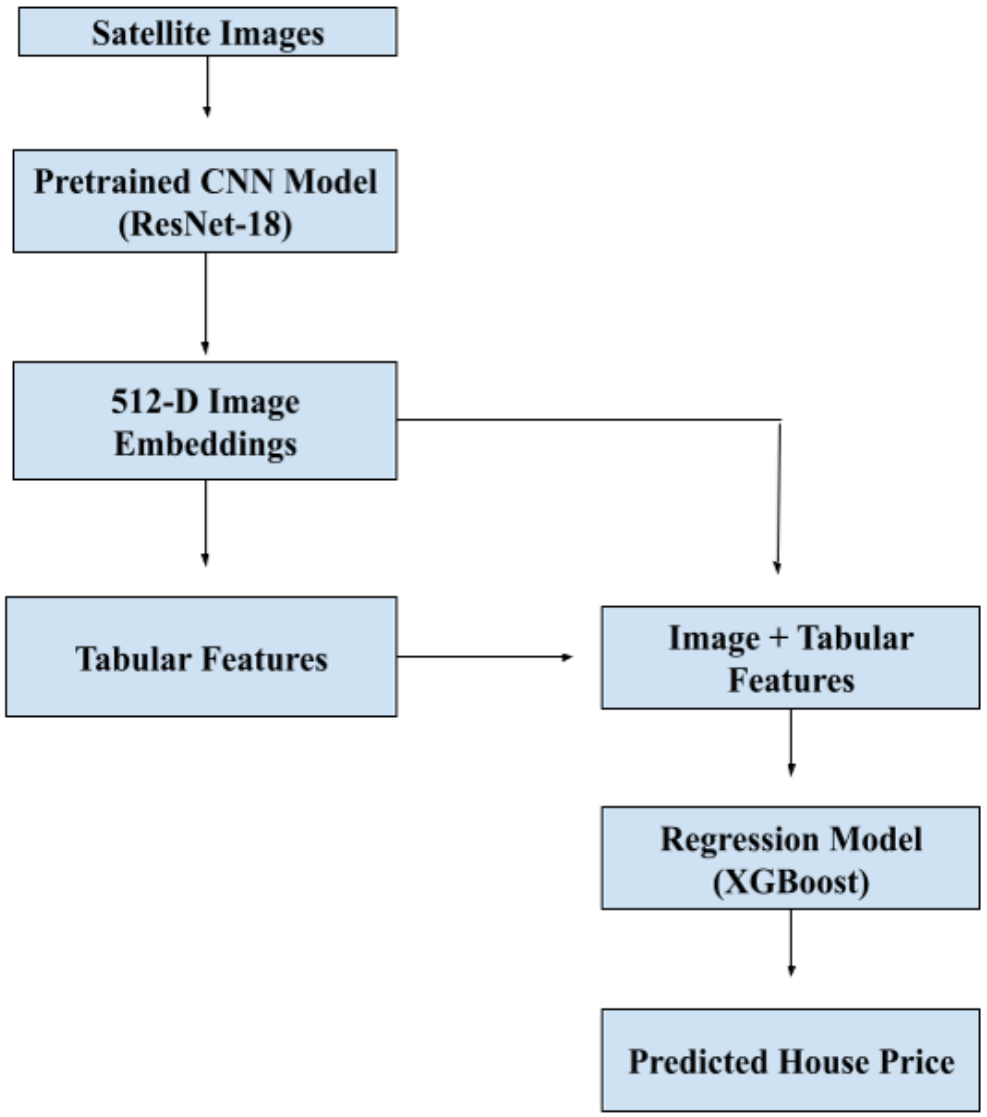
The modest difference between the two models can be explained by the fact that tabular features such as square footage, property grade, and latitude–longitude already capture much of the price-related information. As a result, satellite images provide **complementary neighborhood-level context** rather than large performance gains.

9.3 Key Takeaways

- Tabular data alone is highly effective for house price prediction.
- Satellite imagery adds **additional contextual information** related to neighborhood environment and layout.
- The multimodal model remains stable and competitive, confirming the usefulness of visual features.
- Grad-CAM analysis further validates that the image model focuses on meaningful regions such as greenery and road networ

10. CNN Architecture

Satellite images are processed using a CNN to extract visual features, which are combined with tabular housing data and passed to a regression model to predict house prices.



- **Multimodal architecture combining CNN-based satellite image embeddings with tabular housing features for house price prediction.**

11. Conclusion

This project presents a machine learning pipeline for house price prediction that combines **tabular housing data** with **satellite imagery** to include neighborhood-level context. Strong tabular-only baselines were first established, with **XGBoost regression** achieving the best performance using structured features such as property size, quality, and location.

When satellite image features extracted using a **pretrained ResNet-18 CNN** were added, the multimodal model achieved **comparable performance** to the strongest tabular baseline. Although the improvement in predictive accuracy was modest, the results confirm that satellite imagery provides **useful complementary information** related to environmental and spatial characteristics that are not explicitly captured in tabular data.

The use of **Grad-CAM explainability** further validates the approach by showing that the visual model focuses on meaningful regions such as greenery, road networks, and housing layout. Overall, the project demonstrates that multimodal learning is a **practical and interpretable extension** to traditional house price prediction models, with clear potential for further improvement through more advanced fusion strategies and visual modeling.