Multimodal House Price Prediction

using Tabular Data and Satellite Imagery

Submitted by: Piya Solanki

Enrolment Number: 24113097

Branch: Civil Engineering

Tech stack

- Python, Pandas, NumPy
- PyTorch, Torchvision
- Scikit-learn
- OpenCV, PIL
- Mapbox Static Images API

Date Of submission- 7th Jan 2026

(CDC X Yhills Open Project)

1. OVERVIEW

The objective of this project is to build a **multimodal regression pipeline** to predict residential house prices by combining **structured housing attributes** with **satellite imagery–based visual context**.

Traditional house price models rely heavily on tabular data such as number of bedrooms, square footage, and location coordinates. This project extends that approach by incorporating **environmental context** extracted from satellite images using **Convolutional Neural Networks (CNNs)**.

The pipeline integrates:

- Tabular regression models as strong baselines
- Satellite image feature extraction using a pretrained CNN
- Feature-level fusion of tabular and visual data
- Comparative evaluation between tabular-only and multimodal models

2. DATASET DESCRIPTION

The base dataset consists of historical housing data containing property-level attributes and geographic coordinates.

Tabular Features include:

- Price (target variable)
- Bedrooms and bathrooms
- Living area and lot size
- Condition and construction grade

- Latitude and longitude
- Neighbourhood density indicators

Visual Data:

Using the latitude and longitude of each property, satellite images were programmatically fetched via the Mapbox Static Images API. These images provide visual context of the surrounding environment such as vegetation, proximity to roads, and urban density.

A subset of images was used to ensure computational feasibility while maintaining representative coverage of the dataset.

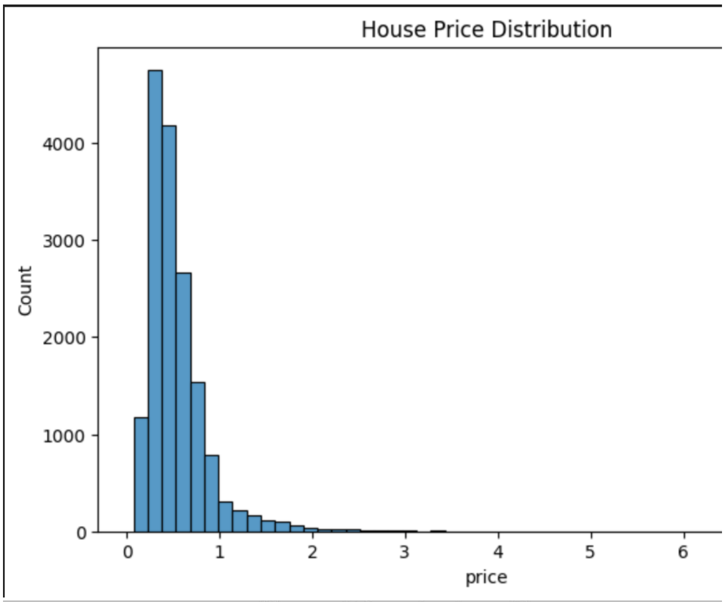
Image location:
data/images/train/
data/images/test/

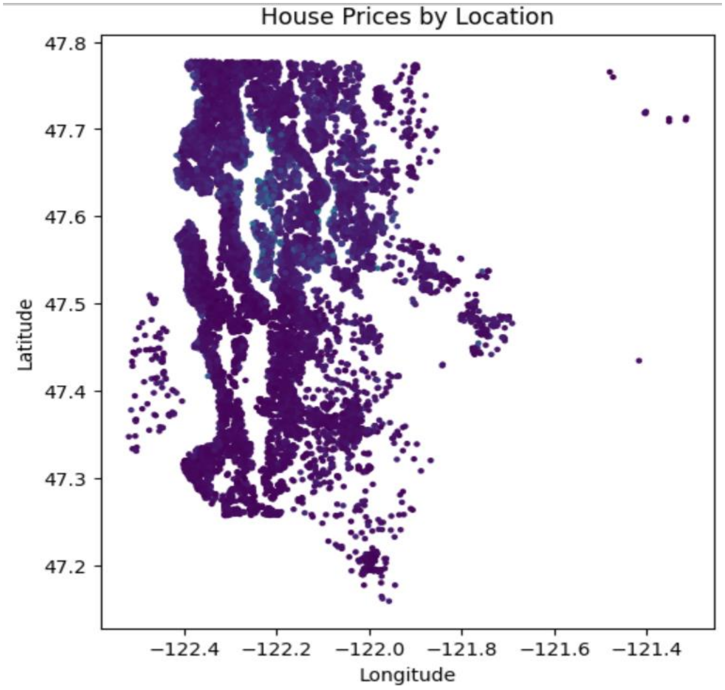
3. EXPLORATORY DATA ANALYSISE (EDA)

Exploratory analysis was conducted to understand the distribution of housing prices and key structural attributes.

The price distribution shows a right-skewed pattern, indicating a majority of mid-range priced houses with a smaller number of high-value properties.

Relationships between price and features such as living area, grade, and waterfront presence were examined, confirming known real-estate trends where larger and better-quality homes command higher prices





4. METHODOLOGY

4.1 Tabular Modelling

The tabular data was cleaned and preprocessed by handling missing values, removing non-numeric fields, and standardizing features where required.

Baseline regression models were trained using only tabular features to establish a reference performance.

4.2 Satellite Image Feature Extraction

Satellite images corresponding to each property location were processed using a pre-trained Convolutional Neural Network (ResNet-18). The CNN was used as a fixed feature extractor, generating high-dimensional embeddings that capture visual characteristics of the neighbourhood.

4.3 Feature Fusion

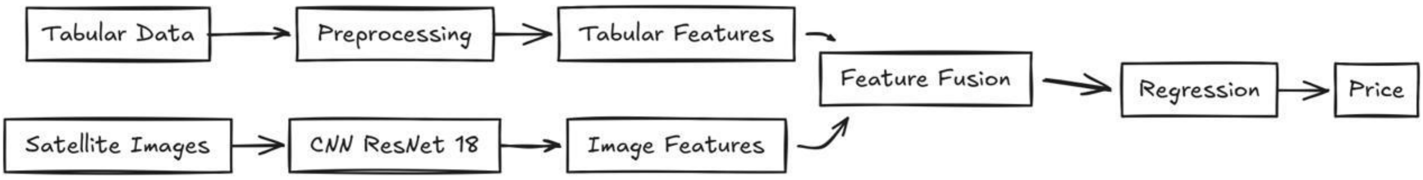
The extracted image embeddings were concatenated with the tabular feature vectors to form a fused multimodal representation.

This fused dataset was then used to train regression models, enabling the model to leverage both numerical property attributes and visual environmental context.

4.4 Multimodal Training

The fused tabular and image feature dataset was used to train a regression model for multimodal house price prediction. A linear regression model was employed to evaluate the effectiveness of late feature fusion. Model performance was assessed using a train–validation split, and evaluation metrics included Root Mean Squared Error (RMSE) and R^2 score.

5. ARCHITECHTURE DIAGRAM



6. RESULTS AND COMPARISONS

Model performance was evaluated using Root Mean Squared Eroor (RMSE) and R^2 score.

Model	RMSE	R^2
Linear regression (Tabular)	191570.99	0.708
Random Forest (Tabular)	129856.31	0.866
Multimodal (Tabular + Images)	284313.88	0.033

7. DISCUSSIONS

The experimental results demonstrate that tabular features remain the dominant predictors of house prices. While the multimodal model incorporating satellite imagery achieved a positive R^2 score, its performance did not surpass the tabular-only baselines. This suggests that naïve late fusion of visual and tabular features may be insufficient to capture complex interactions between spatial context and property attributes.

8. LIMITATIONS AND FUTURE WORK

The primary limitation of this study is the use of a limited number of satellite images and a simple feature fusion strategy. Future work could explore end-to-end multimodal learning, attention-based fusion mechanisms, and higher-resolution satellite imagery to better capture neighbourhood-level context.

9. CONCLUSIONS

This project successfully implemented a complete multimodal regression pipeline for house price prediction. Although tabular models outperformed the multimodal approach, the integration of satellite imagery demonstrates the feasibility of incorporating visual environmental context into real estate valuation models. The findings highlight both the potential and challenges of multimodal learning in practical settings.