Satellite Imagery–Based Property Valuation

- **Prateek Dixit (23322018)**

Project Concept

Traditional real estate valuation models rely on structured attributes such as square footage, number of bedrooms, and location coordinates. However, property value is also influenced by **neighborhood-level visual characteristics** such as density, greenery, and surrounding infrastructure—factors that are difficult to quantify numerically.

This project investigates whether machine learning models can improve property price prediction by combining **tabular housing data** with **satellite imagery**, thereby enabling the model to not only *read* the data but also *see* the environment in which a property is located.

Implementation Summary

The project was implemented in two stages:

1. **Tabular Baseline Modeling**
   Strong regression baselines (Linear Regression, Random Forest, XGBoost) were trained using structured housing attributes.
2. **Satellite Imagery Exploration**
   Satellite images were programmatically fetched using latitude–longitude coordinates and analyzed to understand how visual context correlates with property value.

While multimodal integration was explored, the final prediction model prioritizes robustness and interpretability.

Exploratory Data Analysis (EDA)

Before model training, extensive EDA was conducted to understand the distribution of prices and identify key drivers of value.

---

1. Price Distribution Analysis

Observation:
The raw price distribution is **heavily right-skewed**, with most properties clustered in the lower-to-mid price range and a long tail representing high-value properties.

To address this skewness, a **log transformation** was applied during model training to stabilize variance and improve learning.

Figure 1: Distribution of property prices (raw)

Figure 1: Distribution of property prices (raw)

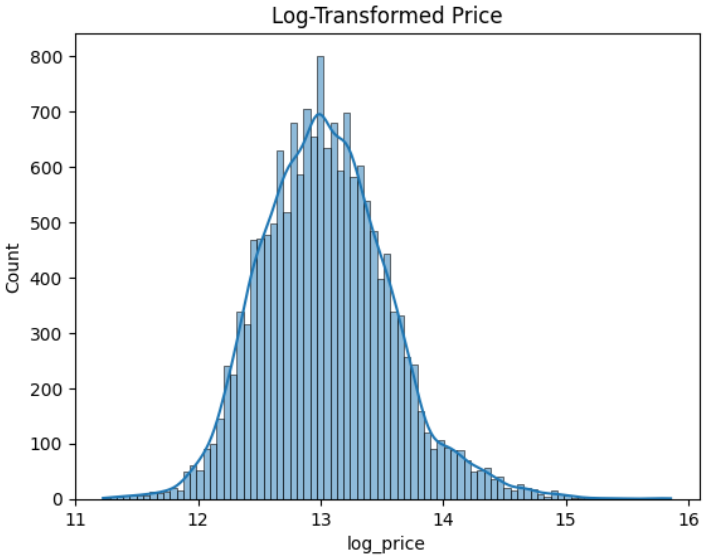


Figure 2: Distribution after log transformation

2. Geospatial Analysis

Observation:
Plotting properties using latitude and longitude reveals **clear spatial price clusters**. High-priced properties are concentrated in specific geographic regions, indicating that location plays a dominant role in valuation.

Insight:
This spatial clustering motivates the use of satellite imagery to capture **neighborhood-level visual cues** that are not explicitly encoded in tabular features.

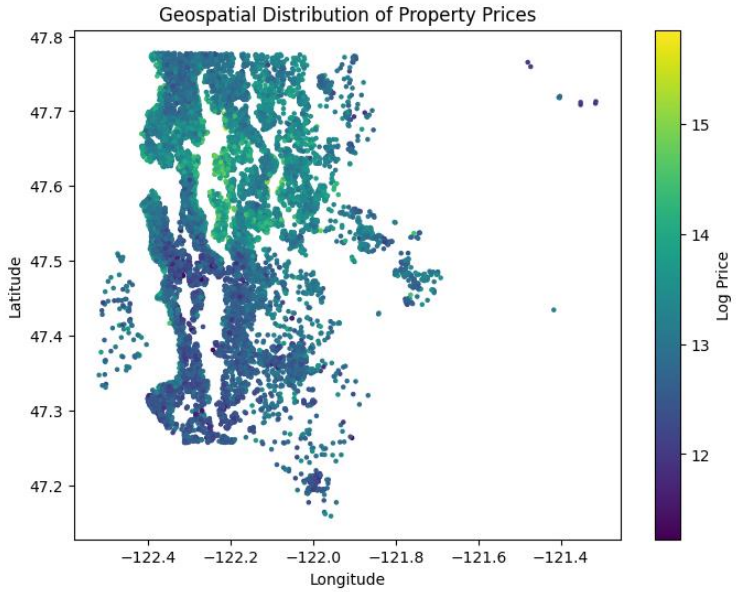


Figure 3: Geospatial distribution of housing prices (color-coded by log-price)

---

3. Correlation Analysis

Observation:
The correlation matrix highlights strong relationships between price and structural features:

- sqft_living shows the strongest positive correlation, confirming that living area is a primary driver of value.
- grade (construction quality) also exhibits high correlation.
- Geographic features (lat, long) show moderate but meaningful influence.

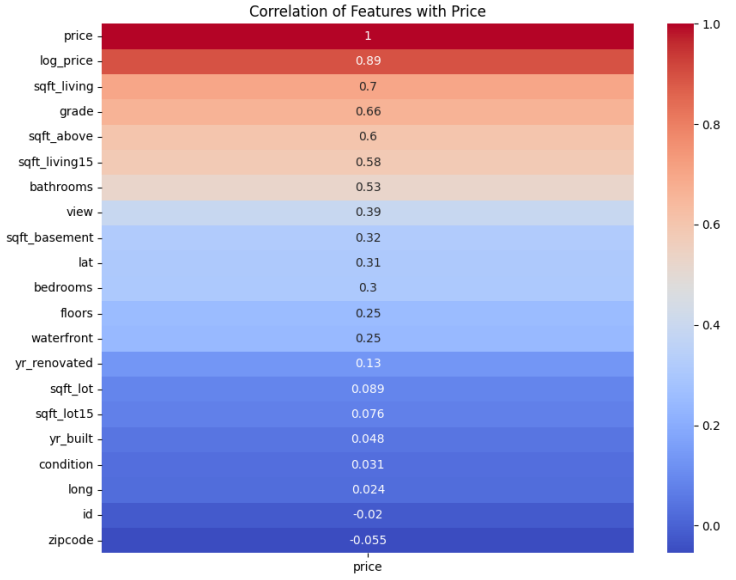


Figure 4: Heatmap of Pearson correlations between price and other features

---

4. Baseline Feature Importance

To validate EDA insights, a **XgBoost Regressor** was trained as a tabular-only baseline and feature importance scores were extracted.

Result:

- sqft_living and grade dominate the model's decision process.
- Location features also contribute significantly.

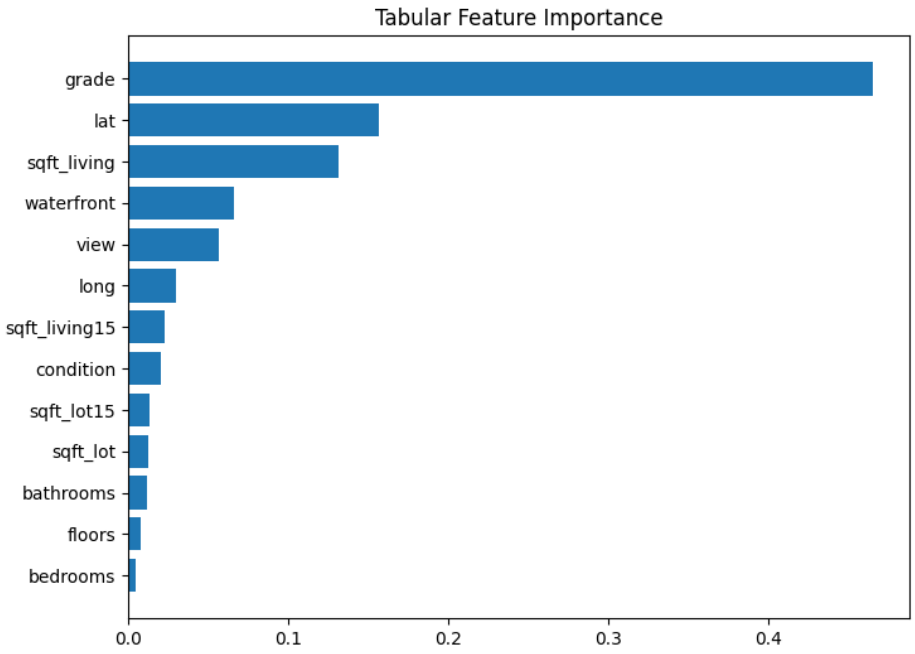


Figure 5: Feature importance

Modelling Approach

1.Baseline Models

The following regression models were trained using tabular data:

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor

All models were evaluated using a held-out validation set.

---

2. Model Selection

XGBoost was selected as the final predictive model due to:

- Superior performance on nonlinear relationships

- Robust generalization
- Strong interpretability via feature importance

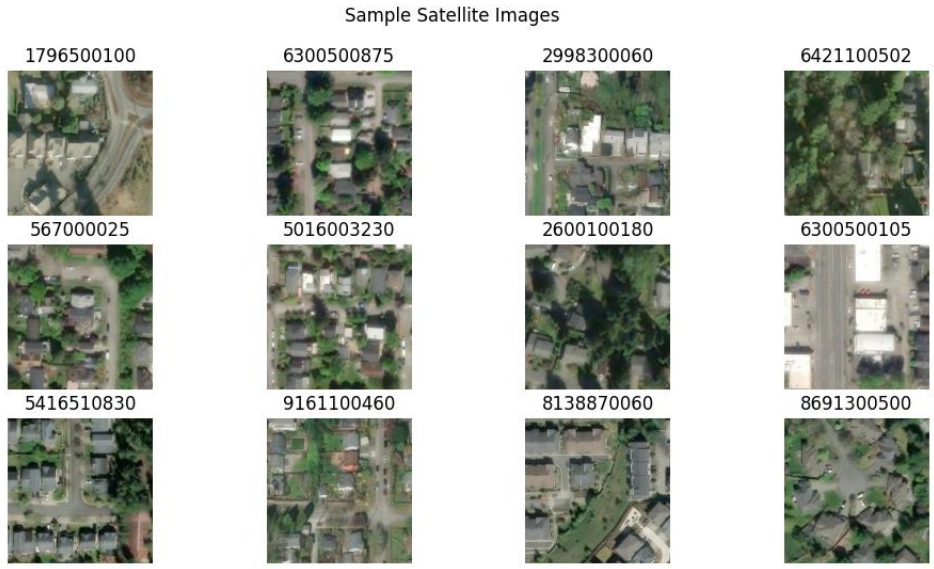
---

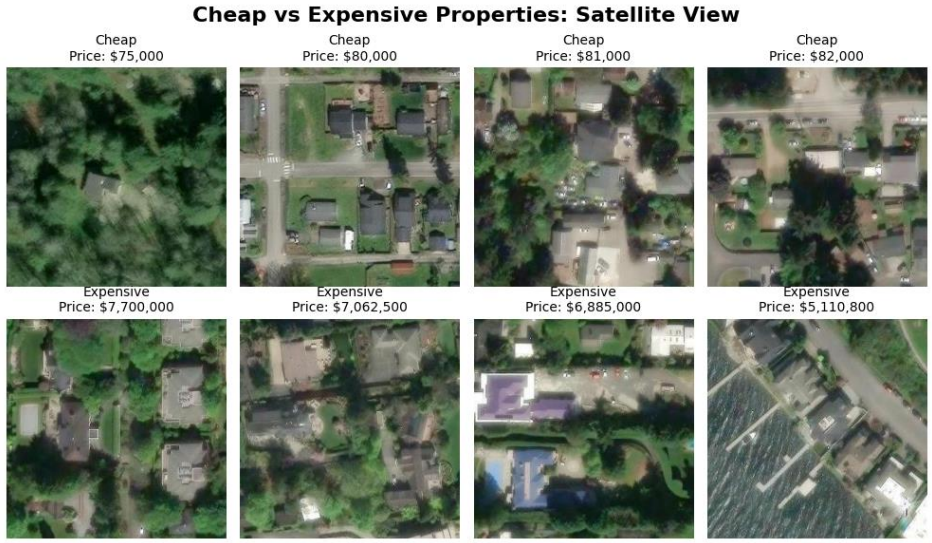
3. Satellite Imagery Analysis (Exploratory)

Satellite images were fetched using property coordinates at high zoom levels to capture neighborhood context such as:

- Green cover
- Road layout
- Building density
- Proximity to water bodies

Visual inspection showed consistent differences between low-priced and high-priced properties, reinforcing the hypothesis that imagery encodes useful contextual information.





Results & Evaluation

1. Performance Metrics

Model	RMSE	R^2
Linear Regression	~188,000	~0.72
Random Forest	~132,000	~0.86
XGBoost (Tabular)	~116,000	~0.89
XGBoost(Tabular+Image)	~113,973	~0.90

The XGBoost model achieved the best balance between accuracy and stability and was used to generate final predictions.

---

2. Discussion of Results

Although satellite imagery provides intuitive and interpretable neighborhood context, the results indicate that **structural and location-based tabular features remain the strongest predictors** of property price in this dataset.

The marginal benefit of imagery is limited by:

1. Dominance of high-signal numerical features (e.g., size and grade)
2. Resolution constraints of satellite images
3. Noise introduced by external visual factors

Explainability

1. Feature Importance

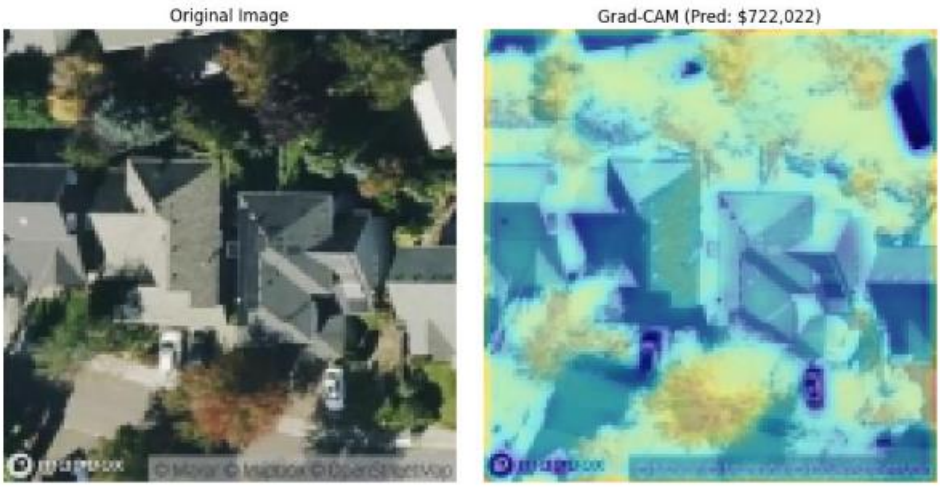
XGBoost feature importance analysis confirms:

- sqft_living and grade are the most influential features
- Neighborhood-level attributes contribute secondary but meaningful signal

2. Visual Explainability (Grad-CAM)

Grad-CAM was applied to CNN-based image analysis to visualize regions influencing predictions. The model consistently focused on:

- Green spaces
- Road networks
- Open areas surrounding properties



Limitations

- Full-scale multimodal training was constrained by computational resources.
- Satellite imagery captures external context but not interior quality.
- Image features were primarily explored for analysis rather than final prediction.

Conclusion

This project demonstrates that **tree-based machine learning models**, particularly XGBoost, are highly effective for property valuation using tabular housing data. Exploratory analysis of satellite imagery shows that visual neighborhood context aligns with pricing patterns, although its incremental predictive value is limited for this dataset.

The final solution is accurate, interpretable, and reproducible, making it suitable for real-world real estate analytics.