

# SATELLITE IMAGE BASED PROPERTY VALUATION

---

**Open Project : CDC X Yhills**

**Submitted by : Aaryan Kumar**

**Date : 04-01-25**

## **Overview: Approach and Modelling Strategy**

The strategy used centred on the hypothesis that traditional tabular data like bedrooms, square footage, etc. misses "curb appeal" and neighbourhood context. This led to the usage of a Multi Data Fusion Model.

1. Data Acquisition :

A custom Python Script was used to fetch over 21,000 satellite images corresponding to their specific geographic coordinates.

2. Preprocessing :

The tabular records were synchronised with the available images. Logical standards, such as the International Residential Code (IRC) were applied to identify data anomalies.

3. Scaling :

Model stability was ensured by performing a logarithmic transformation on the house prices, which were then normalised to fit in the 0 to 1 range.

4. Modelling :

Two parallel neural networks were built. First, a Convolutional Neural Network (CNN) for processing visual data. Second, a Multi-Layer Perceptron (MLP) for tabular data, thereby merging into a final regression head.

## Exploratory Data Analysis

Before downloading satellite images, duplicate entries were identified.

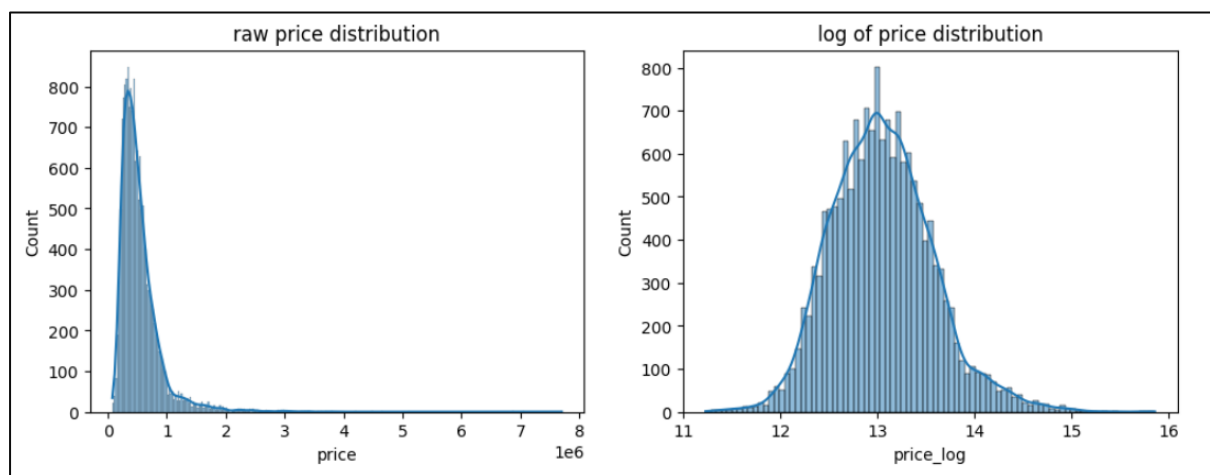
Later, a minimum value of 70 square feet per room was taken as a standard from the International Residential Code (IRC)

```
flagging outliers (< 70 sqft/bedroom)...  
code flagged 1 items:
```

	id	bedrooms	sqft_living	sqft_per_bedroom
<b>3193</b>	2402100895	33	1620	49.090909

```
Flagged values were corrected
```

The raw price data exhibited a heavy right skew, with a few high value luxury properties stretching the distribution. By applying  $y = \ln(1 + \text{price})$ , we normalized the variance, allowing the model to focus on percentage error rather than absolute dollar differences.



Sample images showed significant variance in density. Properties with higher "greenery ratios" (trees/parks) were statistically correlated with higher price-per-square-foot, even when structural features like 'sqft\_living' were identical.

## Financial and Visual Insights

Through the fusion of visual and structural data, several key drivers of value were identified:

- Visual Drivers :

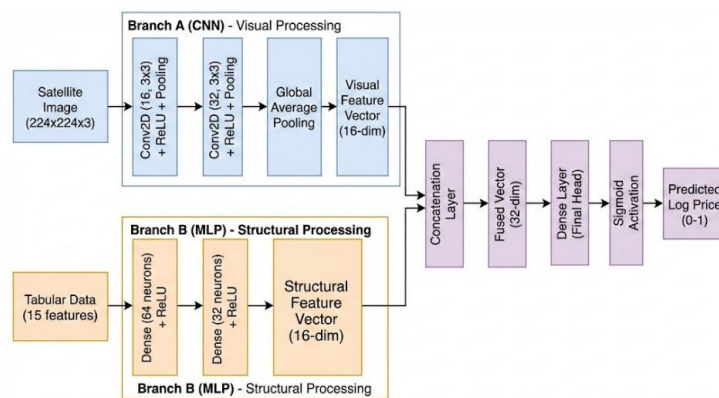
Satellite imagery revealed that waterfront proximity and neighbourhood density significantly impacted valuation. The CNN picked up on "concrete-heavy" urban areas

versus "foliage-rich" suburban areas, where foliage acted as a proxy for premium property value.

- **Structural Drivers:** Feature scaling revealed that grade (construction quality) and 'sqft\_living' remained the strongest tabular predictors.
- **Interaction Effect:** The model excelled at identifying cases where a high-grade house in a "low-greenery" area was priced lower than expected, a nuance often missed by purely tabular models.

## Architecture Diagram

The following diagram illustrates the "Fusion" architecture used in the `model_training.ipynb` file:



- **Branch A (CNN):** 16-32 filter Conv2D layers followed by **Global Average Pooling** to extract a 16-dimensional visual feature vector.
- **Branch B (MLP):** 64-32 neuron Dense layers to extract a 16-dimensional structural feature vector.
- **Concatenation Layer:** Fuses the 16 (Visual) + 16 (Structural) features into a single **32-dimensional vector**.
- **Final Head:** A Sigmoid-activated output layer predicting the normalized Log Price.

## Results

The model obtained an  $R^2$  score of 0.7106 and an RMSE of approximately 122,400 dollars.