Satellite Imagery Based Property Valuation

- CDC X Yhills OPEN PROJECTS 2025-2026

Anurag Sain , Indian Institute of Technology Roorkee

Introduction

Property valuation is a critical problem in real estate analytics, traditionally addressed using structured attributes such as house size, number of rooms, and location-based statistics. While effective, these approaches fail to capture environmental and neighborhood-level visual context, such as greenery, water proximity, and urban density.

This project develops a multimodal regression pipeline that integrates tabular housing features with satellite imagery extracted using geographic coordinates. By combining numerical and visual information, the model aims to improve price prediction accuracy and interpretability.

To maintain clarity, the system design is presented using multiple diagrams, each focusing on a specific stage of the pipeline.

Dataset Overview

The dataset consists of residential housing records containing both structural and geographic attributes.

Tabular Features

- Structural: bedrooms, bathrooms, floors
- Size-based: sqft_living, sqft_lot, sqft_above, sqft_basement
- Neighborhood context: sqft_living15, sqft_lot15
- Quality indicators: condition, grade
- Location indicators: view, waterfront
- Geographic coordinates: lat, long

The target variable is price, which shows strong right skewness. To stabilize variance and improve learning performance, the target variable was transformed using log1p(price)

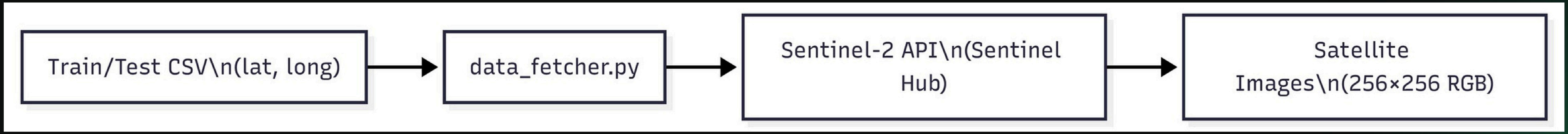
Satellite Image Collection

Satellite imagery was programmatically

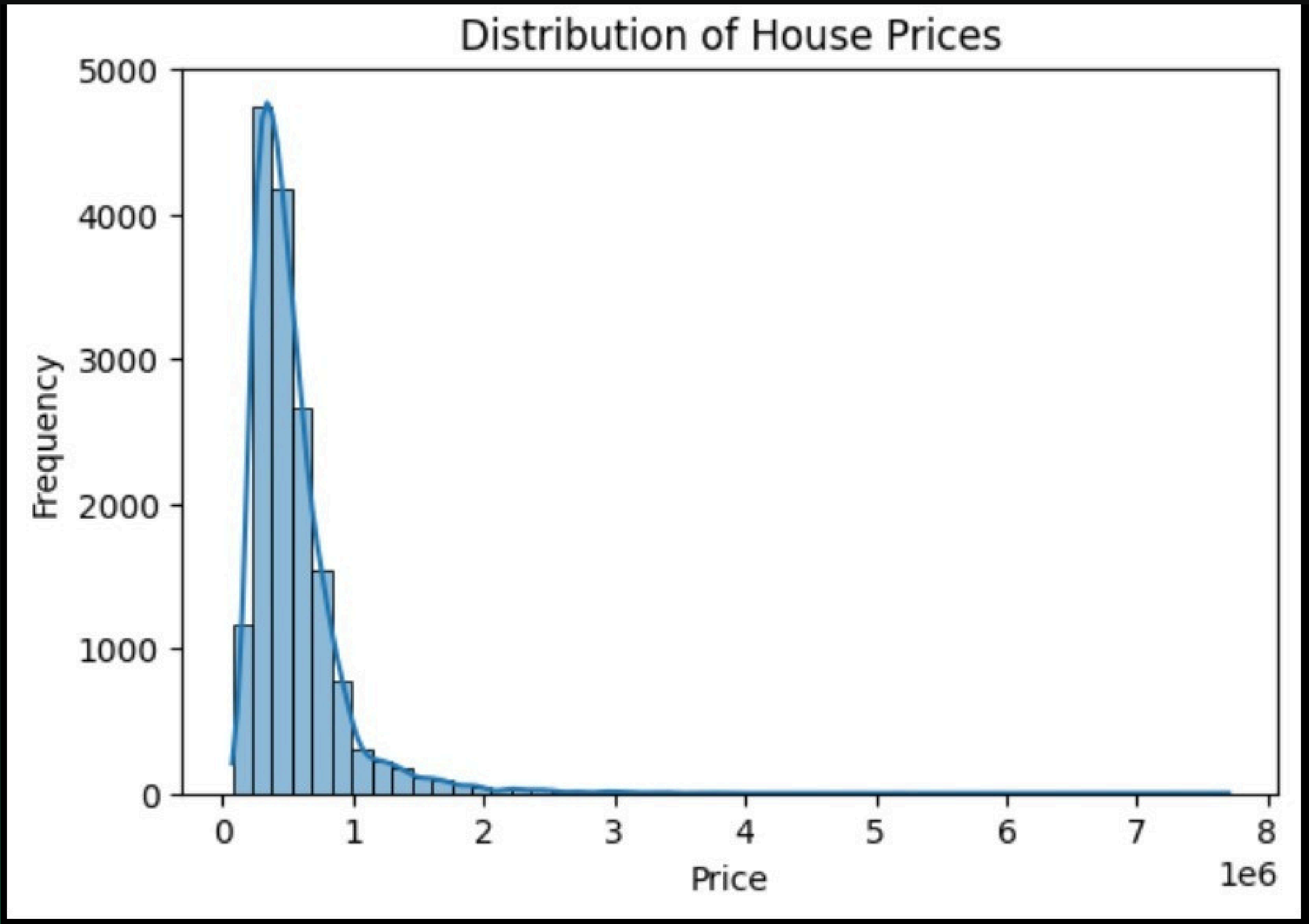
downloaded usingSentinelHub’s Sentinel-2 L2A dataset, based on latitude and longitude values associated with each property.

Description:

- Trainingand test CSV files provide latitude and
- longitude data_fetcher.py queries Sentinel Hub’s
- API RGB satellite images (256×256) are stored locally The process is fully automated and reproducible This ensures scalable and consistent visual data generation without manual intervention.



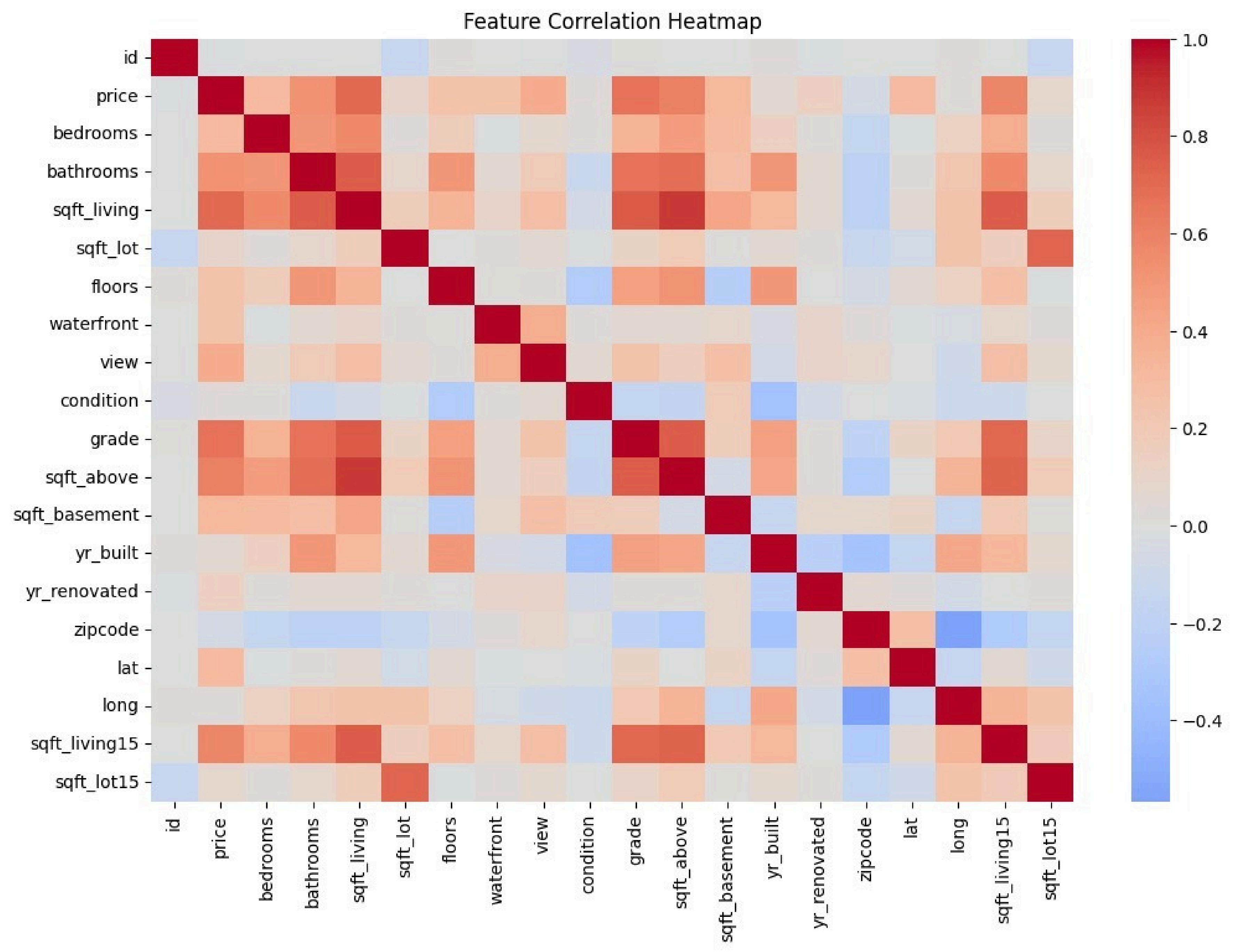
Train/Test CSV\n(lat, long) → data_fetcher.py → Sentinel-2 API\n(Sentinel Hub) → Satellite Images\n(256×256 RGB)



Price Distribution Analysis

Interpretation

- The distribution is heavily right-skewed
- Most properties fall in lower-to-mid price ranges
- A small number of high-value properties form long tails
- This validates the need for log transformation of the target variable.



Feature Correlation Heatmap

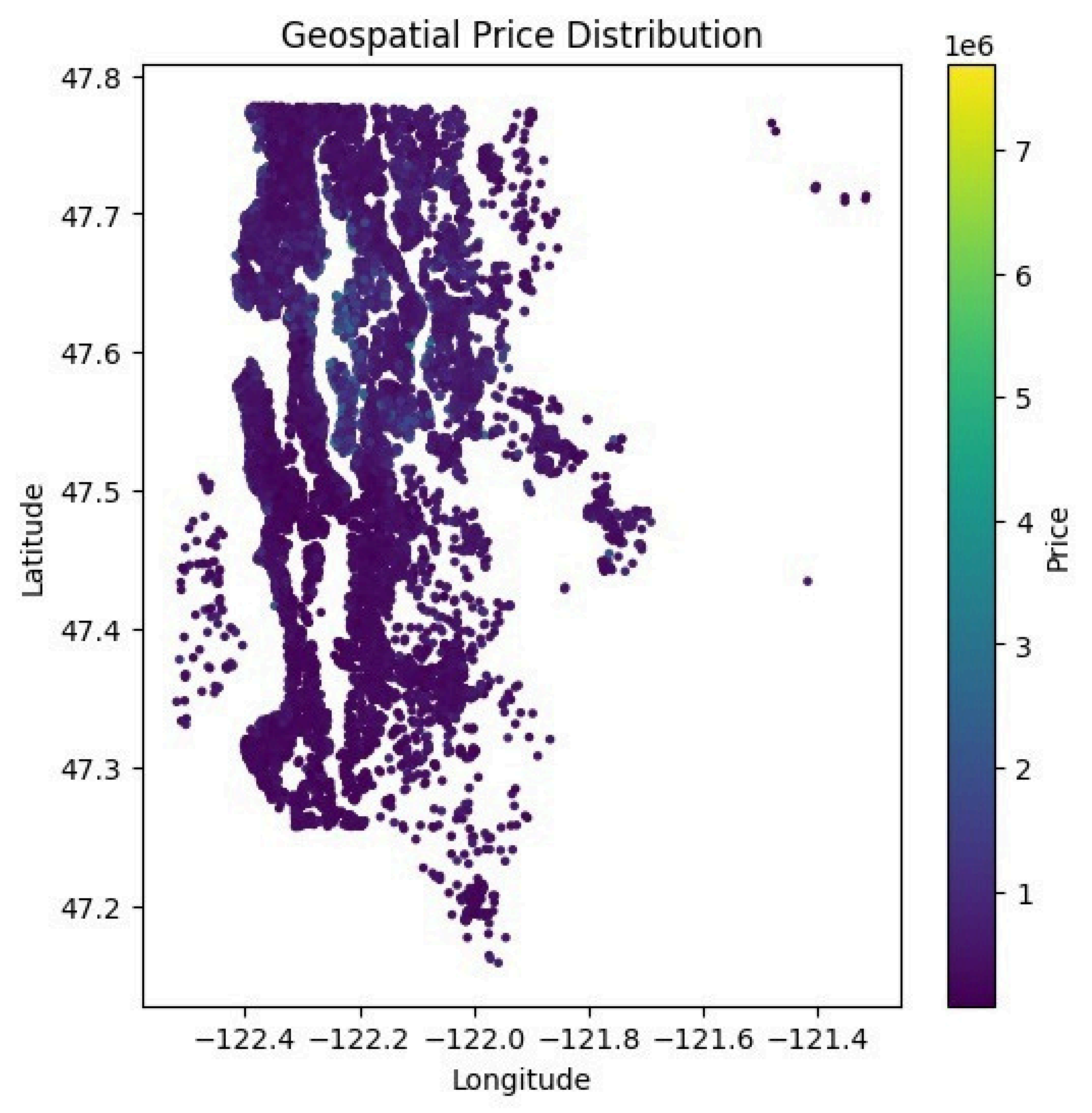
Key Observations:

- Strong positive correlation between price and:
- sqft_living, grade, bathrooms
- Moderate correlation with view and waterfront Weak correlation with condition and year_built
- This confirms that size and construction quality are dominant price drivers.

Geospatial distribution of property prices.

Interpretation:

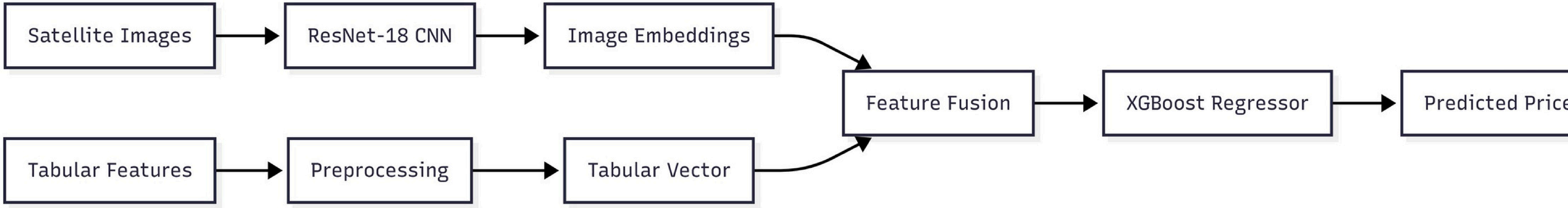
- High-value properties cluster in specific geographic regions
- Coastal and waterfront areas exhibit higher prices Inland and dense urban regions show lower valuations
- This motivates the inclusion of satellite imagery to capture spatial context.



Model Architecture

Explanation:

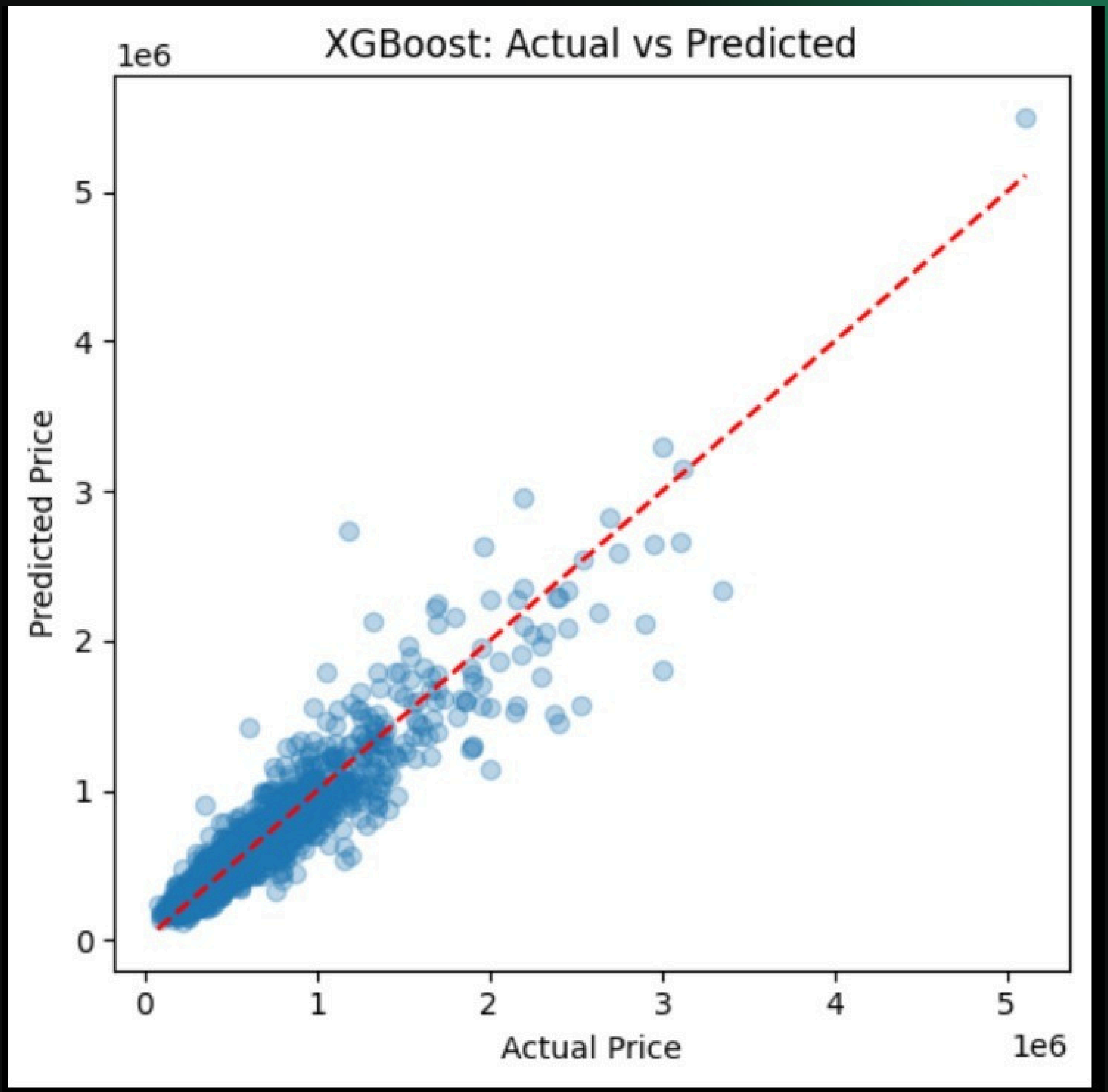
- The modeling pipeline consists of two parallel
- components:
- Tabular Branch
- Data cleaning and scaling
- Structured numerical feature vector
- Image Branch
- Pretrained ResNet-18 CNN (ImageNet weights)
- Final classification layer removed Extraction of 512-dimensional image embeddings
- The two representations are concatenated and passed to an XGBoost regressor for final price prediction.



Actual vs Predicted Prices

Interpretation:

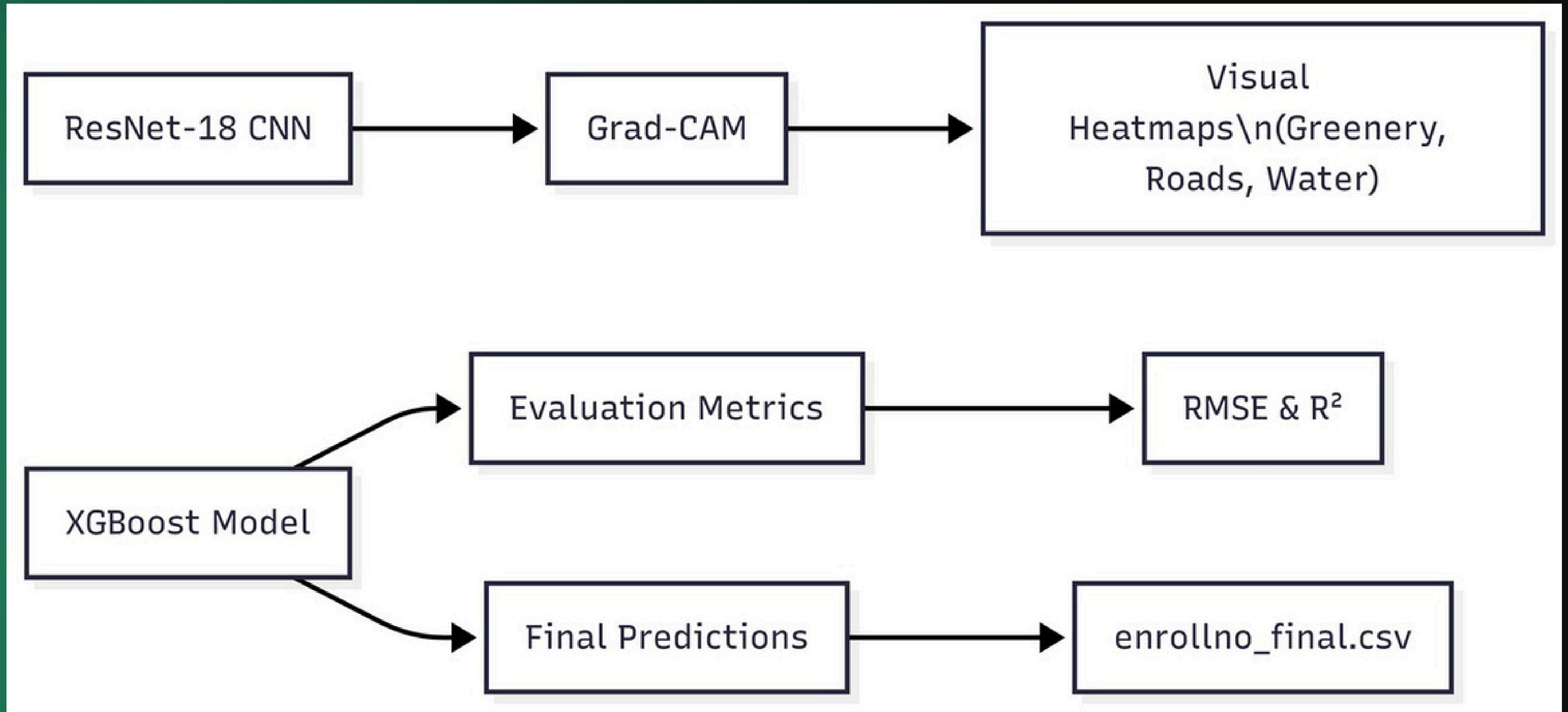
- Points closely follow the diagonal reference line
- Indicates strong agreement between predictions and ground truth
- Slight underestimation observed for very high-priced properties
- Overall performance shows good generalization



Visual Explainability

Explanation:

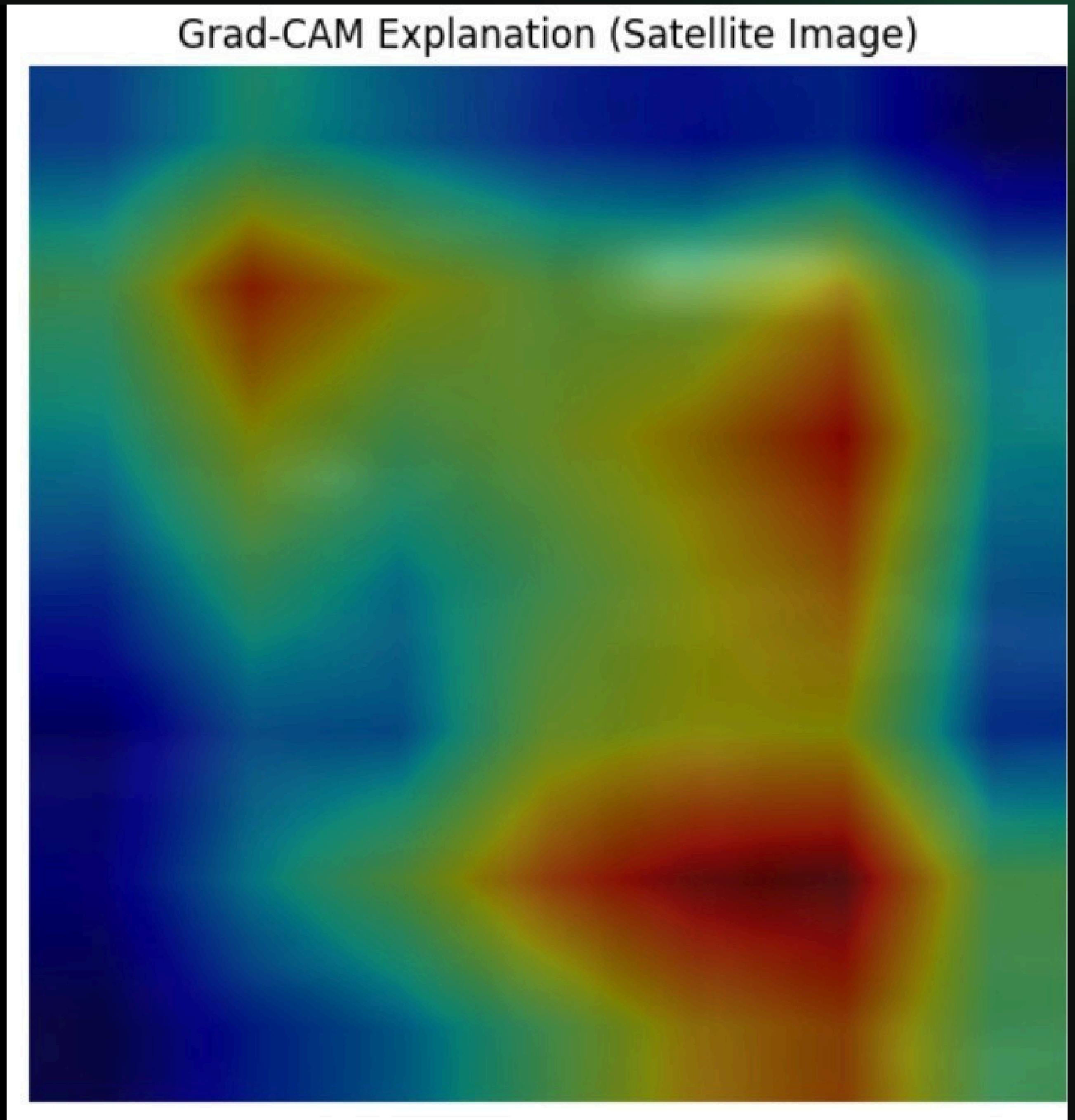
- Grad-CAM was appliedtotheCNNtoidentify
- spatialregions contributing topredictions.
- Observed high-importance regionsinclude: Vegetation and greenery Water bodies Road networks and surrounding urban layout
- This confirms the model learns meaningful environmental features, improving transparency and trust.



ResNet-18 CNN → Grad-CAM → Visual Heatmaps\n(Greenery, Roads, Water)

XGBoost Model → Evaluation Metrics → RMSE & R²

XGBoost Model → Final Predictions → enrollno_final.csv



Grad-CAM Explanation (Satellite Image)

Model Performance Comparison

Model	RMSE	R^2
Random Forest (Tabular Only)	**126,369.70**	**0.873**
XGBoost (Multimodal)	**117,588.97**	**0.890**

Key Findings:

- 7% reduction in RMSE
- Improved variance explanation
- Satellite imagery adds measurable predictive value

Prediction Output

- Format (strict): id, predicted_price
- All predictions were inverse-transformed to the original price scale.
- Predictions were generated on the unseen test dataset and saved as: 23115018_final.csv

Conclusion & Future Work

Conclusion

This project demonstrates that integrating satellite imagery with structured housing data significantly improves property valuation performance. The multimodal approach captures environmental and neighborhood characteristics that are otherwise ignored in traditional models.

Future Improvements:

- Fine-tuning CNN on satellite-specific datasets
- Higher resolution imagery
- Attention-based multimodal fusion
- Temporal satellite analysis

Thank You

Email anurag_s@ee.iitr.ac.in