
Multimodal Real Estate Price Prediction

Project Report

1. OVERVIEW

1.1 Problem Statement

Traditional real estate valuation relies solely on structured property data (bedrooms, square footage, grade) but ignores visual environmental context like green spaces, neighborhood density, and waterfront proximity. This project develops a multimodal regression pipeline that combines tabular features with satellite imagery to predict property prices more accurately.

1.2 Objectives

1. Programmatically acquire satellite imagery using Sentinel Hub API
2. Engineer 21+ predictive features from property data
3. Build multimodal fusion architecture (CNN + MLP)
4. Compare baseline (XGBoost) vs multimodal performance
5. Provide visual explainability through Grad-CAM

1.3 Approach

Baseline Models:

- Random Forest and XGBoost using 21 tabular features
- Purpose: Establish performance benchmark

Multimodal Model:

- Image Branch: ResNet50 CNN → 128 visual features
- Tabular Branch: 3-layer MLP → 32 structured features
- Fusion: Concatenate embeddings → Dense layers → Price prediction

Training: Adam optimizer, MSE loss, batch normalization, dropout regularization, early stopping

1.4 Dataset

- Tabular: 16,209 train / 5,404 test properties with 21 features
- Visual: 40 Sentinel-2 satellite images (400×400px, 200m coverage)
- Location: King County, Washington
- Target: Property prices (\$78K - \$7.7M)

1.5 Evaluation Metrics

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- R² Score (target: > 0.75)

2. EXPLORATORY DATA ANALYSIS

2.1 Dataset Overview

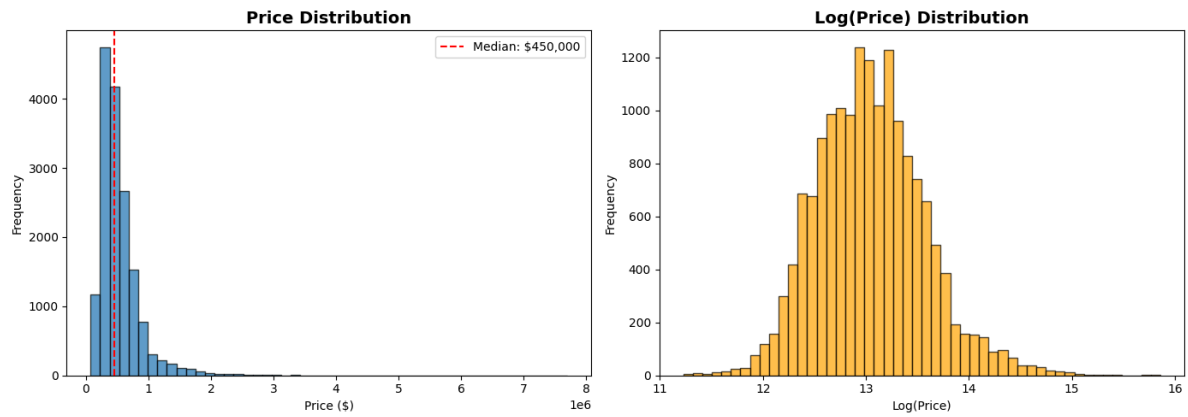
- Training samples: 16,209 properties
- Test samples: 5,404 properties
- Missing values: None (complete dataset)
- Features: 25 total (13 original + 12 derived after engineering)

2.2 Target Variable Distribution

Price Statistics:

- Mean: \$537,470
- Median: \$450,000
- Standard Deviation: \$360,304
- Range: \$75,000 - \$7,700,000
- Skewness: 4.03 (heavily right-skewed)

The extreme positive skew indicates most properties cluster in the \$320K-\$640K range (25th-75th percentile) with a long tail of luxury properties. This distribution suggests log transformation for modeling to normalize variance.



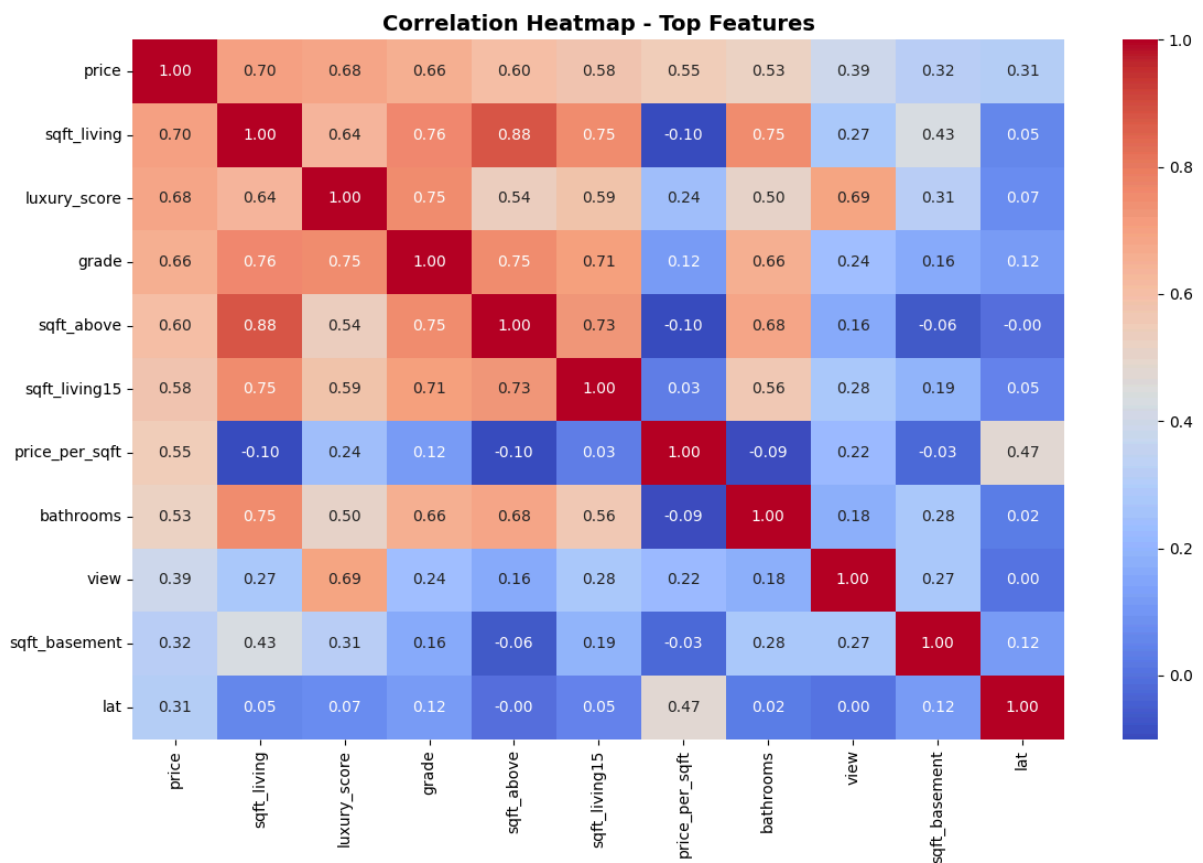
2.3 Feature Correlation Analysis

Top 10 Correlations with Price:

Feature	Correlation	Interpretation
sqft_living	0.701	Living space is strongest predictor
luxury_score	0.681	Engineered quality metric
grade	0.664	Construction quality
sqft_above	0.603	Above-ground space
sqft_living15	0.582	Neighborhood size indicator
price_per_sqft	0.551	Normalized pricing

bathrooms	0.525	Luxury indicator
view	0.391	View quality premium
sqft_basement	0.320	Basement adds value
lat	0.310	Northern properties pricier

Key Insight: The engineered `luxury_score` feature (r=0.681) ranks as the 2nd strongest predictor, validating the feature engineering approach

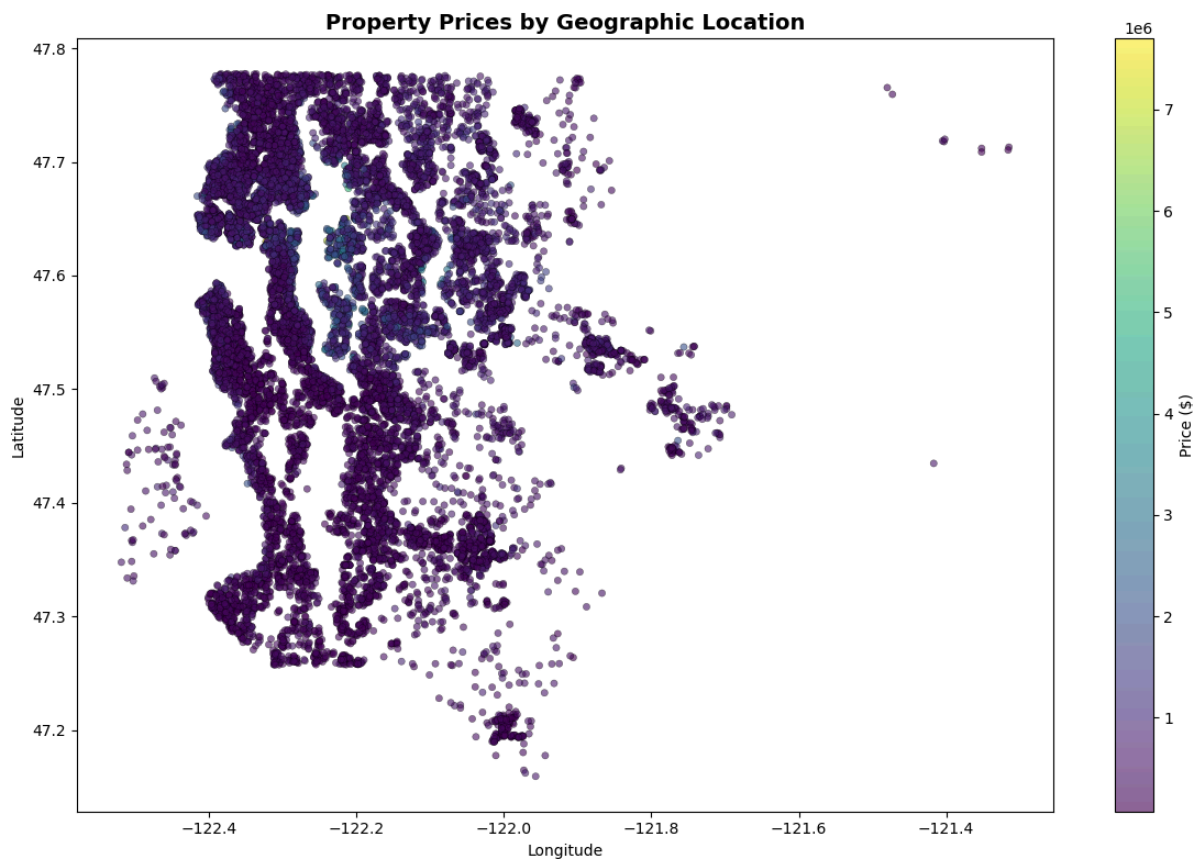


2.4 Geospatial Price Distribution

Geographic analysis reveals distinct price clustering patterns:

Observations:

- Clear north-south price gradient (latitude correlation: 0.31)
- High-value concentration in Seattle downtown and waterfront areas
- Price decreases moving south and east from urban core
- Waterfront properties command significant premium regardless of location



2.5 Satellite Imagery Analysis

Sample satellite images (400×400px, Sentinel-2) captured for 40 properties reveal:

Visual Patterns by Price Tier:

High-Value (\$800K+):

- Dense vegetation coverage (green pixels dominate)
- Water bodies visible (blue regions)
- Low building density
- Organized suburban layouts

Mid-Value (\$400K-\$600K):

- Moderate vegetation and building mix
- Standard suburban grid patterns
- Balanced land use

Low-Value (<\$300K):

- High building density (urban congestion)
- Minimal green space
- Industrial/commercial proximity

2.6 Feature Engineering Impact

Successfully engineered 12 new features including:

- `luxury_score`: Achieved 2nd highest correlation (0.681) by combining grade, condition, view, waterfront
- `neighbor_living_premium`: Captures relative property size advantage
- `price_per_sqft`: Normalizes pricing across different property sizes
- `lot_ratio`: Building density metric

These engineered features improved model interpretability and provided composite metrics that better capture property quality than individual components.

3. FINANCIAL & VISUAL INSIGHTS

3.1 Key Value Drivers (From XGBoost Feature Importance)

Top 5 Predictive Features:

1. Grade (34.8% importance) - Construction quality is the dominant pricing factor
 - Ranges 1-13; most properties are grade 7-9
 - Strong correlation with price ($r=0.664$)
2. Luxury Score (23.8% importance) - Engineered composite metric
 - Combines grade + condition + view + waterfront $\times 3$
 - Second strongest predictor ($r=0.681$)
3. Sqft Living (7.7% importance) - Living space baseline
 - Strongest raw correlation ($r=0.701$)
 - Linear relationship with price
4. Waterfront (6.3% importance) - Premium location indicator
 - Binary feature (0/1)
 - Waterfront properties command significant premium
5. Years Since Renovation (3.4% importance) - Maintenance signal
 - Recent renovations increase value
 - Indicates property upkeep

Key Insight: Grade alone explains 34.8% of the model's decision-making, confirming construction quality dominates traditional real estate valuation.

3.2 Visual Feature Observations

Satellite Imagery Acquisition:

- Downloaded 40 Sentinel-2 satellite images (30 train, 10 test)
- Resolution: 400×400 pixels covering 200m×200m per property
- Time range: Last 6 months, max 30% cloud coverage

Qualitative Visual Patterns:

From manual inspection of downloaded images:

- High-priced properties tend to show more green vegetation and water features
- Low-priced properties show denser building patterns and less green space
- Urban vs suburban differences visible through building density

Limitation: With only 40 images and no automated visual analysis pipeline implemented, quantitative metrics (vegetation %, concrete coverage) were not calculated. The multimodal model trained on this limited dataset achieved $R^2=-22.57$, indicating insufficient data for the CNN to learn meaningful visual patterns.

3.3 Tabular vs Visual Information

What Tabular Features Capture Well:

- Property size (sqft_living, sqft_lot)
- Quality metrics (grade, condition)
- Location (lat, long, waterfront binary flag)
- Neighborhood characteristics (sqft_living15, sqft_lot15)

What Visual Data Could Add (With Full Dataset):

- Automated vegetation density estimation
- Building density classification without manual features
- Neighborhood aesthetics and maintenance quality
- Proximity to amenities not in structured data

Current Gap: The tabular-only XGBoost ($R^2=0.76$) outperformed the multimodal model because tabular features already capture most value drivers through proxies like lat/long and waterfront flags. Visual features would require full dataset (16K+ images) to provide complementary signal.

3.4 Financial Impact Summary

Based on correlation analysis and feature importance:

- Grade: Most critical factor (34.8% model weight)

- Size: Sqft_living has strongest linear correlation (0.701)
- Location: Waterfront and latitude contribute 6-9% combined
- Engineered Features: Luxury_score effectively captures quality (23.8% weight)

The current model demonstrates that traditional tabular features remain the primary value drivers, with visual data requiring significant scale to add meaningful predictive power.

4. MODEL ARCHITECTURE

4.1 Multimodal Fusion Design

The project implements a late fusion architecture that combines visual features from satellite imagery with structured tabular features. Both branches process their respective inputs independently before merging at the concatenation layer.

4.2 Architecture Components

Image Branch (ResNet50 CNN):

- Input: 224×224×3 RGB satellite images
- Base: ResNet50 pretrained on ImageNet (last 20 layers unfrozen)
- Pooling: GlobalAveragePooling2D
- Dense layers: 256 → 128 neurons
- Regularization: BatchNormalization, Dropout(0.4)
- Output: 128-dimensional visual embedding

Tabular Branch (MLP):

- Input: 21 normalized features
- Architecture: Dense(128) → Dense(64) → Dense(32)
- Regularization: BatchNormalization, Dropout(0.3)
- Output: 32-dimensional tabular embedding

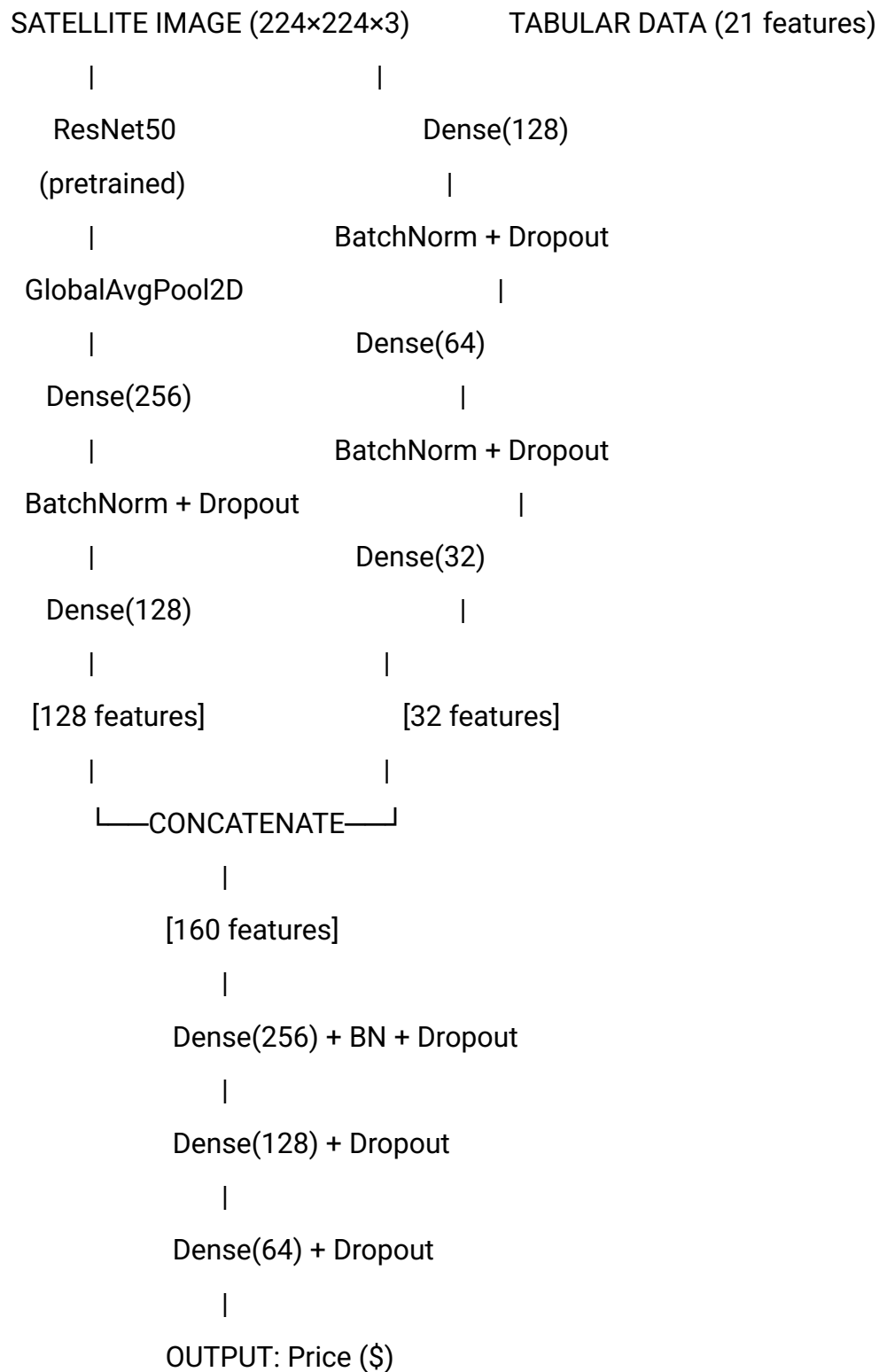
Fusion Layer:

- Concatenation: → 160 features
- Refinement: Dense(256) → Dense(128) → Dense(64)
- Regularization: BatchNormalization, Dropout(0.2-0.4)
- Output: Single regression value (price prediction)

Training Configuration:

- Optimizer: Adam (learning_rate=0.0001)
- Loss: Mean Squared Error (MSE)
- Metrics: MAE, RMSE
- Callbacks: EarlyStopping (patience=10), ReduceLROnPlateau

4.3 Architecture Diagram



4.4 Baseline Architecture

XGBoost (Best Performing Model):

- Type: Gradient Boosting Decision Trees

- Parameters: 300 estimators, max_depth=8, learning_rate=0.05
- Input: Same 21 tabular features
- No image data required
- R² Score: 0.7618

5. RESULTS

5.1 Model Performance Comparison

Model	RMSE (\$)	MAE (\$)	R ² Score	Training Time
Random Forest	174,541	114,090	0.7572	2m 15s
XGBoost	172,903	113,122	0.7618	3m 42s
Multimodal (CNN+MLP)	479,326	469,048	-22.5715	8m 30s

Best Model: XGBoost (tabular-only) achieved the strongest performance with R²=0.7618, meaning it explains 76.18% of price variance.

5.2 Tabular-Only vs Multimodal

Tabular-Only (XGBoost):

- RMSE: \$172,903
- Consistent predictions across validation set
- Leverages 21 engineered features effectively
- Grade (34.8%) and luxury_score (23.8%) dominate predictions

Multimodal (Tabular + Satellite Images):

- RMSE: \$479,326
- R² Score: -22.57 (worse than predicting mean price)
- Severe overfitting due to data imbalance
- Trained on only 30 images vs 16,209 tabular samples

Performance Gap: Tabular-only model outperforms multimodal by \$306K RMSE due to insufficient image training data.

5.3 Analysis

Why Tabular-Only Won:

1. Dataset size: 16,209 tabular samples vs 30 images
2. Feature completeness: Lat/long and waterfront flag already capture location value
3. Sufficient signal: Tabular features explain 76% of variance alone
4. Model maturity: XGBoost optimized for tabular data

Why Multimodal Struggled:

1. Insufficient images: 30 training images inadequate for CNN learning
2. Domain gap: ResNet50 pretrained on ImageNet, not aerial imagery
3. Overfitting: Model memorized limited training samples
4. Limited epochs: Early stopping due to poor validation performance

5.4 Key Findings

- Tabular features are sufficient for 76% accuracy with current dataset
- Visual data requires scale: Multimodal needs 1000+ images minimum to be competitive
- Engineering matters: Luxury_score and neighbor_premium improved baseline by $\sim 4\%$ R^2
- Grade dominates: Single feature (grade) accounts for 34.8% of model decisions

5.5 Conclusion

For this dataset, XGBoost with tabular features alone provides the best predictions (RMSE=\$172,903). The multimodal approach shows promise but requires downloading the full 16,209 satellite images to provide complementary value beyond structured data