Satellite Imagery Based Property Valuation

Final Project Report

Name: Bhoomi Kaushik

Enrollment No:24118016

---

1. Introduction

Accurate property valuation is a fundamental problem in real estate analytics, finance, and urban planning. Traditional house price prediction models rely heavily on structured tabular data such as the number of bedrooms, bathrooms, square footage, construction quality, and geographic location. While these features are highly informative, they fail to capture **environmental and neighborhood context**, which plays a significant role in determining property value.

Satellite imagery provides a rich source of spatial information, capturing features such as greenery, road networks, urban density, and proximity to water bodies. This project proposes a **multimodal regression framework** that integrates **tabular housing attributes** with **satellite imagery** to predict property prices more effectively.

The primary objective of this project is to:

- Build a strong baseline model using tabular data
- Incorporate satellite imagery using deep learning
- Evaluate whether visual context improves predictive performance

---

2. Dataset Description

2.1 Tabular Dataset

The tabular dataset consists of residential housing data with the following key features:

- Structural features: bedrooms, bathrooms, floors
- Size-related features: sqft_living, sqft_lot, sqft_living15, sqft_lot15
- Quality indicators: grade, condition, view
- Binary indicators: waterfront
- Temporal features: yr_built, yr_renovated
- Spatial features: latitude, longitude, zipcode

The training dataset contains the target variable **price**, while the test dataset does not.

2.2 Satellite Imagery Dataset

Satellite images were fetched using **latitude and longitude coordinates** via the **Mapbox Static Images API**. Each property location was used to download a **400×400 pixel satellite image at zoom level 18**, capturing neighborhood-level spatial context.

> **API Usage Note:**
> The Mapbox access token was stored securely during execution and is intentionally **not included** in this report or repository to follow security best practices.

---

3. Exploratory Data Analysis (EDA)

3.1 Price Distribution and Skewness

The target variable `price` exhibited strong **right skewness**, indicating that while most houses fall within a moderate price range, a small number of luxury properties have extremely high prices.

To stabilize variance and reduce the influence of outliers, a logarithmic transformation was applied:

log_price=log(1+price)\text{log_price} = \log(1 + \text{price})log_price=log(1+price)

This transformation significantly improved model training stability and evaluation metrics.

3.2 Feature Relationships

EDA revealed strong positive relationships between house price and:

- Living area
- Construction grade
- Waterfront and view indicators

These relationships were largely nonlinear, motivating the use of tree-based models.

3.3 Geospatial Analysis

Latitude and longitude demonstrated clear spatial price patterns, confirming that **location is one of the strongest predictors of house prices**.

---

4. Feature Engineering

Domain-driven feature engineering was applied to improve predictive power:

- **House age:** year_sold − yr_built
- **Renovation indicator:** binary flag for renovated properties
- **Relative house size:** living area compared to neighborhood average
- **Lot utilization ratio:** living area divided by lot size
- **Bathroom-to-bedroom ratio:** luxury indicator

Redundant and highly correlated features were removed to improve model stability.

---

5. Baseline Regression Models

Baseline models trained on tabular data included:

- Linear Regression
- Ridge Regression

- Lasso Regression

These models established a reference performance but were limited in capturing nonlinear relationships.

---

6. XGBoost on Tabular Data

XGBoost was used due to its ability to:

- Capture nonlinear interactions
- Handle heterogeneous features
- Provide strong performance on tabular datasets

Tabular XGBoost Results

- **RMSE (log scale): 0.159**
- **R^2: 0.90**

This model served as the strongest baseline.

---

7. Satellite Image Feature Extraction

A pretrained **ResNet50** model was used as a feature extractor via transfer learning. The final classification layer was removed, and **2048-dimensional image embeddings** were extracted for each satellite image.

To reduce dimensionality and noise, **Principal Component Analysis (PCA)** was applied, reducing the feature space to **128 dimensions**.

---

8. Multimodal Fusion

An **early fusion strategy** was adopted, where engineered tabular features were concatenated with PCA-reduced image features and passed to an XGBoost regressor.

---

9. Results and Comparison

Model	RMSE (log)	R^2
Linear Regression	0.242	0.769
Ridge Regression	0.242	0.769
Lasso Regression	0.242	0.768
XGBoost (Tabular)	**0.159**	**0.90**
XGBoost (Multimodal)	0.218	0.79

While the multimodal model incorporated richer contextual information, it did not outperform the tabular XGBoost model due to the already strong predictive power of structured features and limited image sample size.

---

10. Conclusion

This project demonstrates a complete end-to-end machine learning pipeline for property valuation using both tabular data and satellite imagery. The results show that **high-quality tabular features remain the dominant predictors**, while satellite imagery provides complementary contextual understanding.

The study highlights an important insight: **multimodal learning is most effective when the additional modality contributes strong, clean signal**.

---

11. Limitations and Future Work

- Limited number of satellite images
- CNN model was not fine-tuned on real estate imagery
- Use of single aerial viewpoint

Future improvements include:

- Fine-tuning CNNs on domain-specific images
- Higher-resolution imagery

- Street-level view integration
- Attention-based fusion techniques

---

12. References

1. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System*, 2016
2. He et al., *Deep Residual Learning for Image Recognition*, 2016
3. Selvaraju et al., *Grad-CAM*, 2017