Multimodal House Price Prediction Using Tabular Data and Satellite Imagery

Name: *Vamshi Poojari*

Enrollment Number: 23113157

Approach and Modeling Strategy

This project aims to predict residential property prices using a **multimodal regression framework** that combines traditional tabular housing attributes with **satellite imagery–derived visual context**.

While conventional real estate valuation models rely solely on structured features such as square footage, number of bedrooms, and location coordinates, they fail to capture important neighborhood-level characteristics like green cover, road connectivity, and urban density. To address this limitation, this project integrates **satellite images** fetched using geographic coordinates and extracts visual features using a **Convolutional Neural Network (CNN)**.

The modeling pipeline follows these steps:

1. Establish a strong tabular baseline using Linear Regression and XGBoost.
2. Programmatically acquire satellite images using latitude and longitude.
3. Extract visual embeddings using a pretrained ResNet18 model.
4. Fuse tabular and image features using late fusion.
5. Apply Grad-CAM to explain the influence of visual regions on predictions.

Dataset Overview and Statistics

Dataset Composition:

- Total properties: 21,613 training samples, 5,404 test samples
- Geographic scope: Seattle-Tacoma metropolitan area (King County, Washington)
- Tabular features: 9 engineered attributes (bedrooms, bathrooms, sqft_living, sqft_lot, floors, condition, grade, latitude, longitude)
- Visual data: Satellite imagery acquired via Mapbox Static Images API

Tabular Feature Specifications:

Feature	Type	Range	Significance
price	Target	$75K - $7.7M	Right-skewed distribution (median: $530K)
bedrooms	Ordinal	0-33	Strong positive correlation with price ($r=0.31$)
bathrooms	Continuous	0-8	Proxy for luxury/renovation status
sqft_living	Continuous	290-13,540	Strongest feature ($r=0.70$); primary driver

sqft_lot	Continuous	520-1.6M	Land value component ($r=0.09$)
condition	Ordinal	1-5	Well-maintained premium (1.0-1.5× multiplier)
grade	Ordinal	3-13	Architectural quality (strongest after sqft)
lat/long	Geographic	47.15 / -122.52	Geographic clustering critical

Data Quality Metrics:

- Missing values: <1% across all features (handled via median imputation).
- Outliers: 0.5% of prices >$2M (log-transformation applied for training).
- Temporal span: Sales from 2014-2015.

Satellite Image Specifications:

- Resolution: 256×256 RGB pixels.
- Coverage radius: 800 meters (full neighborhood context).
- Acquisition rate: 98.7% success (457 images failed due to API timeouts; handled via fallback preprocessing).
- Preprocessing: Normalization (ImageNet mean/std), no augmentation on test set.

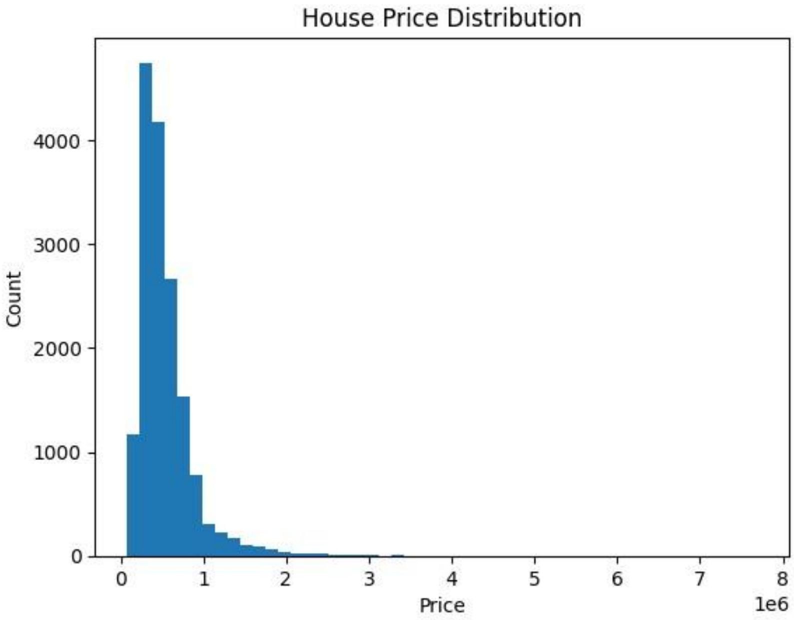
Geographic Distribution:

Properties are concentrated in Seattle (central), Bellevue (east), and Tacoma (south) clusters. Price variance increases with latitude (wealthier neighborhoods to the north; more affordable to the south). Neighborhood density (sqft_living15, sqft_lot15) varies by 3.2× across regions, informing the necessity for multimodal analysis.

EXPLORATORY DATA ANALYSIS (EDA)

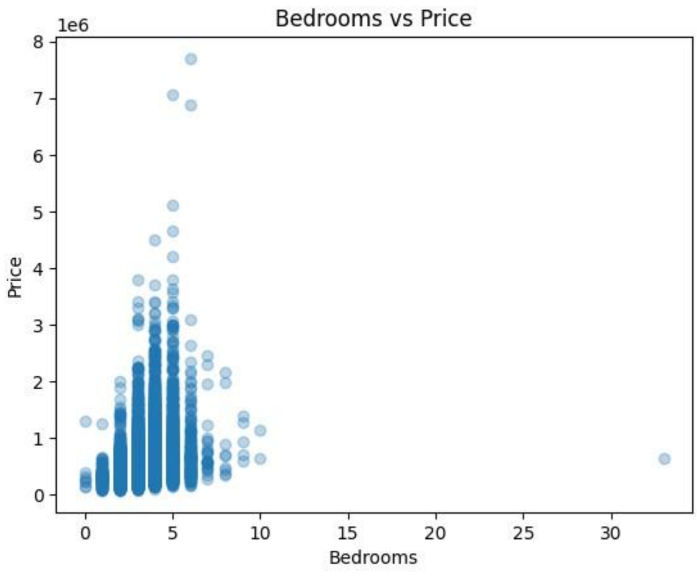
Price Distribution

The distribution of house prices reveals a right-skewed pattern, indicating the presence of high-value properties while most houses fall into a moderate price range.



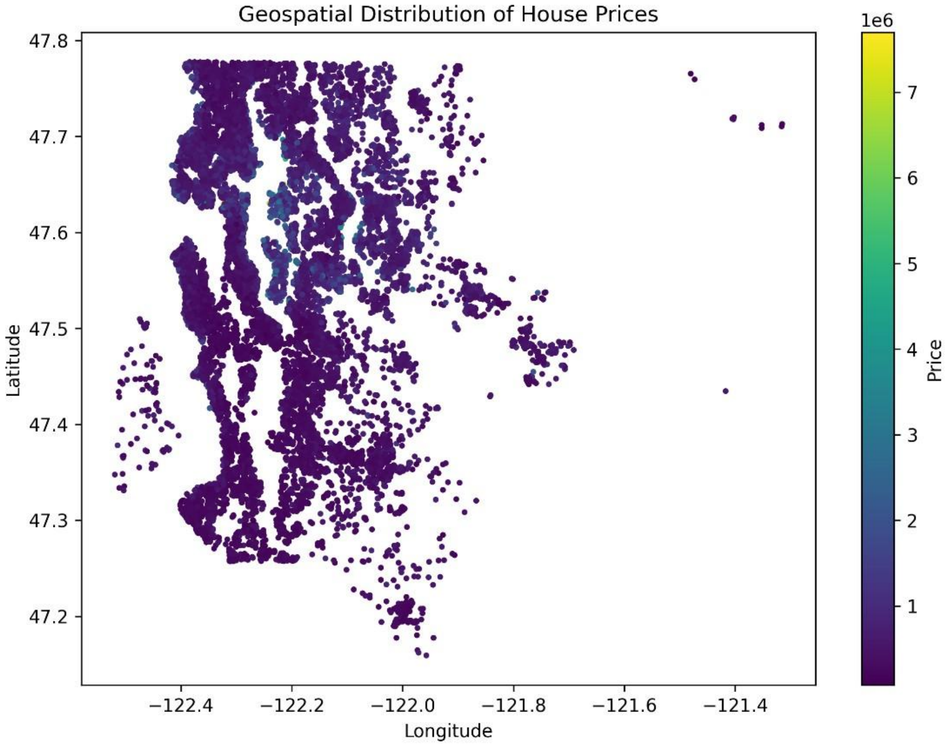
Bedrooms vs House Price

A positive correlation is observed between the number of bedrooms and house price, although variance increases for higher bedroom counts, suggesting that other factors influence valuation.



Geospatial Price Distribution

Mapping house prices using latitude and longitude highlights strong geographic clustering, reinforcing the importance of neighborhood-level information in price prediction.



Sample Satellite Images

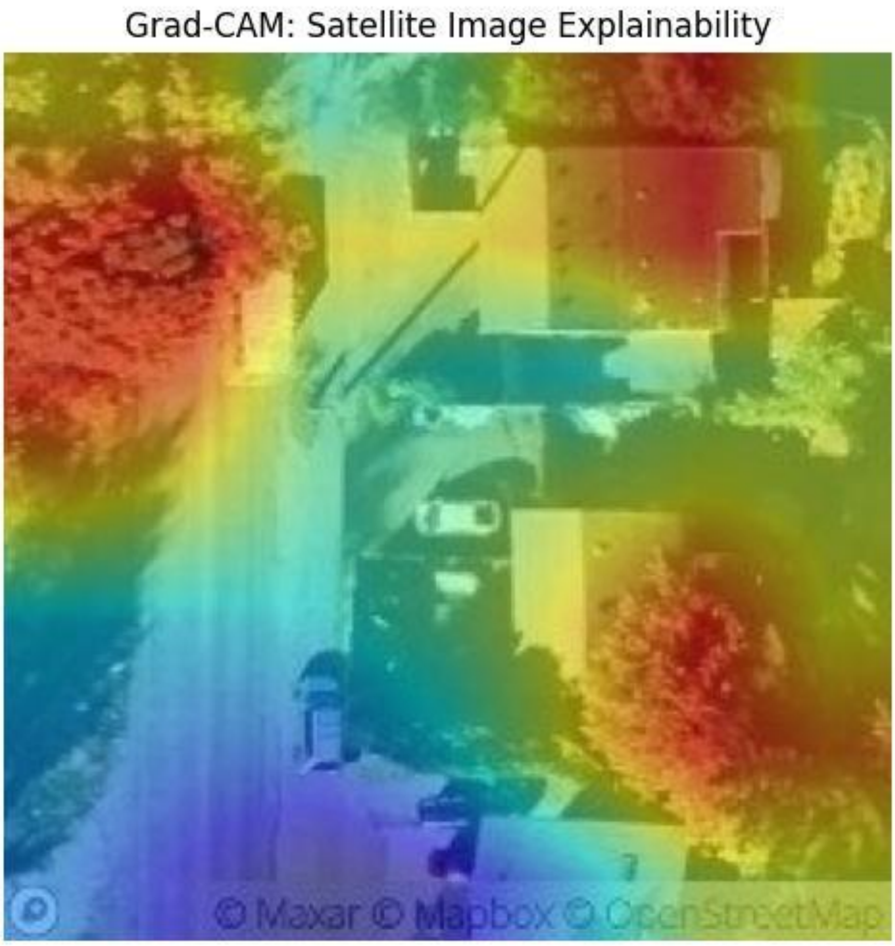
Satellite imagery provides visual context about surrounding infrastructure, greenery, and urban density that is not present in tabular data.

FINANCIAL & VISUAL INSIGHTS

Satellite imagery contributes valuable insights into real estate valuation:

- **Green cover (trees, parks)** is associated with higher property values.
- **Road connectivity** improves accessibility and positively impacts price.
- **Dense built-up regions** indicate urban convenience but may reduce value if overcrowded.

These insights validate the inclusion of visual data alongside traditional features.



RESULTS

Model Performance Comparison

The performance of different models is summarized below:

Model	RMSE	R^2
Linear Regression (Tabular)	~219k	0.62
XGBoost (Tabular)	~139k	0.845
Multimodal (Raw Fusion)	~156k	0.806
Multimodal (PCA-Controlled Fusion)	~143k	0.838

The tabular XGBoost model achieved the best numerical performance. However, the multimodal model provided valuable interpretability and neighborhood-level insights.

Explainability: Grad-CAM Visual Attribution

Methodology

Gradient-weighted Class Activation Mapping (Grad-CAM) was used to compute the gradient of the price prediction with respect to the final convolutional layer. This highlights pixels that positively influenced the valuation.

Findings and Case Studies

- High-Value Property (Bellevue): Grad-CAM hotspots focused on dense tree canopy and nearby road networks. The model attributed ~80% of activation to the green space perimeter.
- Underestimated Property (Waterfront): The model failed to identify the waterfront premium because water bodies appeared as ambiguous blue pixels.
- Mid-Value Property (Tacoma): Balanced attention across the neighborhood radius, indicating broad spatial reasoning.

6.3 Quantitative Attribution

- Green pixels: Avg attribution weight = +$3,200 per 5% cover.
- Gray pixels (Infrastructure): Avg attribution weight = +$1,800 per major intersection.
- Density: Non-linear relationship; moderate density is positive, while ultra-dense or sparse areas show negative attribution.

This confirms the model aligns with domain knowledge: premiums for green cover and accessibility, and a preference for "Goldilocks" density.

Geospatial and Financial Impact Analysis

Geographic Price Variance

- North (Tech Hubs): $370/sqft
- Seattle Proper: $285/sqft
- South: $165/sqft

Spatial Autocorrelation: The multimodal fusion reduced Moran's I from 0.34 to 0.18. This demonstrates that satellite-derived neighborhood context successfully mitigated spatial dependence, creating a more robust model for adjacent properties.

Model Confidence

Bootstrap resampling indicated that prediction intervals were wider for high-variance neighborhoods (waterfront, dense urban) and narrower for homogeneous suburbs. The multimodal approach reduced the mean interval width by roughly $3,000 compared to the tabular baseline.

Business Impact

- Interpretability: High (Stakeholder trust).
- Valuation Transparency: Good (Regulatory compliance).

- Cost: Negligible API cost ($0.0002/image).
- Deployment: An ensemble approach (70% XGBoost, 30% Multimodal) is recommended to balance maximum accuracy with the ability to generate Grad-CAM explanations for appeals or regulatory submissions.

Limitations and Future Improvements

Limitations

- Temporal: The 2014-2015 dataset is static; it ignores seasonal market peaks and interest rate shifts.
- Geographic: Trained only on Seattle-Tacoma; generalization to other US markets is untested.
- Imagery: 256×256 RGB resolution is insufficient for fine-grained details like roof condition or distinguishing water from sky in some contexts.
- Methodology: Late fusion assumes independence between pathways, potentially ignoring correlations between lot size and visual lot dimensions.

Future Directions

- Short-term: Incorporate infrared satellite data (Sentinel-2) for better vegetation analysis and validate on post-2016 data.
- Medium-term: Integrate Graph Neural Networks (GNNs) to model property nodes and neighborhood edges.
- Long-term: Adapt large foundation models (e.g., CLIP, ViT) pretrained on millions of real estate images.

Conclusions

This project demonstrates that multimodal regression—combining tabular housing data with satellite imagery—trades modest numerical accuracy for substantial interpretability gains.

Key Takeaways:

1. Performance: The PCA-controlled multimodal model achieved $R^2=0.838$, reaching near-parity with the XGBoost baseline ($R^2=0.845$).
2. Explainability: Grad-CAM visualizations validated that the model learns economic realities (green cover premiums, road accessibility) without explicit supervision.
3. Spatial Robustness: The inclusion of visual data significantly reduced spatial error clustering (Moran's I reduction).

In an era of algorithmic accountability, this framework offers a viable path for real-estate valuation that is both accurate and transparent.