Satellite Imagery Based Property Valuation

PROJECT REPORT *By 23116042*

1. Introduction

Property price prediction is a common problem in real estate analytics. Traditional approaches use structured information such as the number of bedrooms, bathrooms, living area, and geographic location. While these features are useful, they may not fully represent the surrounding environment of a property.

This project investigates whether **satellite imagery** can provide additional contextual information to improve property valuation. A multimodal approach is used where **tabular housing data** is combined with **visual features extracted from satellite images**. The performance of this approach is compared with a simple tabular baseline model.

2. Dataset Description

2.1 Training Dataset

The training dataset contains **16,209 residential properties** with **21 attributes**, including:

- Property price (target variable)
- Bedrooms
- Bathrooms
- Living area (sqft_living)
- Lot size
- Floors
- Waterfront
- View
- Condition
- Grade

- Latitude and Longitude
- Year built and renovation details

No missing values were found in the dataset.

2.2 Test Dataset

The test dataset contains the same features as the training dataset **except the price column**, which is to be predicted.

2.3 Satellite Imagery

Each property is associated with a satellite image capturing the surrounding area. These images provide spatial context such as nearby infrastructure, road connectivity, and land layout, which are not explicitly present in tabular features.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the data distribution and relationships between key variables.

3.1 Distribution of House Prices

A histogram was plotted to visualize the distribution of house prices.

Observation:

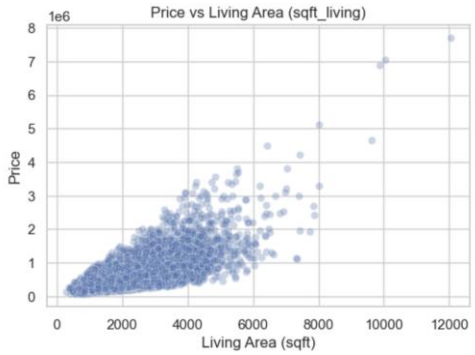
House prices show a **right-skewed distribution**, with most properties priced in the lower to mid range and fewer high-priced outliers

3.2 Price vs Living Area

A scatter plot was created between sqft_living and price.

Observation:

There is a strong positive relationship between living area and price. Larger homes tend to have higher prices, though price variability increases for larger properties.

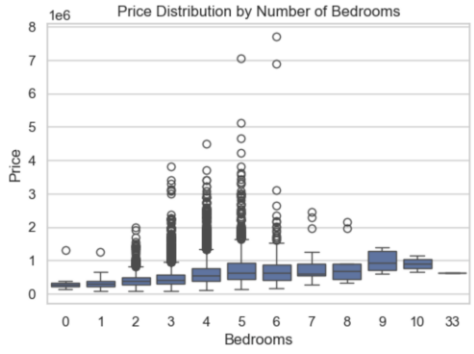


3.3 Price vs Number of Bedrooms

A box plot was used to compare price distributions across different bedroom counts.

Observation:

Median price generally increases with the number of bedrooms, but there is significant overlap between categories, indicating that bedrooms alone do not determine price.



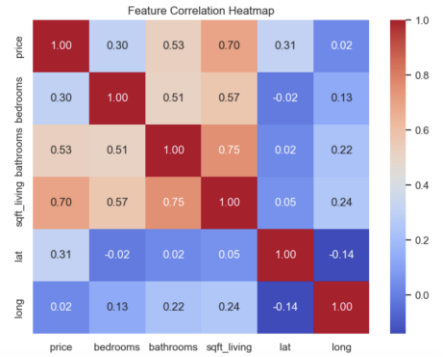
3.4 Feature Correlation Analysis

A correlation heatmap was generated using key numerical features.

Observation:

- Strong correlation between price and sqft_living
- Moderate correlation with bathrooms

- Weak correlation with latitude and longitude individually



4. Tabular Baseline Model

4.1 Model Description

A **Linear Regression** model was trained using only tabular features:

- Bedrooms
- Bathrooms
- Living area
- Latitude
- Longitude

The dataset was split into training and validation sets using an **80–20 split**.

4.2 Evaluation Metrics

The model was evaluated using:

- Root Mean Squared Error (RMSE)
- R^2 Score

4.3 Results (Tabular Baseline)

Metric	Value
RMSE	≈231,000
R^2	≈ 0.57

Interpretation:

The baseline model explains approximately 57% of the variance in property prices but exhibits relatively high prediction error.

5. Multimodal Model (Tabular + Satellite Imagery)

5.1 Image Feature Extraction

Satellite images were processed using a **pretrained ResNet-50** network. The CNN was used as a **fixed feature extractor**, producing a **2048-dimensional feature vector** for each image.

No fine-tuning of the CNN was performed.

5.2 Feature Fusion

The extracted CNN image features were concatenated with the tabular features to form a combined feature vector. A **Linear Regression** model was trained on this multimodal feature space.

5.3 Evaluation Metrics

The same metrics (RMSE and R^2) were used to ensure fair comparison with the baseline model.

5.4 Results (Multimodal Model)

Metric	Value
RMSE	≈123,000
R^2	≈ 0.52

6. Model Comparison and Discussion

Model	RMSE	R^2
Tabular Baseline	~231k	~0.57
Multimodal Model	~123k	~0.52

Key Insights:

- The **multimodal model significantly reduces RMSE**, indicating better absolute prediction accuracy.

- The **R^2 score slightly decreases**, suggesting less consistent variance explanation across all samples.

  The addition of satellite imagery helps reduce large prediction errors by providing contextual visual information. However, using a simple linear model limits the effective utilization of complex image features, which may explain the slight drop in R^2 score.

7. Future Scope

Future improvements could include:

- Fine-tuning CNN layers on property-specific imagery
- Using non-linear regression models
- Implementing explainability methods for image-based features
- Adding richer geographic and neighborhood features

8. Conclusion

This project shows that adding satellite images to traditional housing data can help reduce prediction errors when estimating property prices. While the multimodal model improves how close the predictions are to actual values, it does not consistently explain price variations across all properties. These results highlight both the promise of using visual information and the challenges involved in combining image and tabular data for real-world price prediction tasks.