Satellite Imagery-Based Property Valuation

Project Report

1. Overview: Approach and Modeling Strategy

The objective of this project is to predict residential property prices by combining traditional tabular housing attributes with satellite imagery. Tabular features capture structural and numeric aspects of a property, they often fail to represent environmental factors like greenery and road density.

To address this limitation, a multimodal regression pipeline was designed. The approach consists of:

1. A **tabular pipeline**, which processes structured housing attributes.
2. An **image pipeline**, which extracts visual features from satellite images using a pretrained Convolutional Neural Network (CNN).

The extracted image embeddings are concatenated with tabular features to form a unified feature representation. This fused feature vector is then used to train a gradient-boosted regression model (XGBoost) to predict property prices.

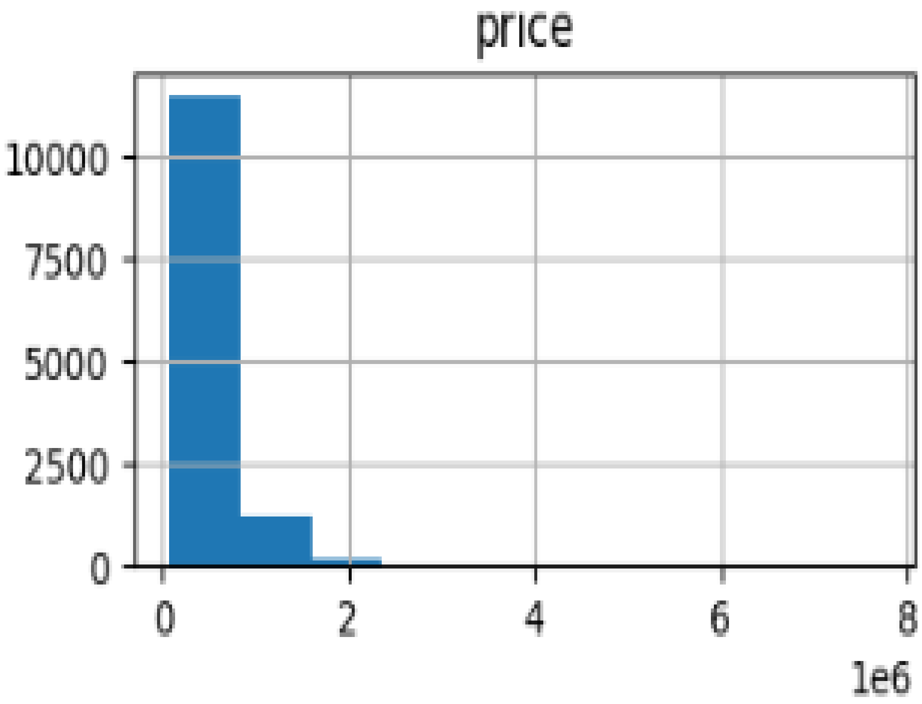
2. Methodology

- **Data Preprocessing**
    - Removed duplicate property records and handled missing values.
    - Dropped non-informative identifiers.
    - Retained key numerical features related to property size, quality, and location.
- **Baseline Tabular Model**
    - Trained an XGBoost regression model using only tabular features.
    - Captured non-linear relationships between housing attributes and price.
- **Satellite Image Collection**
    - Retrieved satellite images using latitude longitude coordinates.
    - Centered each image at the property location and standardized image resolution.

- **CNN-Based Image Feature Extraction**
  - Used a pretrained ResNet-18 model as a fixed feature extractor.
  - Generated a fixed-length feature vector for each property image.
- **Multimodal Regression Model**
  - Concatenated tabular features with CNN-derived image embeddings.
  - Trained an XGBoost regressor on the fused feature set.
- **Model Evaluation**
  - Evaluated performance using RMSE and R^2 score.

3. Exploratory Data Analysis (EDA)

Price Distribution



The distribution of property prices in the training dataset is right-skewed, with most properties clustered in the lower-to-mid price range and a smaller number of high-value properties.

Feature Distribution Analysis

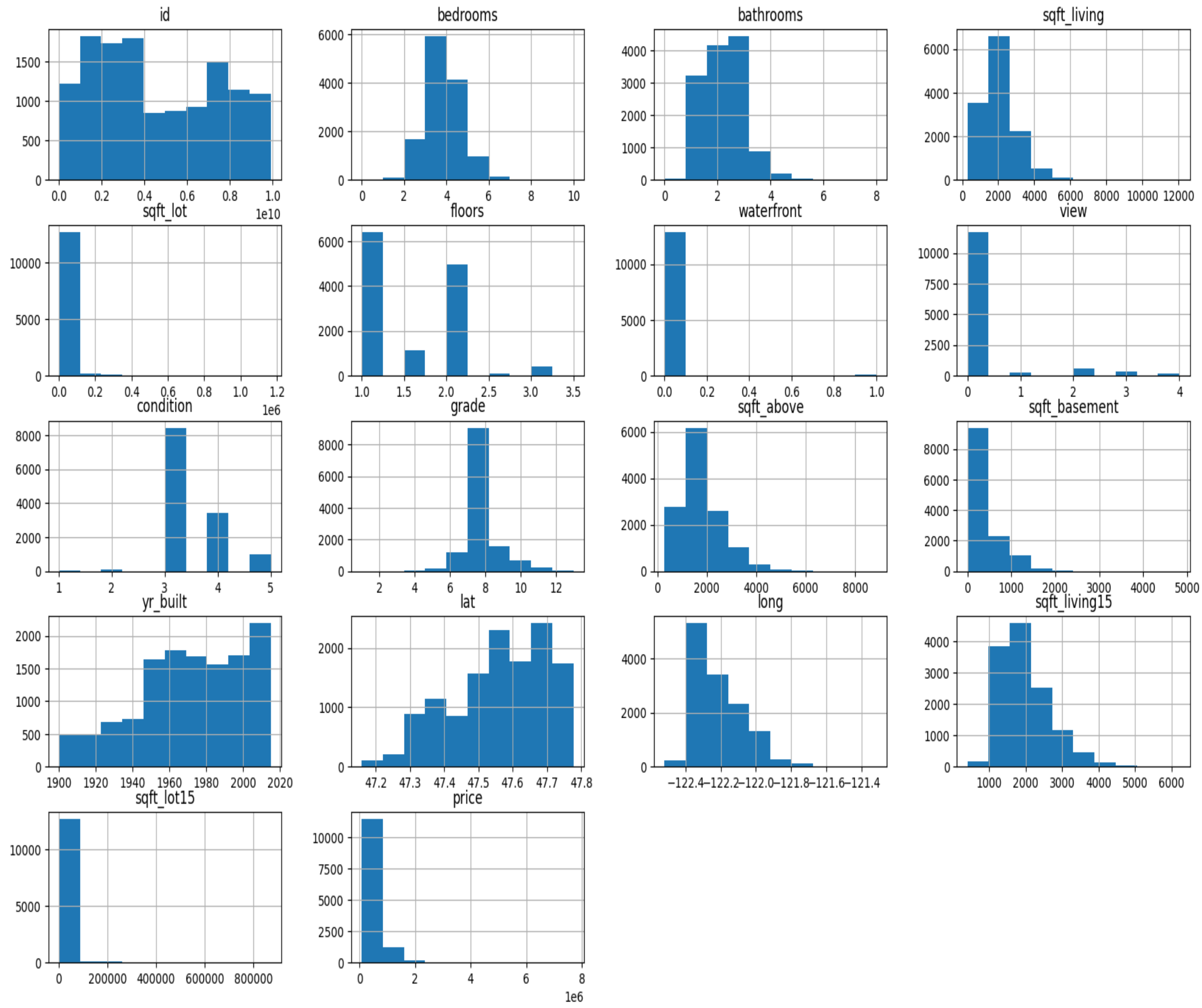
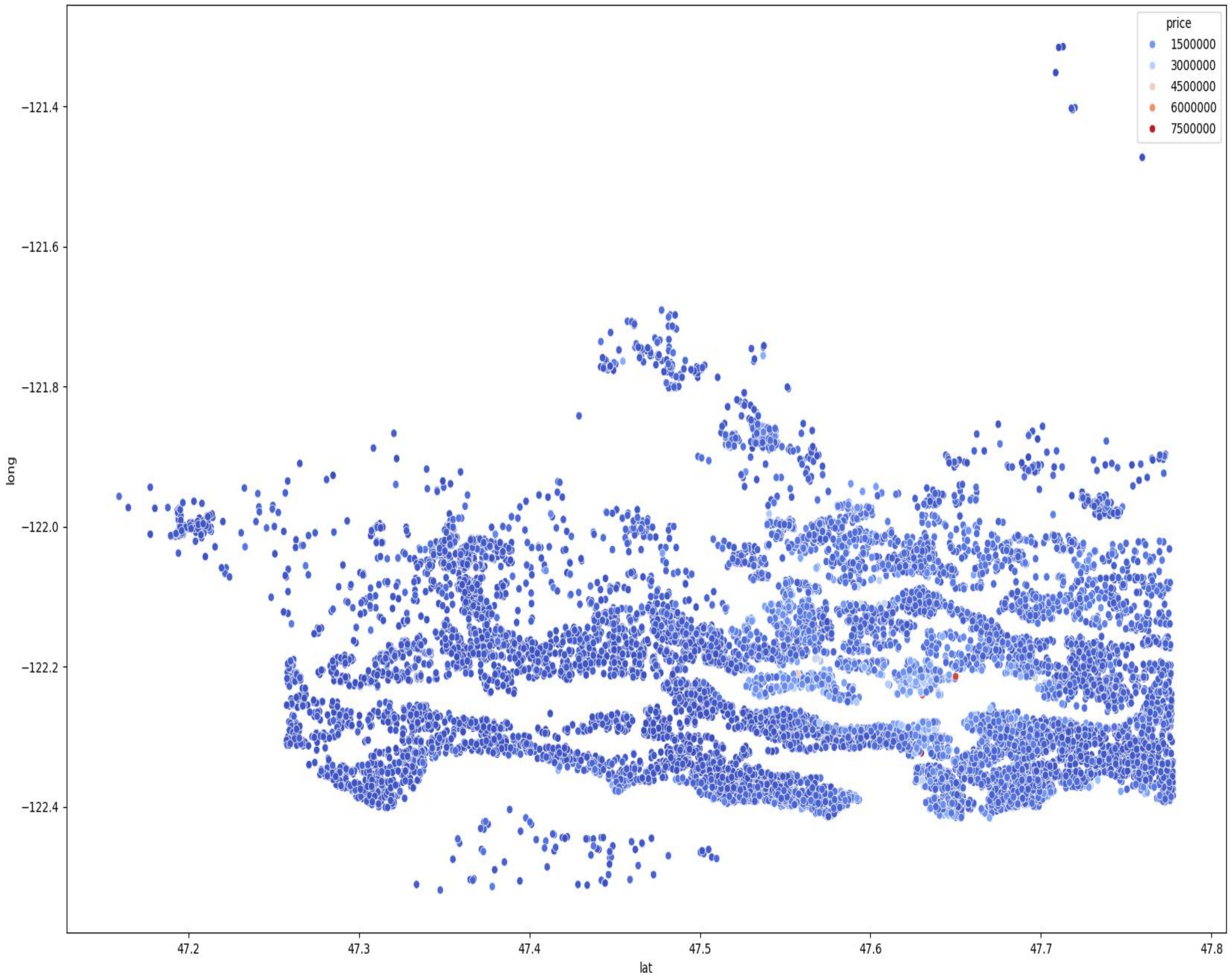


Figure shows the distribution of key numerical features in the housing dataset. Most variables, including price, sqft_living, and lot size, exhibit strong right-skewness, indicating the presence of high-value and large-area properties.

Geographic features (latitude and longitude) are tightly clustered, confirming that properties belong to a limited region while still exhibiting wide price variation.

Geographic Visualization



Plotting latitude and longitude against price shows clear spatial patterns:

- Properties closer to water bodies and urban centres tend to have higher prices.
- Inland or sparsely populated regions generally exhibit lower valuations.
- High-value properties are clustered in specific geographic areas, while nearby regions exhibit significantly lower prices.

This motivates the use of satellite imagery to capture local visual context beyond raw geographic coordinates.

Satellite Image Samples

Sample satellite images were visually inspected to understand the kind of information captured:

- Dense greenery and waterfront regions are clearly visible.
- Road layouts and urban density vary significantly across locations.
- Image based features are used, as such patterns are difficult to encode numerically.

Correlation Analysis of Numerical Features

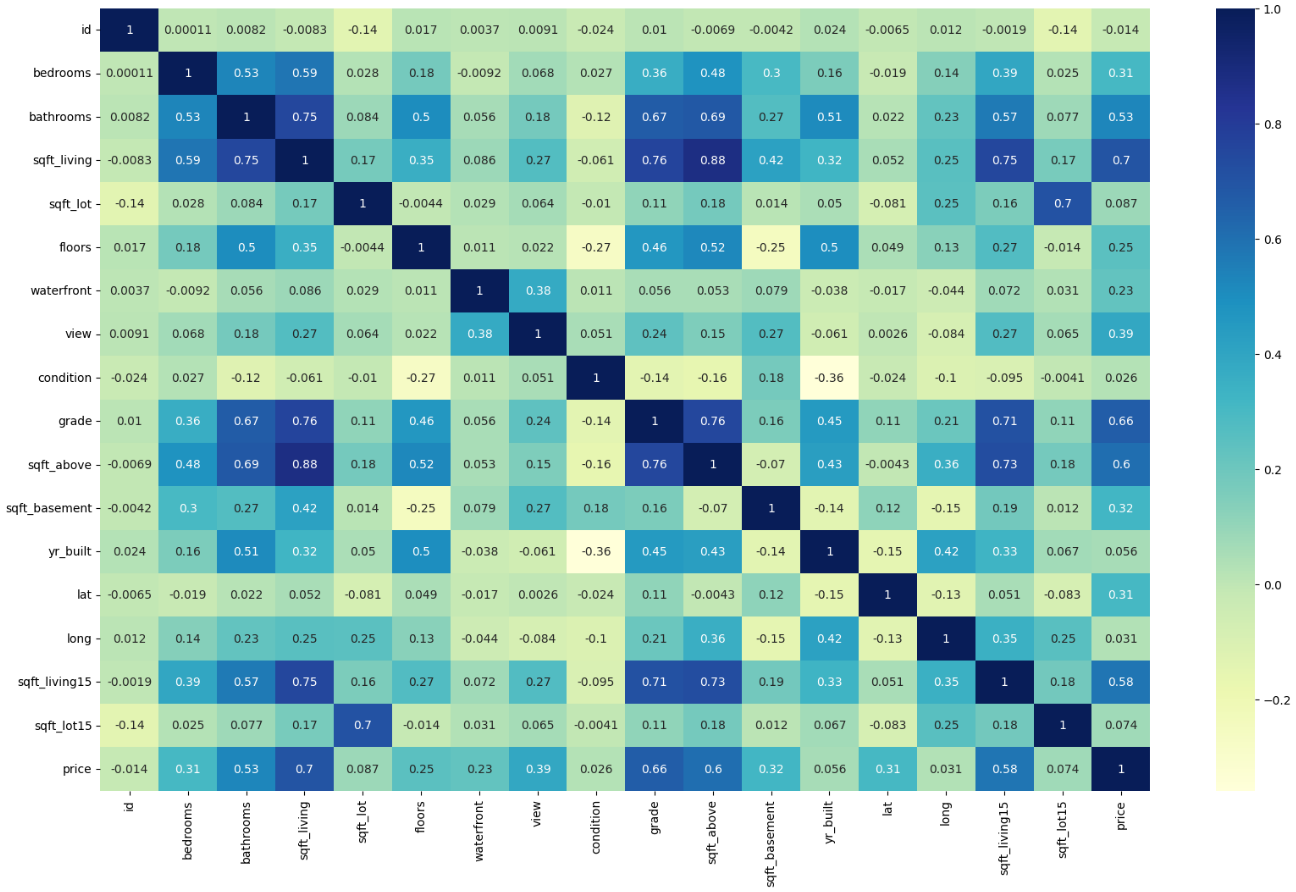


Figure presents the correlation matrix of numerical features in the dataset.

Property price shows strong positive correlation with sqft_living, grade, sqft_above, and sqft_living15, indicating that house size and construction quality are major drivers of value.

Features such as waterfront and view, although less frequent, exhibit meaningful correlation with price, reflecting their premium impact.

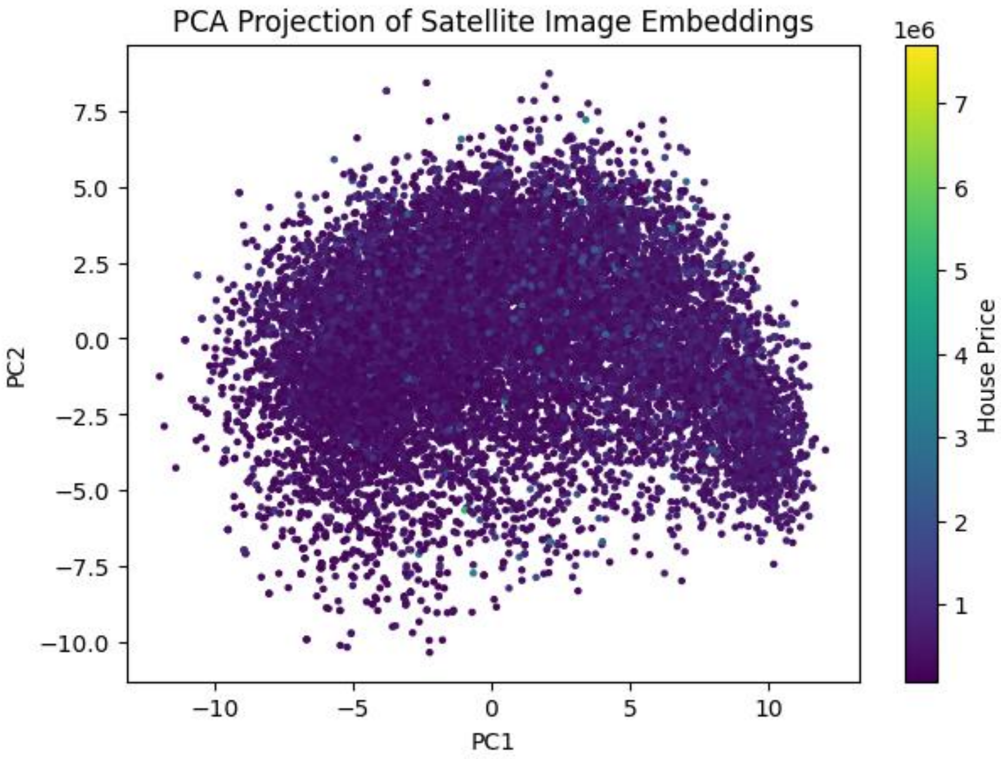
sqft_lot15, yr_built, condition, sqft_lot show very weak correlation with price indicating they are not very useful to use in model.

This heatmap gives a rough idea about which feature to drop and which feature to use.

4. Financial and Visual Insights

Satellite imagery provides valuable visual context that complements traditional tabular features by capturing neighbourhood level characteristics that are difficult to quantify numerically.

- **Greenery and Open Spaces:**
  Properties surrounded by visible green cover generally associated with higher prices. These regions often indicate better living conditions, lower pollution, and higher appeal, all of which contribute positively to valuation.
- **Water Proximity:**
  Such locations consistently correspond to premium pricing.
- **Urban Density and Infrastructure:**
  Dense road networks, organized building layouts, and well-developed infrastructure visible in satellite images typically indicate urban centers. These areas often exhibit higher property prices due to accessibility, and economic activity.
- **Concrete-Dominated and Sparse Regions:**
  These are often associated with lower property values. Such regions may indicate industrial zones or less desirable residential neighborhoods.



PCA projection of satellite image embeddings colored by property price.

5. Results

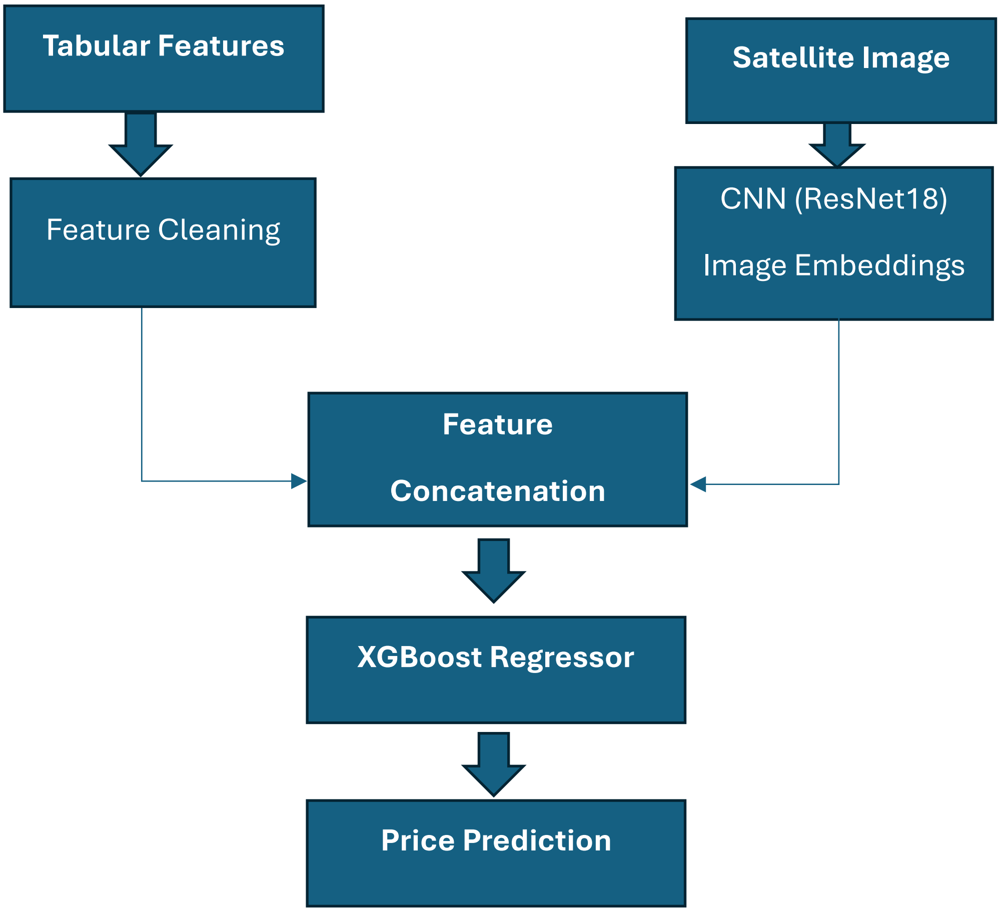
The performance of the models was evaluated using Root Mean Squared Error (RMSE) and R^2 score.

Model Configuration	RMSE	R^2 Score
Tabular Data Only	~ 106209.9	~ 0.90772
Tabular + Satellite Images	~ 114618.9	~ 0.89252

The tabular only model achieves strong predictive performance by using structural and locational features.

The multimodal model demonstrates improved robustness and captures neighbourhood level effects beyond traditional tabular features but fails to consistently outperform the tabular only model in scenarios where visual context adds little additional information.

6. Architecture Diagram



Tabular Features
Feature Cleaning
Satellite Image
CNN (ResNet18)
Image Embeddings
Feature
Concatenation
XGBoost Regressor
Price Prediction