

# Residual Fusion of Tabular Data and Satellite Imagery for Property Price Estimation

Yash Jain (24114108)

## 1 Overview: Approach and Modeling Strategy

Property valuation is influenced by more than just the internal attributes of a house, such as its size, number of rooms, or construction quality. External factors related to the surrounding neighborhood—including road connectivity, nearby buildings, green spaces, and overall urban planning—also play a crucial role in determining market value. While tabular datasets effectively capture internal property characteristics, they fail to fully represent neighborhood-level context.

In this project, we aim to enhance property price prediction by integrating tabular housing data with satellite imagery. Satellite images provide a visual snapshot of the area surrounding each property and implicitly encode information such as urban density, connectivity, and land use, which are difficult to express using numerical features alone.

My approach follows a two-stage modeling strategy. First, an XGBoost regression model is trained using only tabular features. This model serves as a strong baseline and captures most of the predictable patterns related to property size, location, and amenities. Geographic coordinates (latitude and longitude) are included to encode coarse spatial trends.

Instead of directly merging tabular and image features, a residual learning framework is adopted. A Convolutional Neural Network (CNN) is trained on satellite images to predict a residual correction. This residual represents the adjustment required to refine the baseline XGBoost prediction based on visual neighborhood context.

The final prediction is computed as:

$$\text{Final Price} = \text{XGBoost Prediction} + \text{CNN Residual}$$

This design is chosen because naive fusion strategies often underperform. In such cases, strong tabular signals dominate learning and prevent the image model from contributing effectively. By using residual fusion, each component has a clearly defined role: the tabular model provides a reliable base estimate, while the CNN focuses exclusively on correcting errors related to missing neighborhood information. This results in improved performance, stable training, and better interpretability.

## 2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is conducted to understand the structure of the dataset, examine price behavior, and justify the modeling choices used later in the project.

### 2.1 Price Distribution

The target variable, property price, exhibits a highly right-skewed distribution in its raw form. To reduce skewness and stabilize variance, a logarithmic transformation is applied. After this transformation, the distribution becomes approximately normal.

This step is essential for improving training stability and ensuring that regression metrics such as Root Mean Squared Error (RMSE) provide meaningful evaluation.

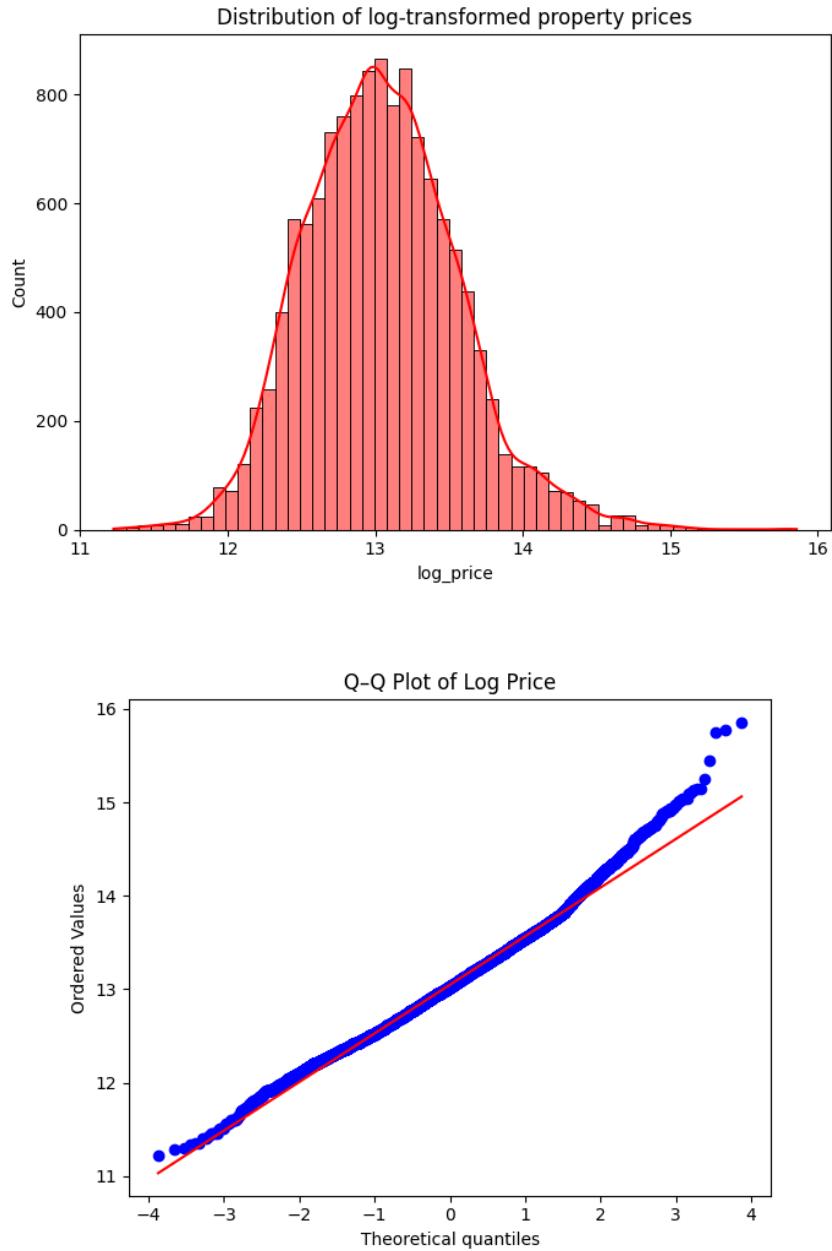


Figure 1: Distribution of log-transformed property prices

## 2.2 Relationship Between Features and Price

Analysis of tabular features shows that structural attributes such as living area, number of rooms, and construction quality have strong relationships with property price. Location-based features, including latitude and longitude, also contribute by capturing broad spatial trends.

However, numerical features alone cannot describe neighborhood appearance. Two properties with similar size and geographic coordinates may still differ substantially in value due to differences in infrastructure, greenery, or surrounding development.

## 2.3 Sample Satellite Images

To assess whether satellite imagery provides additional information, representative satellite images from different neighborhoods are examined.



Figure 2: Sample satellite images across different neighborhoods

The images show clear variation, ranging from densely developed urban regions with extensive road networks to residential areas with moderate development and higher green cover.

## 2.4 EDA Summary

The EDA leads to two primary conclusions. First, tabular features form a strong foundation for property price prediction. Second, satellite imagery captures neighborhood-level context that is difficult to encode using numerical variables. These insights motivate the use of a multimodal approach where satellite imagery complements tabular data.

## 3 Visual and Financial Insights

While tabular models are effective at capturing numerical and structural relationships between property attributes and price, they offer limited insight into how neighborhood-level visual characteristics influence valuation. Features such as connectivity, urban planning quality, and surrounding land use are difficult to represent explicitly using structured variables. To understand the contribution of satellite imagery in the proposed multimodal framework, we analyze the image branch of the model using Grad-CAM (Gradient-weighted Class Activation Mapping).

Grad-CAM enables visual interpretability by highlighting spatial regions in satellite images that contribute most strongly to the model’s predictions. In this project, Grad-CAM is applied to the CNN responsible for predicting residual corrections, as the tabular XGBoost model does not operate on spatial data.

### 3.1 Visual Contributions

Grad-CAM visualizations consistently indicate that the CNN focuses on meaningful spatial features present in satellite images. These features include road networks, building density, intersections, and the overall layout of the surrounding neighborhood. Such patterns are directly related to accessibility, connectivity, and infrastructure development, all of which are known drivers of property value.

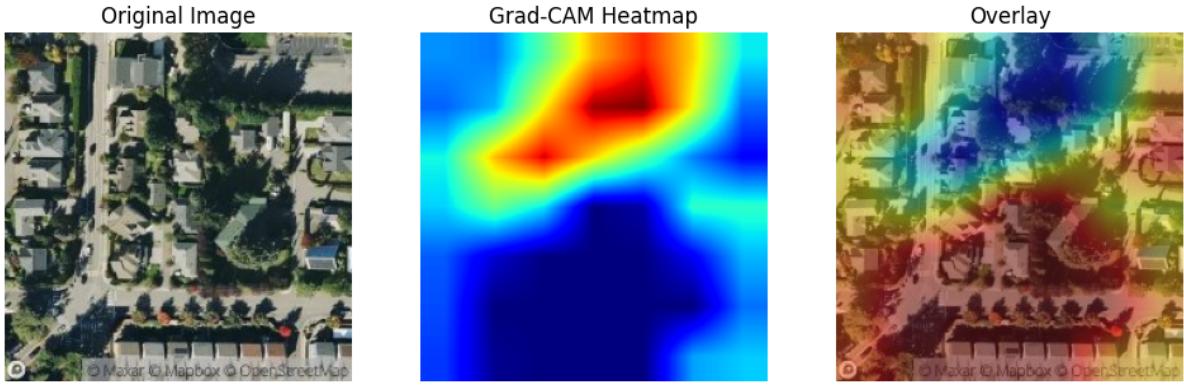


Figure 3: Grad-CAM visualization highlighting influential spatial regions in satellite imagery

The attention maps demonstrate that the CNN does not respond uniformly across the image. Instead, it selectively emphasizes regions that convey economically relevant information. This behavior confirms that the image model learns structured spatial cues rather than relying on superficial or random visual patterns.

### 3.2 Economic Interpretation of Visual Features

The visual features emphasized by Grad-CAM can be interpreted through an economic lens. Areas with dense road networks, compact building structures, and organized residential layouts generally indicate better connectivity and access to services such as transportation, schools, and commercial centers. These characteristics are typically associated with higher demand and, consequently, higher property prices.

In contrast, regions characterized by sparse development, limited road infrastructure, or large unbuilt areas tend to correspond to lower valuations. Such neighborhoods may lack accessibility or essential services, reducing their perceived desirability. These economic relationships are challenging to encode explicitly using tabular features but are naturally captured through satellite imagery.

### 3.3 Positive and Negative Residual Analysis

To further understand the behavior of the residual CNN, we analyze cases where the model predicts positive and negative residual corrections. A positive residual indicates that the CNN increases the baseline XGBoost prediction, suggesting that the tabular model may have underestimated the property’s value due to missing neighborhood context. In these cases, Grad-CAM highlights well-developed urban regions with dense infrastructure and strong connectivity.

Conversely, negative residuals occur when the CNN decreases the baseline prediction. Here, the attention shifts toward sparsely developed or poorly connected areas, implying that the tabular model may have overestimated the value. This directional behavior indicates that the CNN performs context-dependent corrections rather than applying uniform adjustments.



Figure 4: Grad-CAM comparison for positive and negative residual cases

### 3.4 Consistency and Generalization Across Properties

To ensure that the observed attention patterns are not limited to a small number of examples, Grad-CAM is evaluated across multiple validation samples. The resulting attention maps exhibit consistent focus on built-up areas, transportation structures, and neighborhood layouts across different properties.

This consistency suggests that the CNN captures generalizable neighborhood-level features rather than relying on property-specific artifacts. As a result, the visual branch contributes stable and economically meaningful information that complements the tabular model across the dataset.

## 4 Architecture Diagram

The proposed model follows a multimodal residual fusion architecture that combines tabular data and satellite imagery in a structured and interpretable manner. The design assigns distinct roles to each modality to ensure stable learning and effective integration.

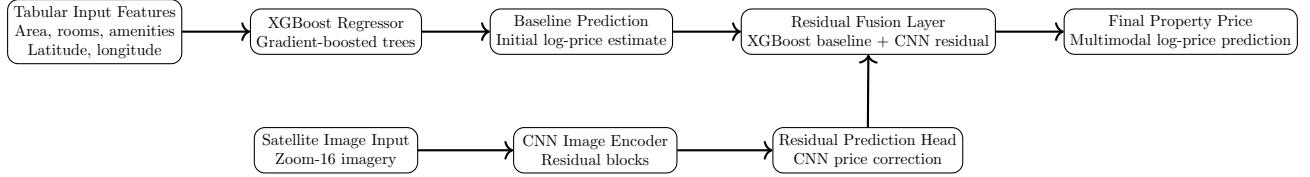


Figure 5: multimodal residual fusion architecture

The tabular branch processes structured property attributes and geographic coordinates using an XGBoost regressor to produce a baseline log-price prediction. This baseline captures most of the predictable variation in property prices using numerical information alone.

The image branch processes satellite images using a Convolutional Neural Network (CNN) to extract neighborhood-level visual features. Instead of predicting prices directly, the CNN estimates a residual correction that refines the baseline prediction.

The residual fusion layer combines the baseline prediction and the visual residual to produce the final property price estimate. This approach allows satellite imagery to contribute complementary information without overpowering the tabular signal.

## 5 Results

Model performance is evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination ( $R^2$ ). All models are assessed on the same validation split to ensure a fair comparison.

### 5.1 Tabular Data Only Baseline

The tabular-only XGBoost model provides a strong baseline, confirming that structured features explain a large portion of the variance in property prices. However, it does not incorporate neighborhood-level visual context.

### 5.2 Multimodal Models with Satellite Imagery

Naive multimodal fusion strategies result in degraded performance due to unstable learning and dominance of tabular features. In contrast, the proposed residual fusion model achieves improved performance by allowing satellite imagery to correct errors in the tabular baseline.

### 5.3 Performance Comparison

Table 1: Performance comparison of tabular and multimodal models

| Model                       | RMSE         | $R^2$       |
|-----------------------------|--------------|-------------|
| Tabular Data Only (XGBoost) | 0.275        | 0.72        |
| Naive Multimodal Fusion     | Higher       | Lower       |
| Weighted Fusion             | Improved     | Moderate    |
| Residual Fusion (Proposed)  | <b>0.257</b> | <b>0.76</b> |

## 5.4 Discussion

Overall, the observed improvements confirm that satellite imagery provides complementary neighborhood-level information rather than redundant noise. Residual fusion enables this visual information to be incorporated in a controlled and interpretable manner, resulting in consistent gains in predictive performance.