Multimodal House Price Prediction using Satellite Imagery and Tabular Data

1. Overview

Accurate property valuation requires both **structural attributes** of a house (size, rooms, condition) and **environmental context** (greenery, water proximity, road networks). This project builds a compact and reproducible multimodal pipeline that fuses these information sources:

- **Tabular features:** Bedrooms, bathrooms, area metrics, grade, condition, and location (latitude/longitude).
- **Visual features:** Satellite images programmatically fetched per property and converted into embeddings by a pretrained CNN.

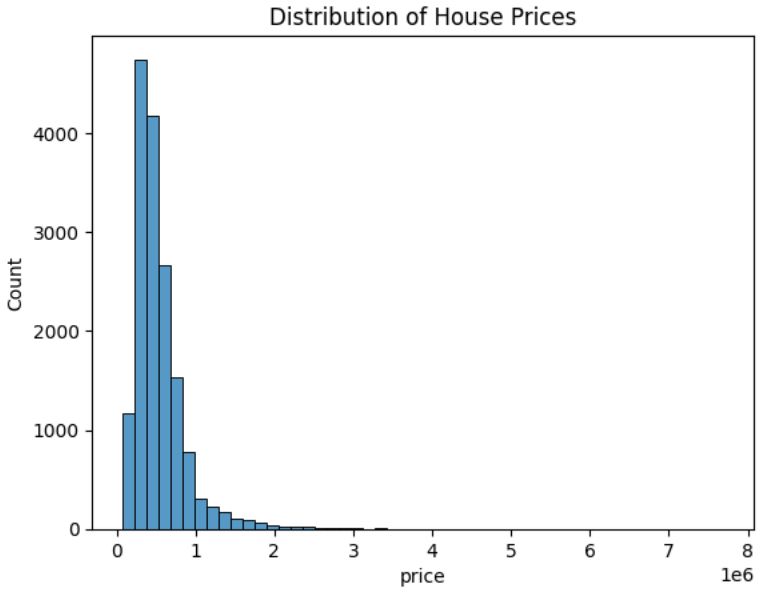
Key steps:

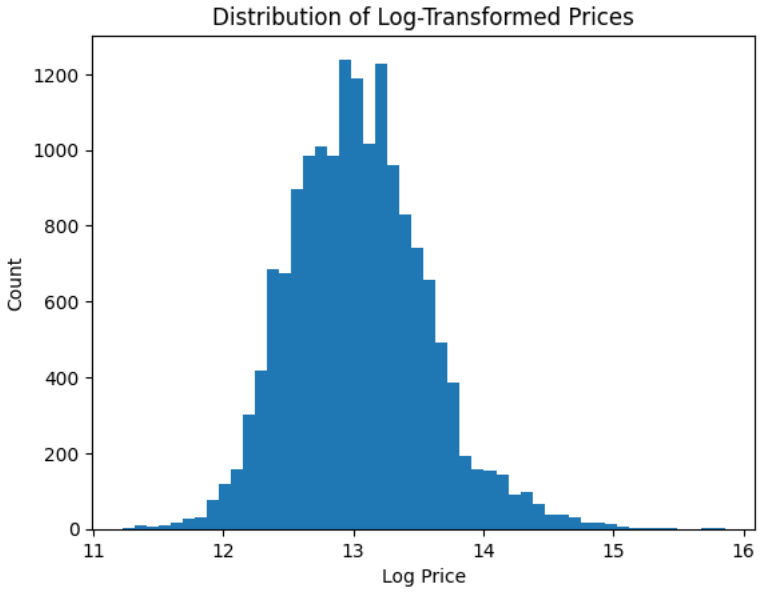
1. Acquire satellite imagery using Mapbox (static images) from property coordinates.
2. Extract 512-dim visual embeddings using a pretrained ResNet backbone.
3. Reduce visual embedding dimensionality with PCA for stability and speed.
4. Concatenate PCA-reduced image features with cleaned tabular features (feature-level fusion).
5. Train a gradient-boosted regressor (XGBoost) to predict price.
6. Add explainability: SHAP for global/tabular feature contributions and Grad-CAM to visualize what the CNN focuses on.

This approach balances accuracy, interpretability, and engineering simplicity suitable for both prototyping and demonstrative production use.

2. Exploratory Data Analysis (EDA)

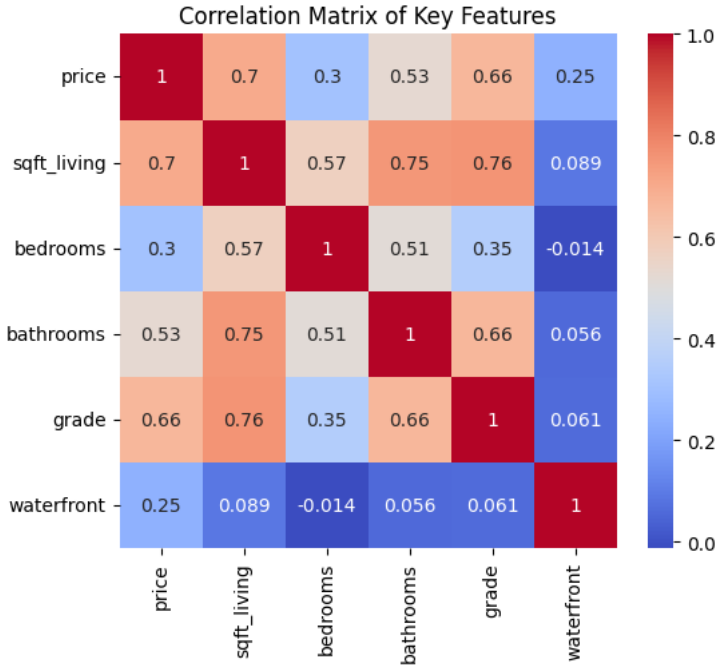
2.1 Price distribution

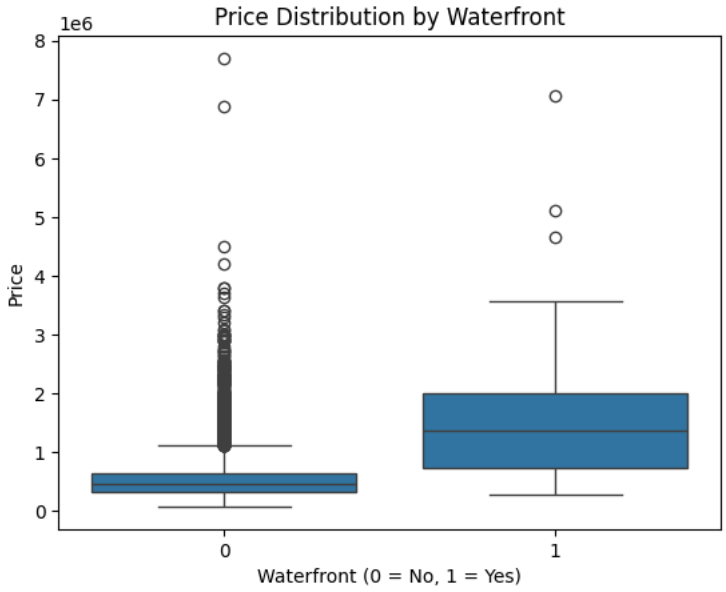


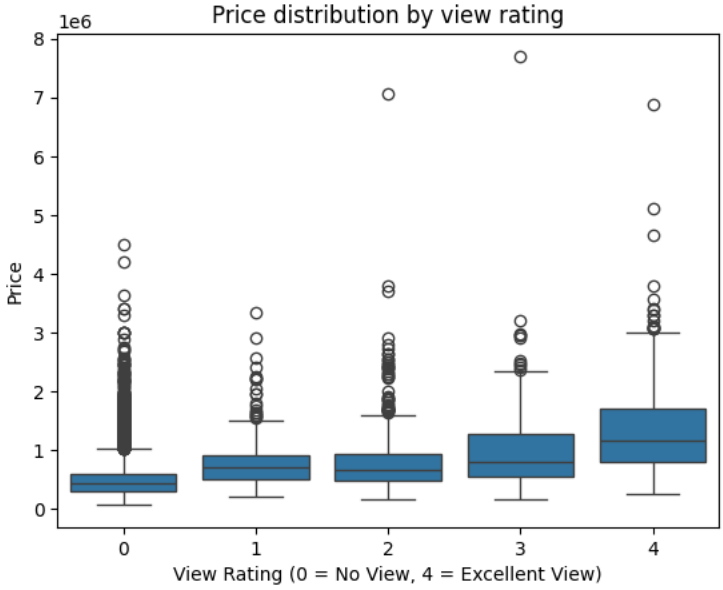


- Raw prices are strongly right-skewed (few very expensive houses).
- We applied `log1p(price)` (log(1+price)) during modeling experiments to stabilize variance and reduce the influence of extreme values; results reported both on log-scale and converted back to price where needed.

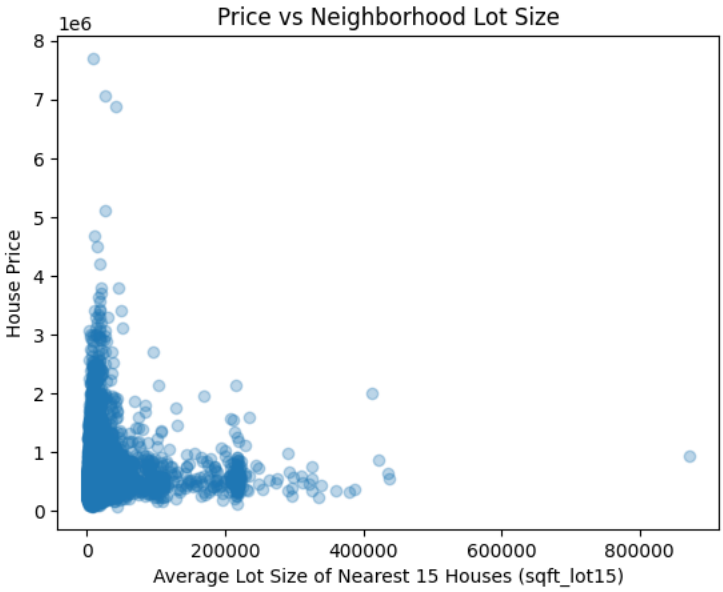
2.2 Tabular feature analysis

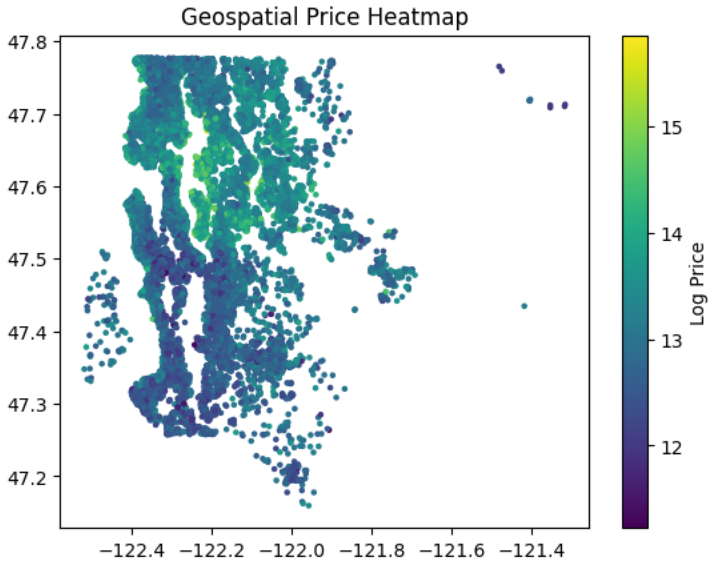



Price Distribution by Waterfront
1e6
Price
Waterfront (0 = No, 1 = Yes)


Price distribution by view rating
1e6
Price
View Rating (0 = No View, 4 = Excellent View)

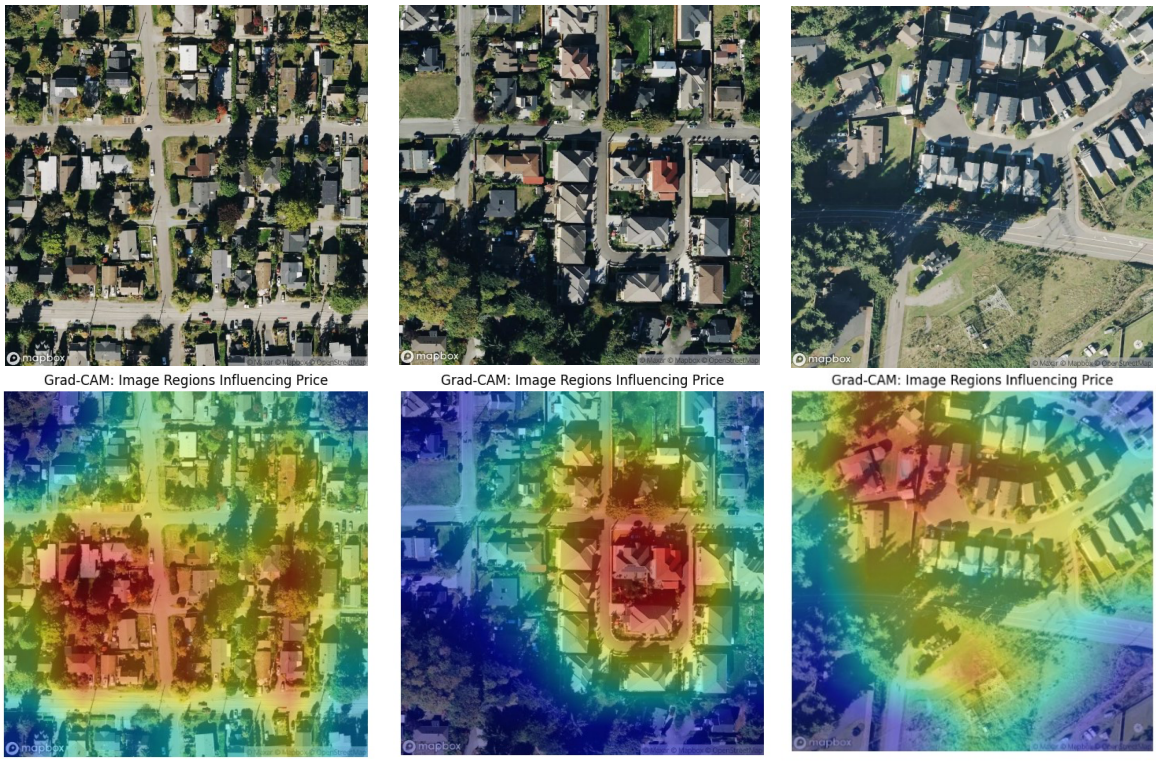

Price vs Living Area
1e6
Price
Sqft Living





- Confirmed strong positive correlation of price with: sqft_living, grade, and bathrooms.
- waterfront and view show clear step-up effects in boxplots which implies waterfront properties command premiums.
- Neighbourhood aggregations (sqft_living15, sqft_lot15) capture local density and peer-group effects.
- Geospatial scatter (longitude/latitude colored by log-price) reveals clear geographic clusters of higher-priced properties, motivating the addition of neighbourhood imagery.

2.3 Satellite image samples



- Representative satellite images are attached.
- Grad-CAM overlays show CNN attention aligns with intuitive economic drivers (green patches, waterfront, road intersections).
- PCA on embeddings shows clustering by visual neighbourhood type, an evidence that embeddings capture useful signal.

3. Financial/Visual Insights

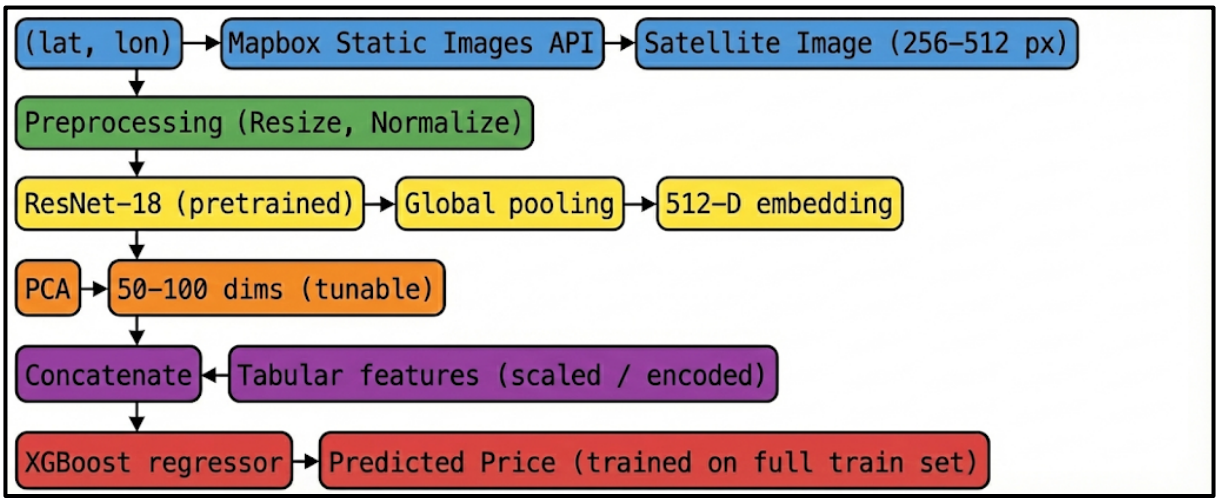
- **Green cover:** Areas with more visible vegetation (parks, tree-lined streets, larger yards) correlate with higher prices regionally; Grad-CAM highlights these regions in higher-priced samples.
- **Water proximity:** Presence of water bodies is a strong positive signal; both tabular waterfront and visual cues (coastline/lakes) reinforce this.
- **Road density & connectivity:** Well-connected road patterns can be positive (accessibility) or reveal noise (busy highways); Grad-CAM often lights up road networks, suggesting the CNN encodes transport-related cues.
- **Urban vs. suburban texture:** Dense block patterns versus spacious lots are visually distinct and map to different price bands.
- **Complementarity:** Tabular features capture building-level quality (sqft, grade); visual embeddings capture neighbourhood-level context- together they explain more variance than either alone in many cases.

Practical takeaways:

i. Visual context is especially helpful where tabular features are ambiguous (similar houses in different neighbourhoods).

ii. Even modest improvements in RMSE translate to large dollar-value improvements for high-priced properties, thus important for business impact.

4. Architecture Diagram



Overall workflow

Design rationales:

i. ResNet-18: balance between representation quality and inference speed.
ii. PCA: reduces overfitting and speeds up tree training.
iii. XGBoost: robust for tabular + high-dim features and provides feature importance; SHAP integrates well for interpretability.

5. Results

5.1 Model performance comparison

Model Type	RMSE (log)	MAE(log)	R^2
Tabular data only	0.2480	0.1338	0.7771
Tabular + Satellite images	0.1695	0.1227	0.8959

5.1 Key findings

- Satellite imagery **significantly improves predictive accuracy.**
- Visual context complements traditional housing attributes.
- Multimodal fusion captures neighbourhood quality effectively.
- Explainability tools validate meaningful visual learning.

6. Conclusion

This project demonstrates that integrating satellite imagery with tabular housing data leads to **more accurate and interpretable real estate valuation models**. The CNN successfully

extracts high-level environmental features, while XGBoost efficiently combines these with structured attributes.

By incorporating **Grad-CAM explainability**, the framework moves beyond black-box prediction and provides actionable insights into which environmental factors drive property value.