1. Overview

The objective of this project is to predict residential property prices by combining traditional tabular housing attributes with satellite imagery. Conventional valuation models rely heavily on structured features such as house size, number of rooms and location coordinates. However, these features do not explicitly capture the surrounding environment, which plays an important role in determining property value.

To address this, a multimodal regression approach is implemented. Satellite images corresponding to each property's geographic coordinates are programmatically downloaded and processed to extract visual features representing neighbourhood context. These visual features are then combined with tabular data to evaluate whether incorporating satellite imagery improves price prediction accuracy.

A baseline model using only tabular features is first developed. This is followed by a multimodal model where image-derived features are fused with tabular features and used for regression. Model performance is compared using RMSE and R^2 score.

2. Exploratory Data Analysis (EDA)

Price Distribution

The target variable (price) shows a right-skewed distribution. Most properties are concentrated in the lower to mid-price range, while a smaller number of high-value properties form a long tail. This skewness motivates the use of ensemble-based regression models that can handle non-linear relationships.

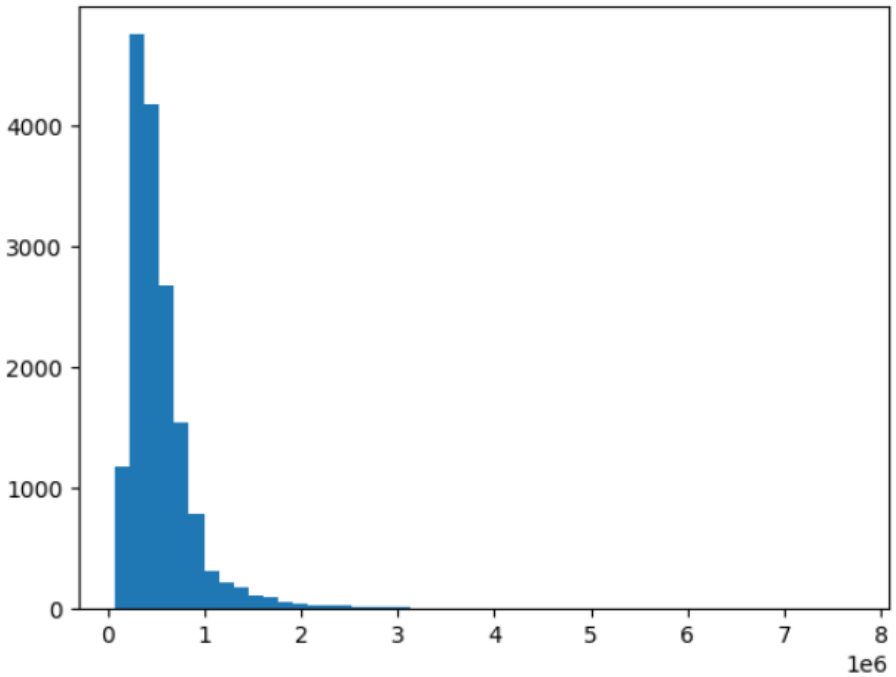

Fig 1: Distribution of property prices in the training dataset

Relationship Between Living Area and Price

A scatter plot between sqft. living and price shows a clear positive relationship. As the living area increases, property prices generally increase, although variance grows for larger houses.

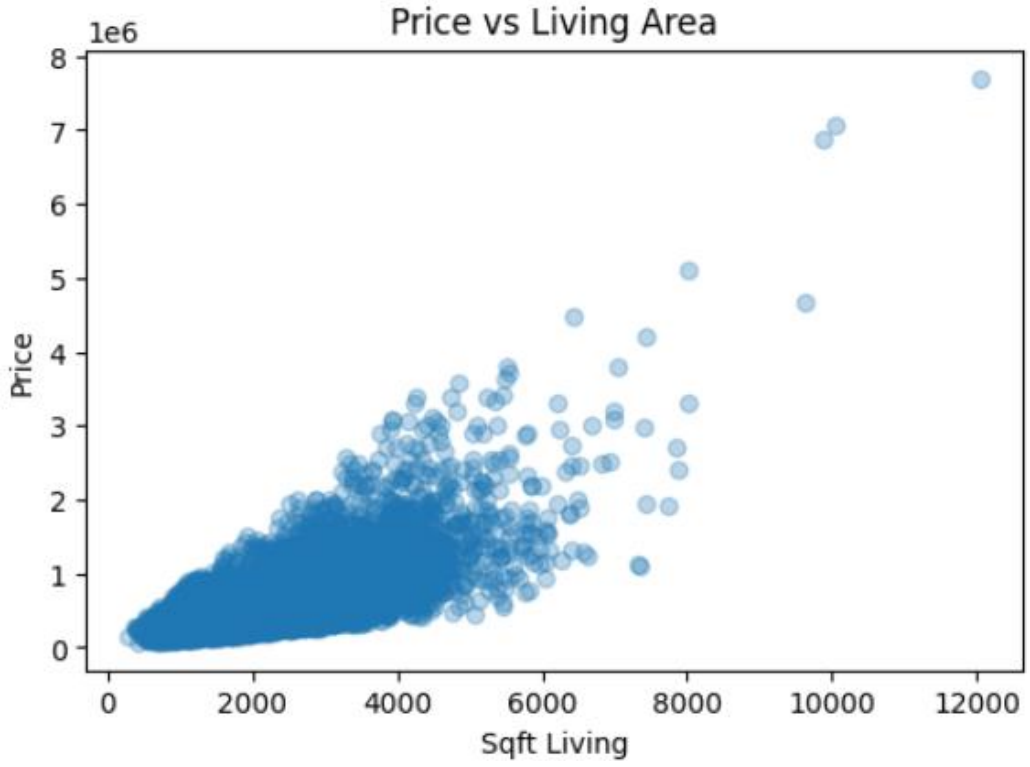

Fig 2: Relationship between living area and property price

Satellite Image Samples

Sample satellite images were visualized to ensure correctness of data acquisition. The images capture rooftops, road networks, vegetation and surrounding urban patterns. Due to cloud coverage and API limitations, only a subset of properties yielded usable images. Although individual structures are not visually distinguishable due to low resolution of obtained images, the satellite tiles capture spatial patterns such as vegetation density, road layout and built-up texture, which are sufficient for CNN-based feature extraction.

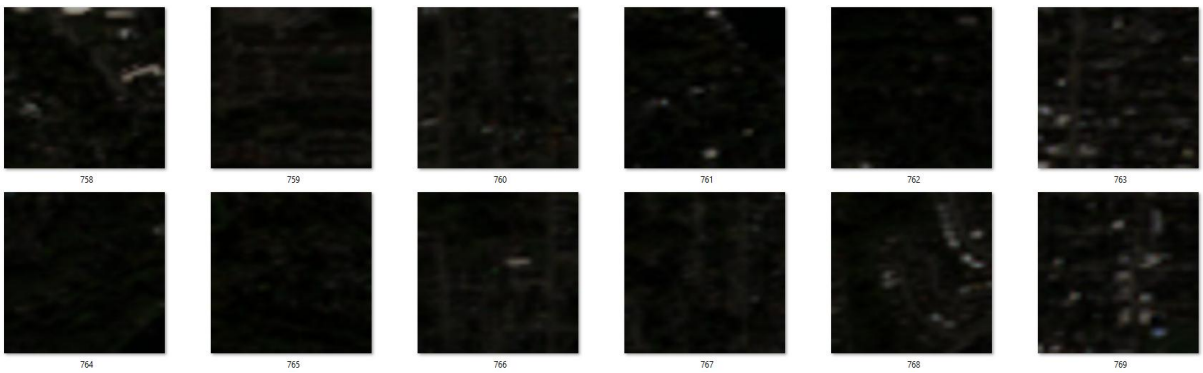

Fig 3: Sample satellite image tiles used for CNN feature extraction

Financial and Visual Insights

Individual buildings and fine-grained urban features are not distinctly recognizable at the selected resolution by visual inspection of the image tiles. Instead, the imagery captures broader patterns such as overall built-up texture, vegetation presence and spatial layout.

However, CNN is able to extract high-dimensional feature representations that implicitly encode such spatial patterns. This provides contextual information related to surrounding land use and neighbourhood characteristics that are not directly represented in the tabular attributes.

The limited number of usable satellite images after cloud filtering reduces the coverage of visual data and introduces sparsity in the multimodal dataset. As a result, while the inclusion of image-based features adds contextual richness, it also increases variability in model performance, leading to a higher prediction error compared to the tabular-only baseline.

3. Model Architecture

The multimodal pipeline consists of two parallel components:

1. Tabular Branch
   - Input: Cleaned numerical housing features
   - Model: Random Forest Regressor

2. Image Branch
   - Input: Satellite images resized to 224×224
   - Feature Extractor: ResNet18 pretrained on ImageNet
   - Output: 512-dimensional image embedding

Outputs from both branches are concatenated (early fusion) and used as input to a regression model for price prediction.

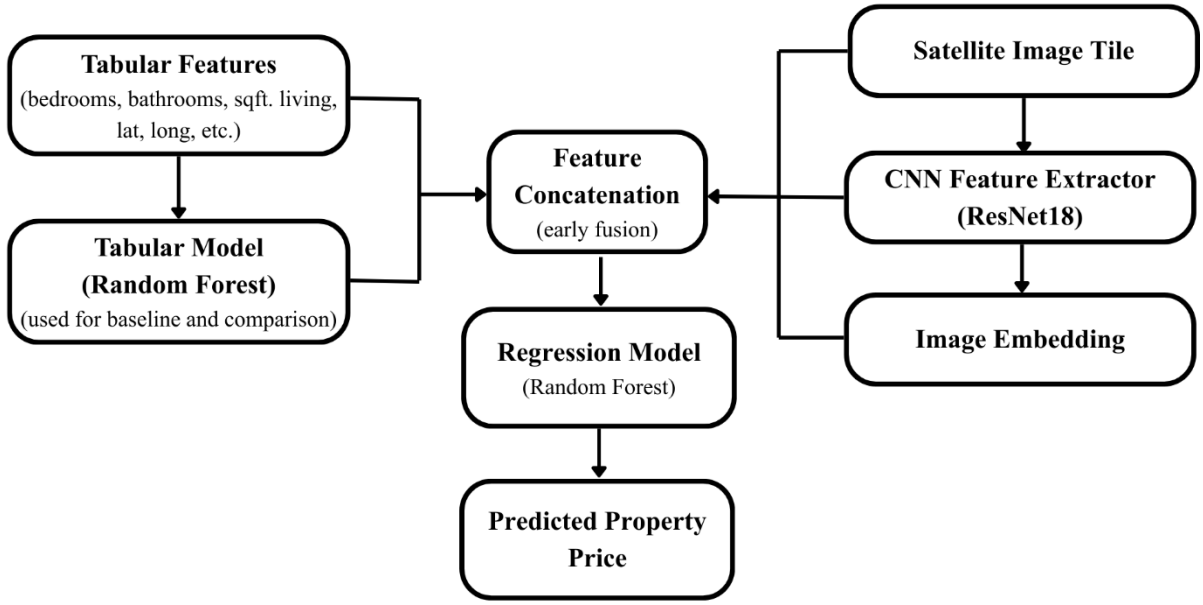

Fig 4: Multimodal model architecture combining tabular and satellite data

4. Results

Even though the multimodal model conceptually sounds better, the tabular-only model achieves stronger performance due to the full availability and reliability of structured features. This is primarily because only 3352 out of 16209 properties had valid satellite images after cloud filtering.

Model Type	RMSE	R^2 Score
Tabular Only	130112	0.865
Tabular + Satellite Images	194147	0.793

Table 1: Multimodal model architecture combining tabular and satellite data

Despite of lower numerical performance, the multimodal approach demonstrates the feasibility of integrating visual context into property valuation pipelines and highlights the practical challenges involved in large-scale image acquisition.

5. Conclusion

This project demonstrates an end-to-end multimodal regression workflow for property valuation using tabular and satellite image data. While tabular features remain dominant in predictive power, satellite imagery provides valuable contextual information. The results highlight the importance of data quality and coverage when working with visual inputs and suggest that improvements in image availability and preprocessing could enhance multimodal performance in future work.