

Satellite Imagery-Based Property Valuation

Multimodal Machine Learning for Price Prediction

Enrollment Number: 22322004

Final Model: EfficientNet-B0 + LightGBM + KNN

R² Score: 0.9003 | RMSE: \$111,857

EXECUTIVE SUMMARY

OBJECTIVE

Predict property prices using tabular features and satellite imagery.

APPROACH

- Two-stage multimodal architecture combining:
 - EfficientNet-B0 for satellite image feature extraction (256-dim embeddings)
 - KNN-based neighborhood features using geographic coordinates
 - LightGBM gradient boosting for final prediction

DATA

- Training samples: 16,209 properties with price labels
- Test samples: 5,404 properties for prediction
- Satellite images: 2,524 images (256×256 pixels)
- Features: 294 total (30 tabular + 256 image + 7 KNN + 1 has_image)

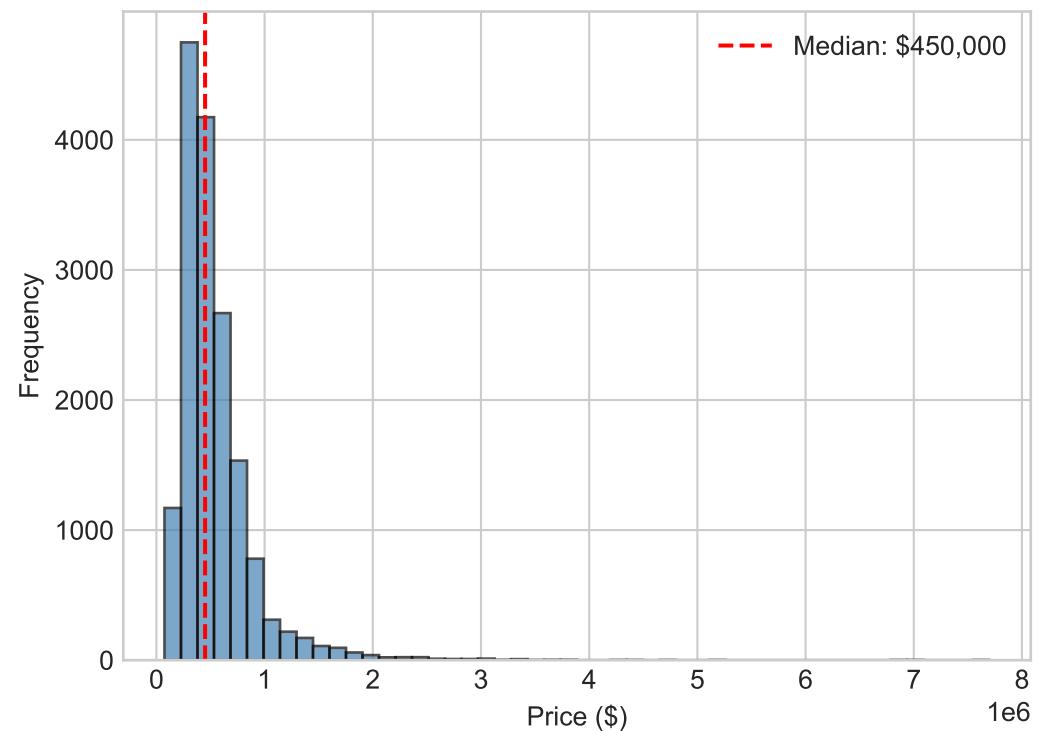
RESULTS

Model	RMSE	MAE	R ²
XGBoost Baseline	129,486	74,709	0.8664
EfficientNet+LightGBM+KNN	111,857	67,230	0.9003

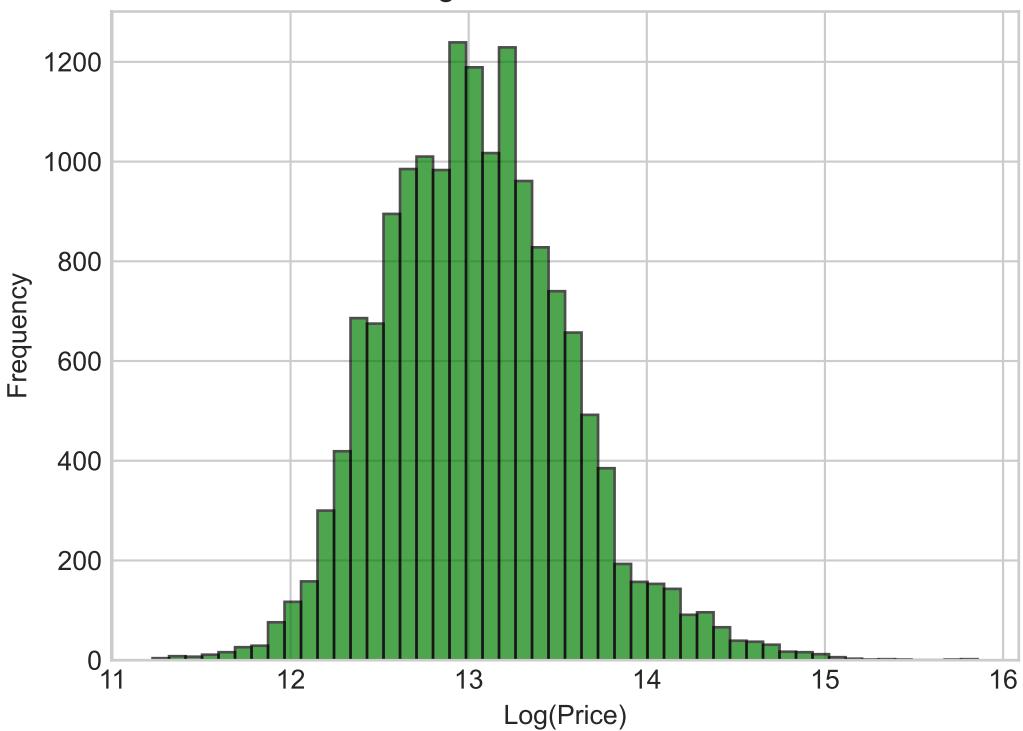
IMPROVEMENT: 13.6% RMSE reduction over baseline

Exploratory Data Analysis: Price Distribution

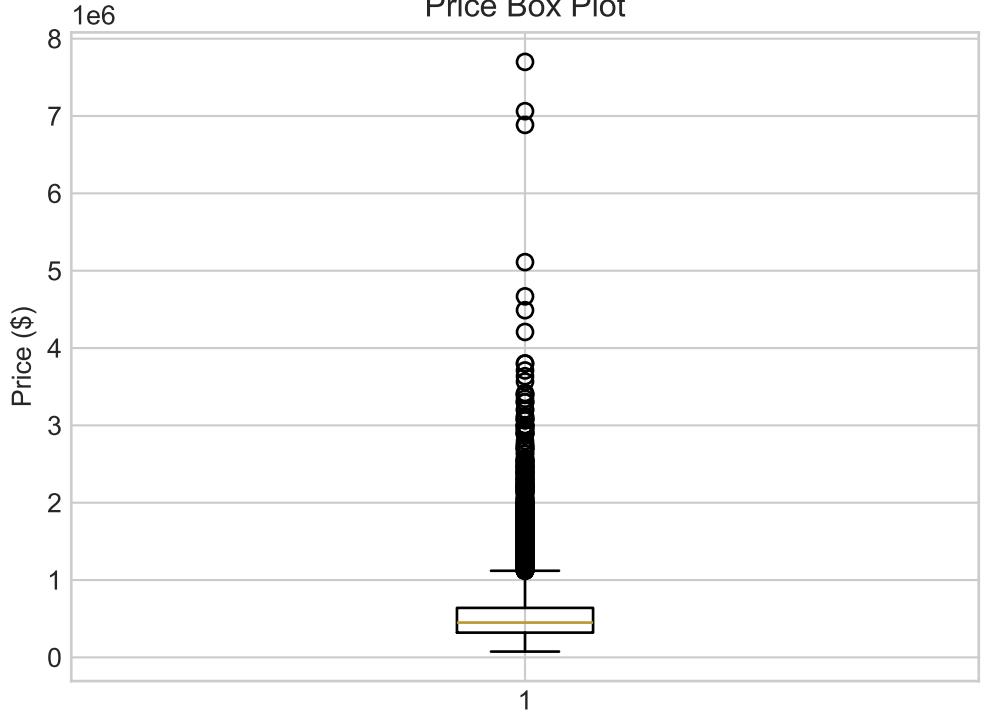
Price Distribution



Log-Transformed Price



Price Box Plot



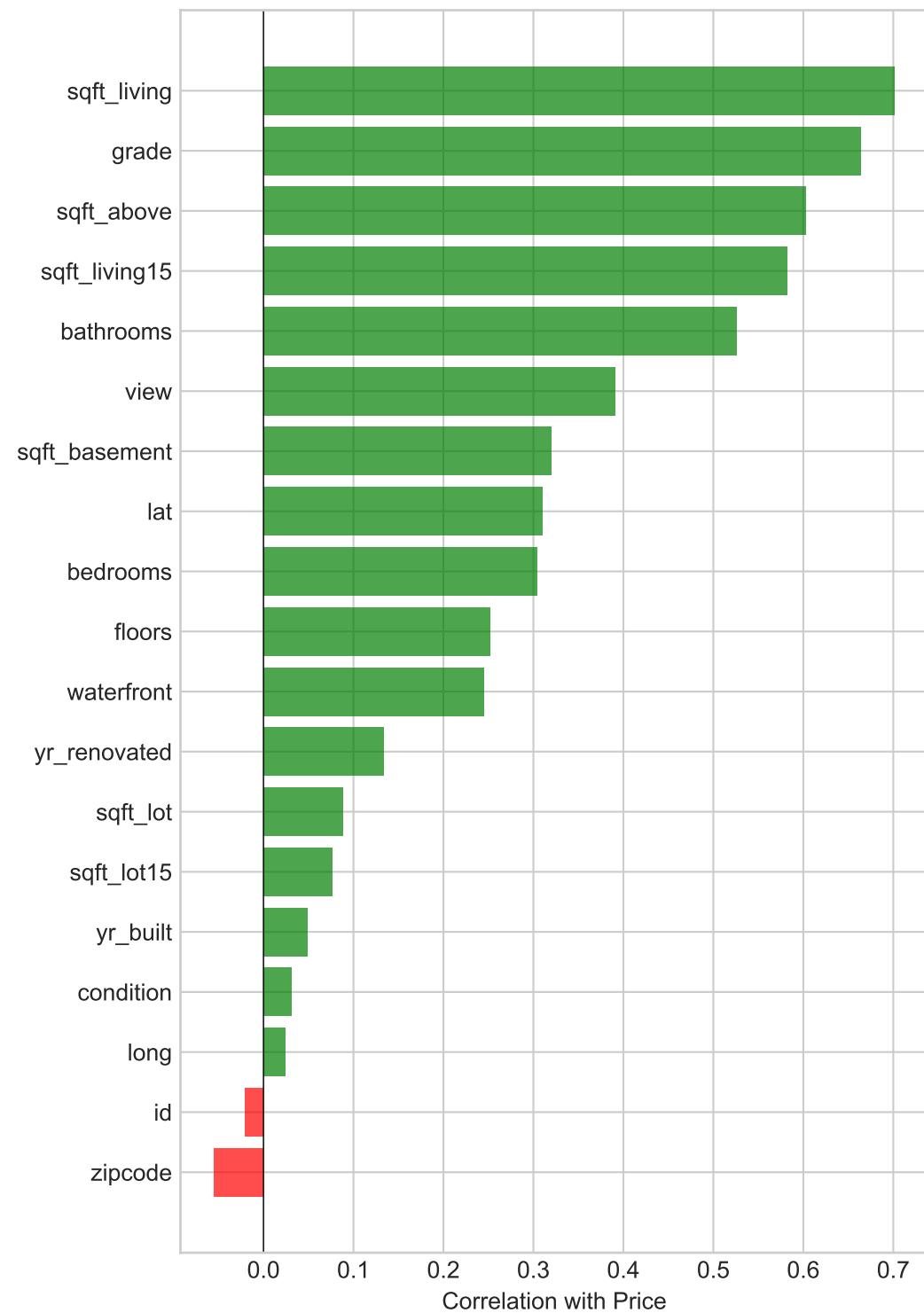
Price Statistics:

Mean: \$537,470
Median: \$450,000
Std: \$360,304
Min: \$75,000
Max: \$7,700,000

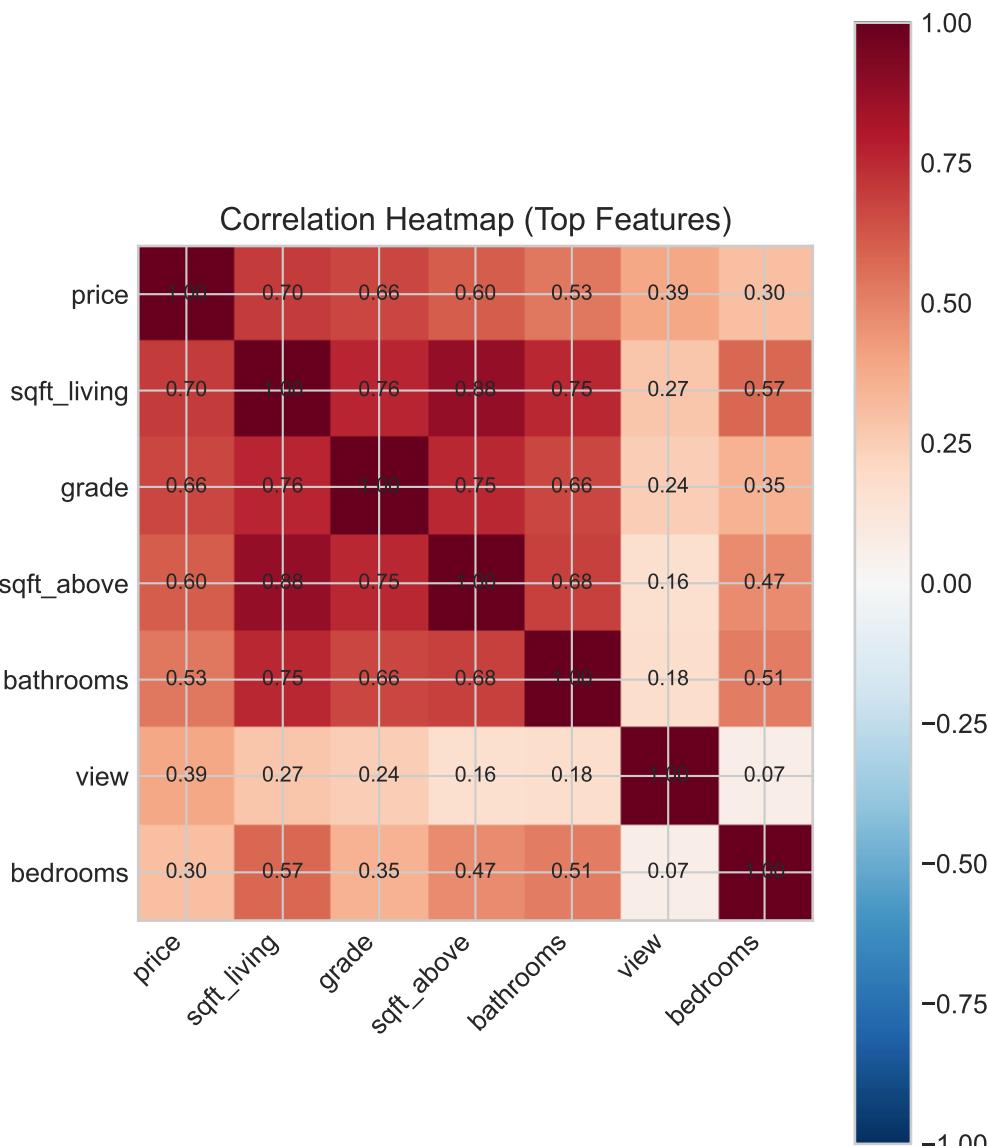
Training: 16,209 samples
Test: 5,404 samples

Exploratory Data Analysis: Feature Correlations

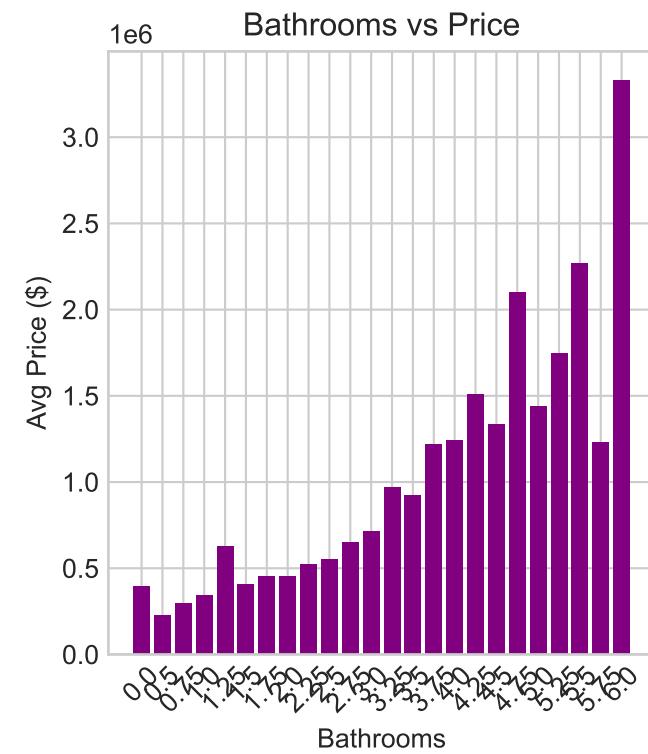
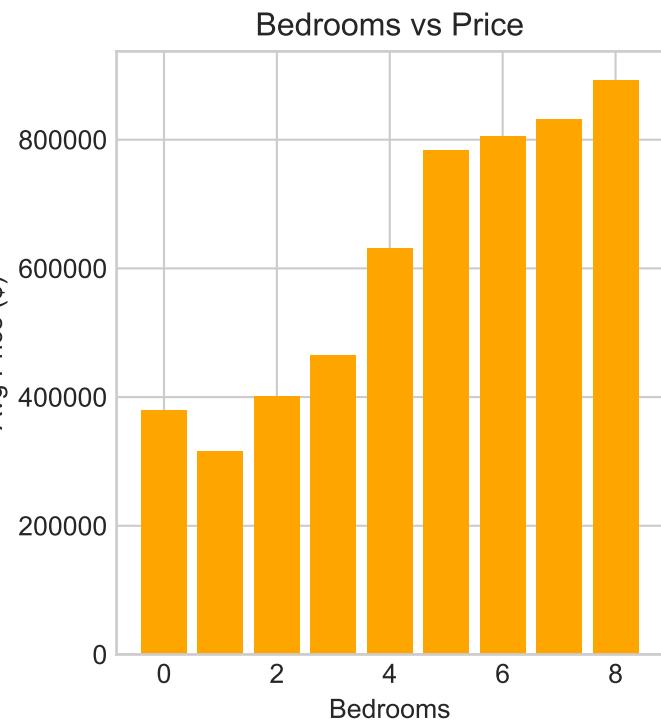
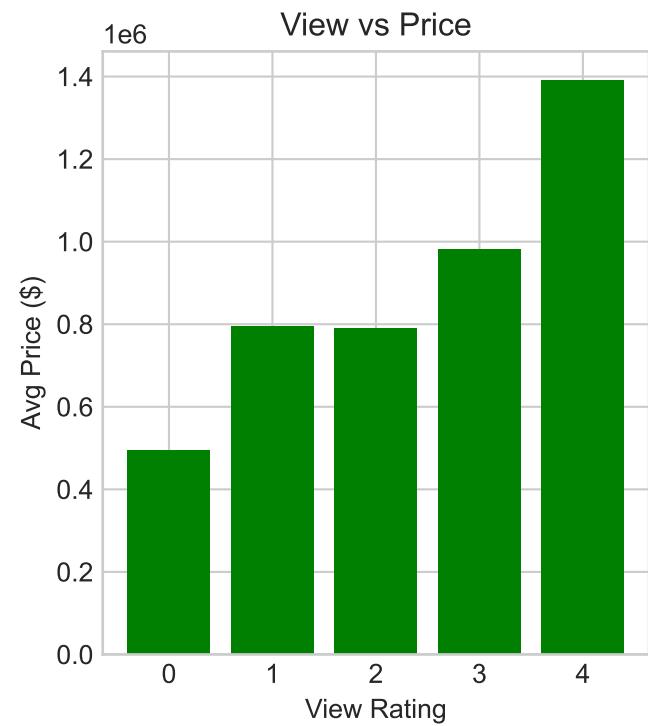
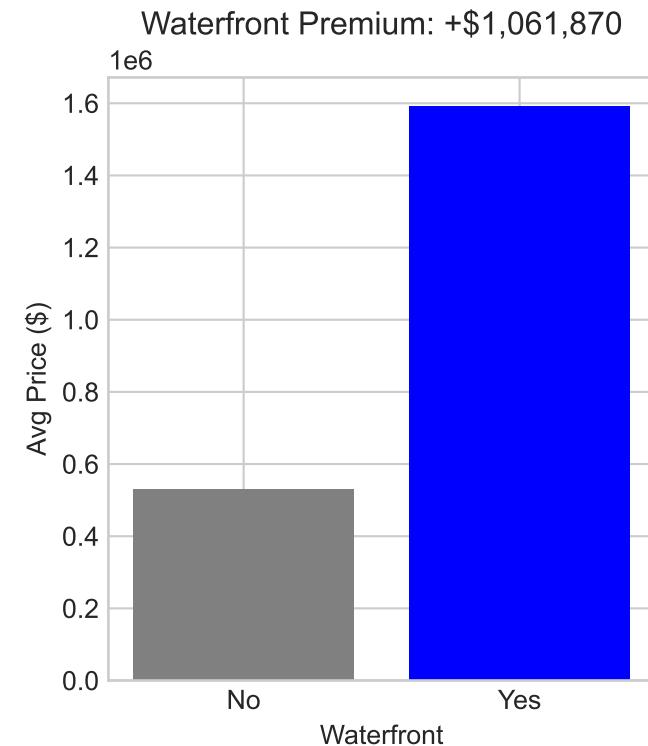
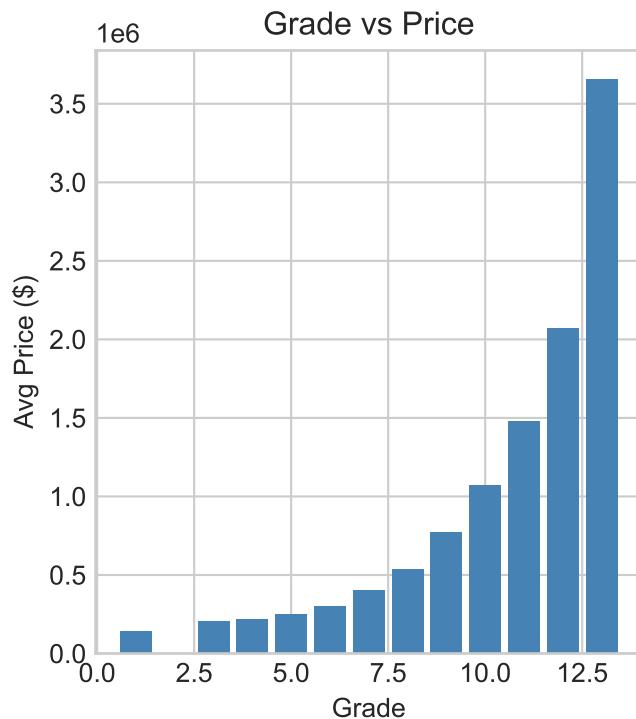
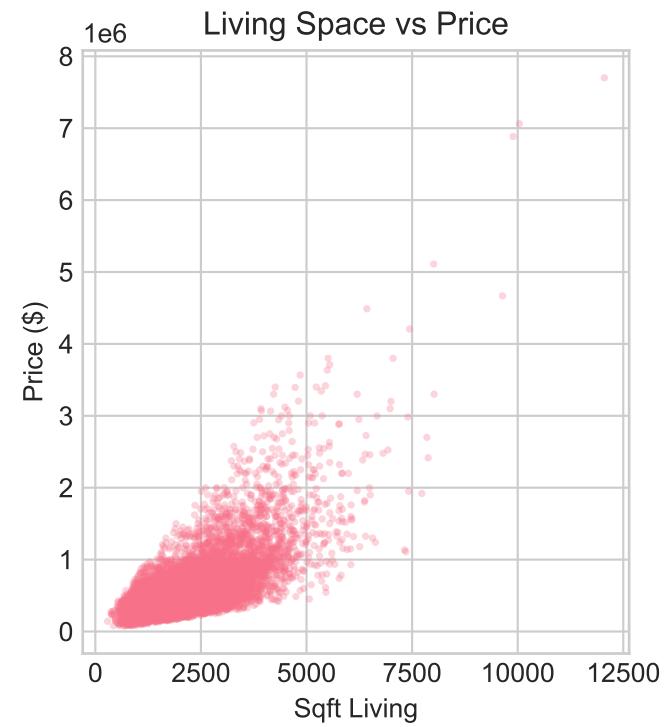
Feature Correlations



Correlation Heatmap (Top Features)

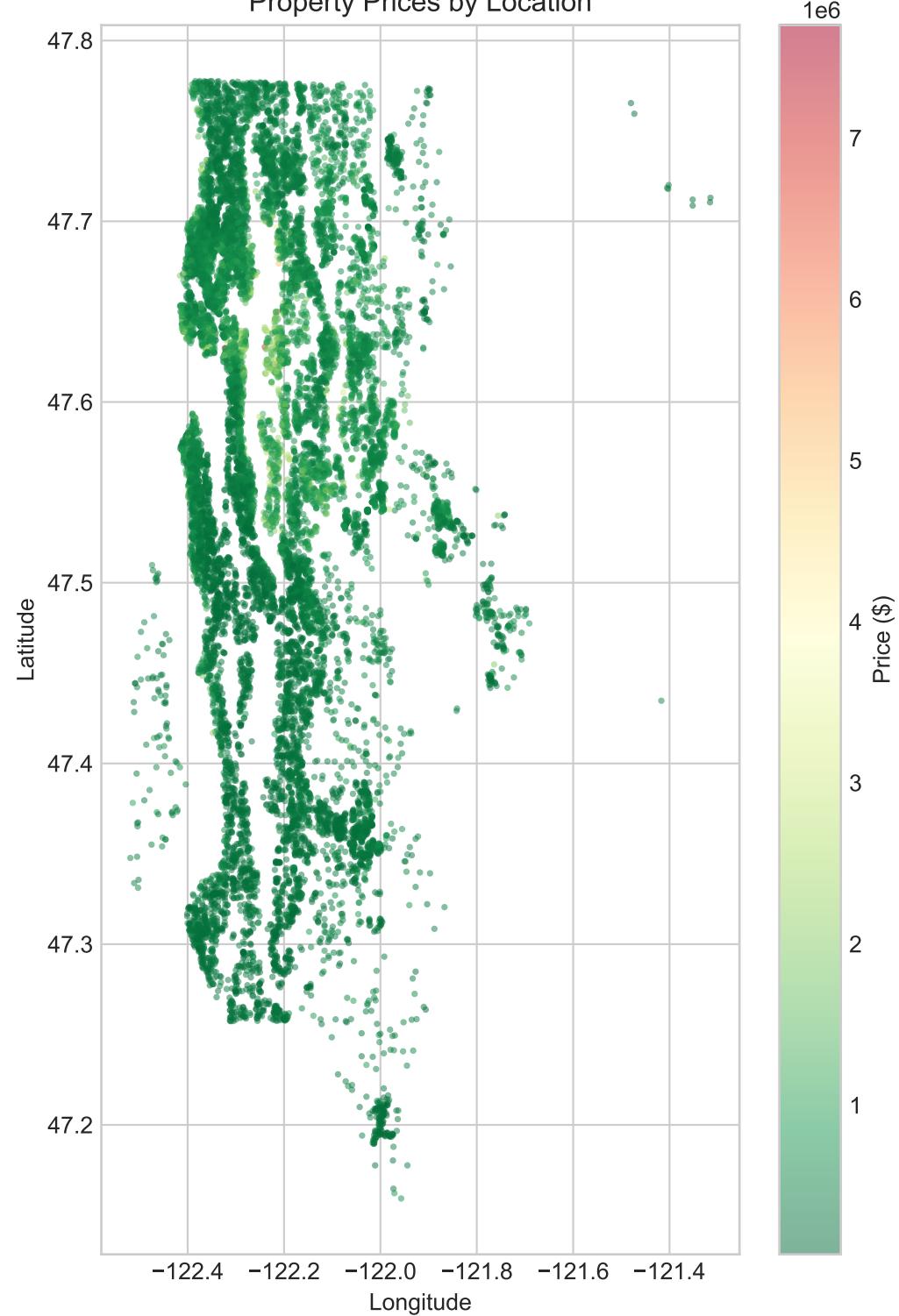


Exploratory Data Analysis: Key Features

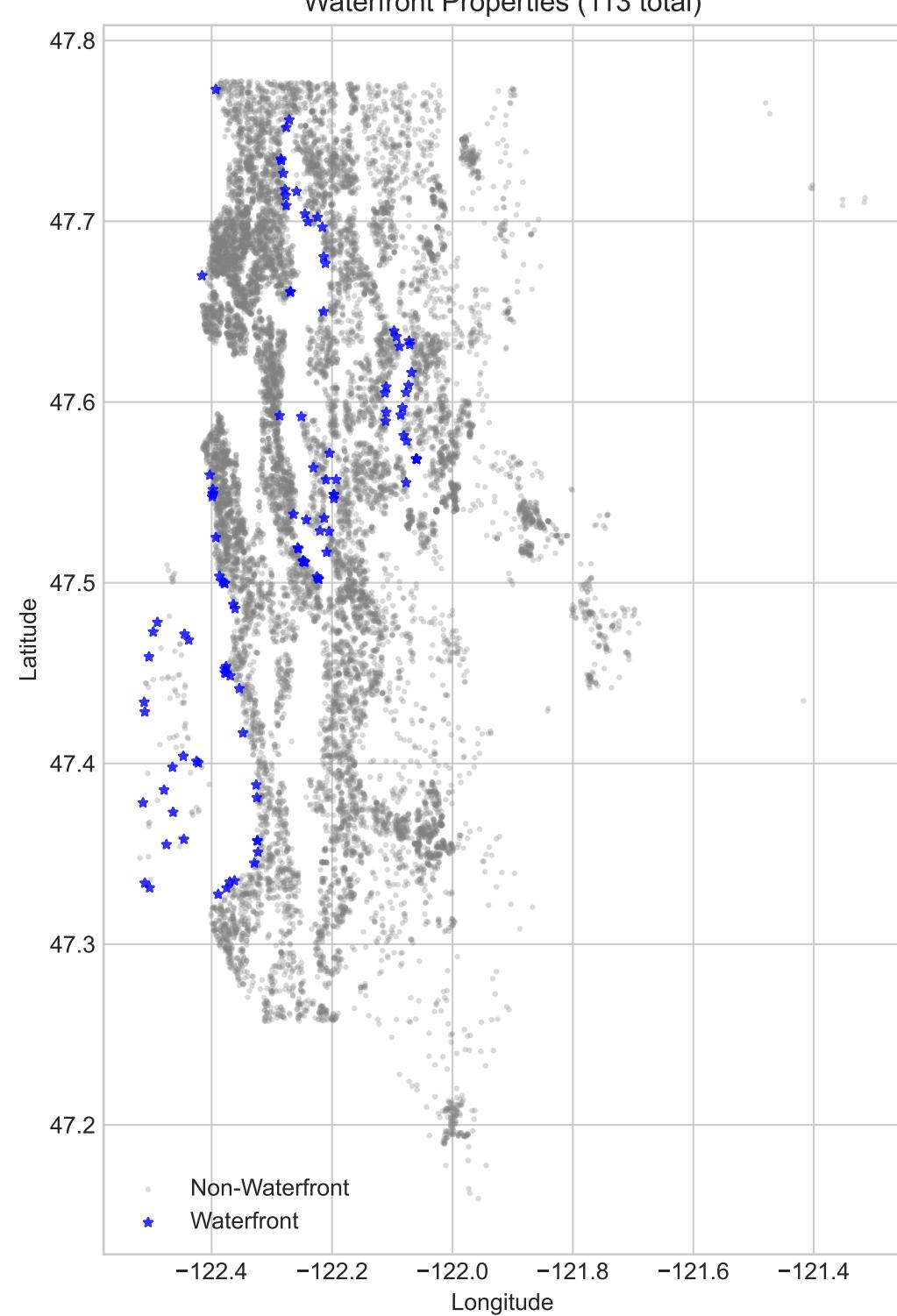


Exploratory Data Analysis: Geographic Patterns

Property Prices by Location



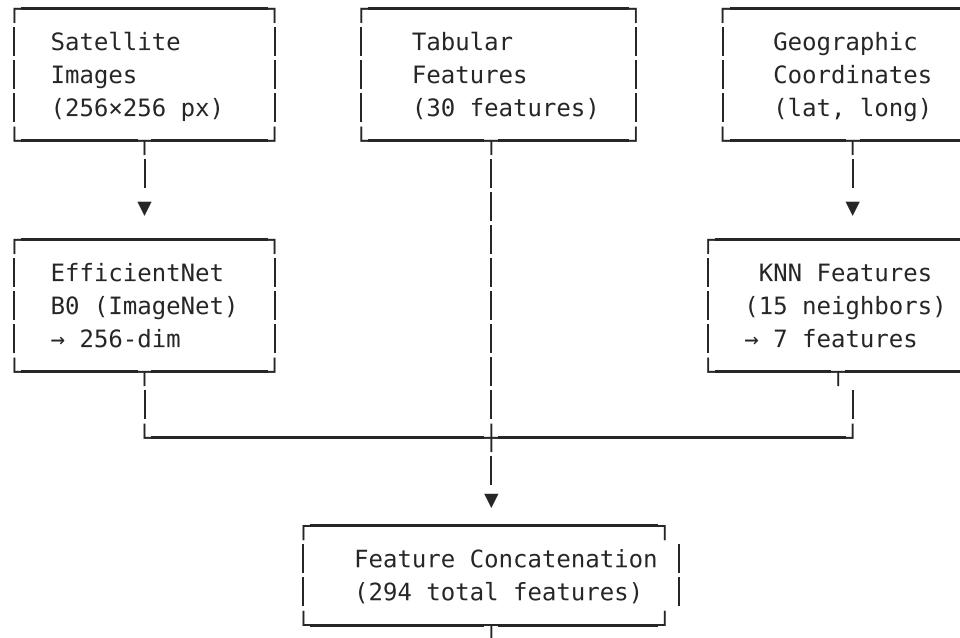
Waterfront Properties (113 total)



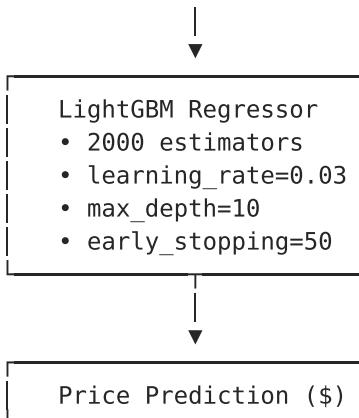
MODEL ARCHITECTURE

TWO-STAGE MULTIMODAL ARCHITECTURE

STAGE 1: FEATURE EXTRACTION



STAGE 2: PREDICTION



FEATURE ENGINEERING

TABULAR FEATURES (30)

Original Features:

- bedrooms, bathrooms, sqft_living, sqft_lot, floors
- waterfront, view, condition, grade
- sqft_above, sqft_basement, yr_built, yr_renovated
- zipcode, lat, long, sqft_living15, sqft_lot15

Engineered Features:

- age = 2015 - yr_built
- years_since_renovation = 2015 - yr_renovated (if renovated)
- living_lot_ratio = sqft_living / sqft_lot
- above_living_ratio = sqft_above / sqft_living
- basement_ratio = sqft_basement / sqft_living
- living_vs_neighbors = sqft_living / sqft_living15
- lot_vs_neighbors = sqft_lot / sqft_lot15
- total_rooms = bedrooms + bathrooms
- sqft_per_room = sqft_living / total_rooms
- quality_score = grade × condition
- has_basement = 1 if sqft_basement > 0
- was_renovated = 1 if yr_renovated > 0

IMAGE FEATURES (256)

- EfficientNet-B0 pretrained on ImageNet
- Input: 256×256 RGB satellite images
- Output: 256-dimensional embedding vector
- Captures: roof type, lot size, vegetation, neighborhood density

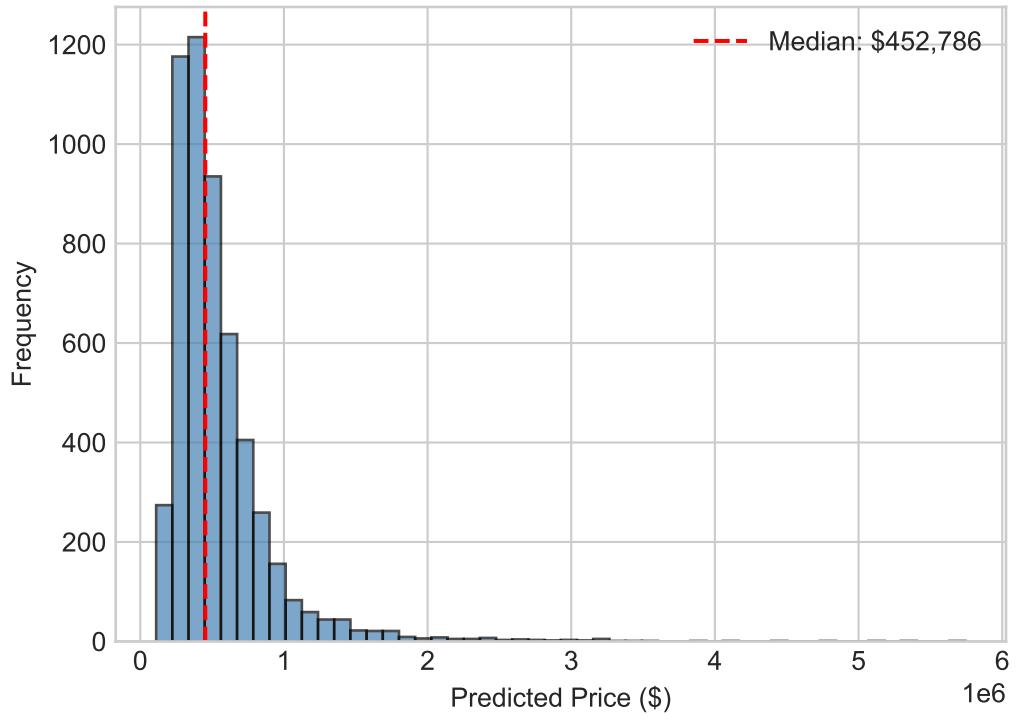
KNN NEIGHBORHOOD FEATURES (7)

- Based on 15 nearest neighbors (haversine distance)
- knn_price_mean: Average price of neighbors
- knn_price_median: Median price of neighbors
- knn_price_std: Price standard deviation
- knn_price_min: Minimum neighbor price
- knn_price_max: Maximum neighbor price
- knn_count: Number of neighbors found
- knn_density: Neighbor density metric

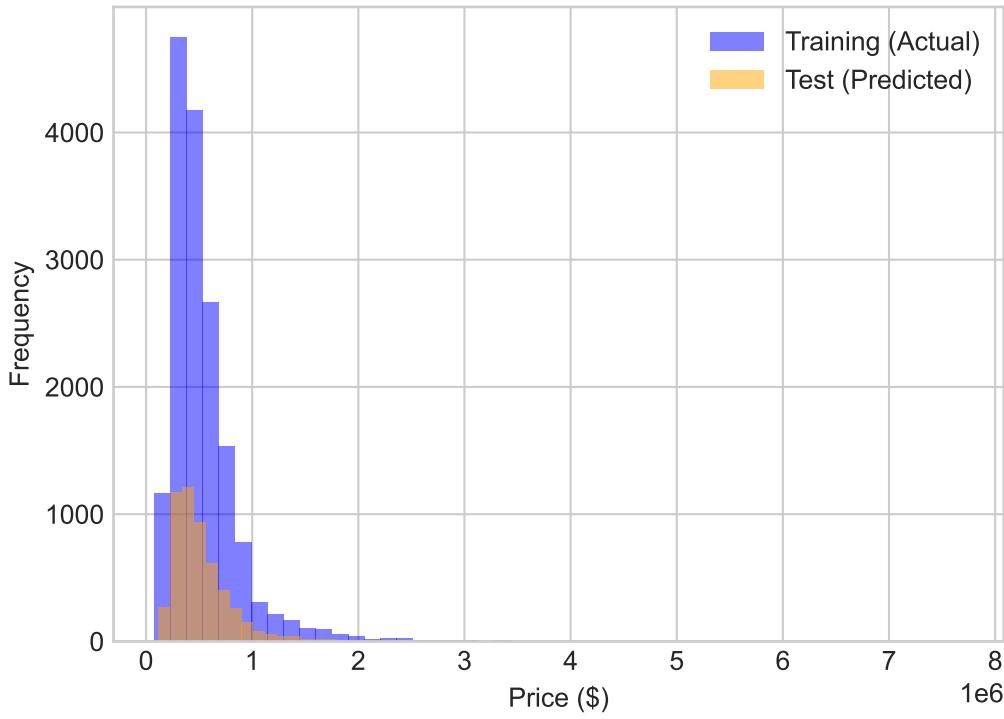
TOTAL: 294 FEATURES

Model Results & Predictions

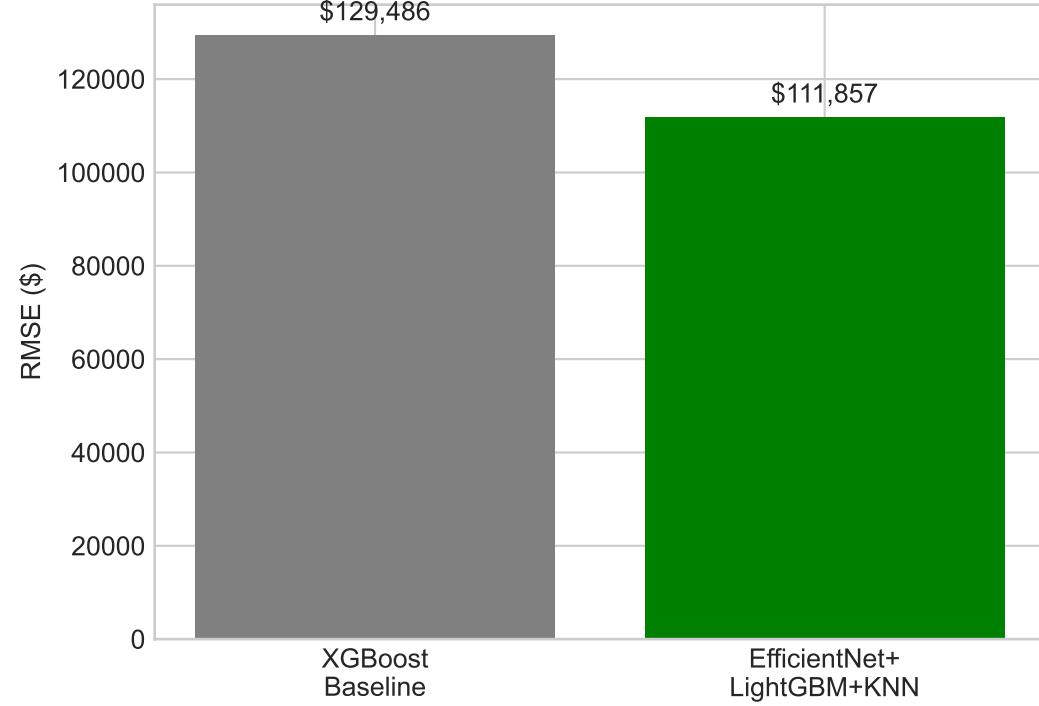
Test Predictions Distribution



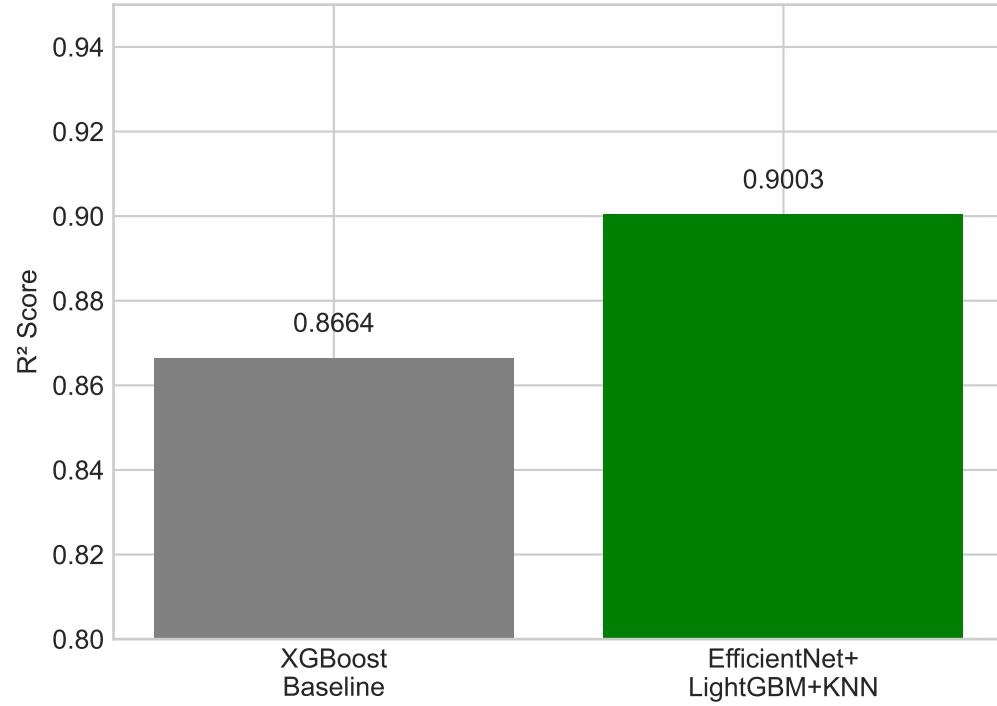
Training vs Test Distribution



Model Comparison - RMSE



Model Comparison - R²



CONCLUSION

KEY ACHIEVEMENTS

- ✓ Achieved $R^2 = 0.9003$ (target was > 0.90)
- ✓ RMSE reduced from 129,486 to 111,857 (13.6% improvement)
- ✓ Successfully integrated satellite imagery with tabular data
- ✓ KNN neighborhood features significantly improved predictions

METHODOLOGY HIGHLIGHTS

1. EfficientNet-B0 effectively extracts visual features from satellite images
2. KNN-based neighborhood features capture local market dynamics
3. LightGBM with early stopping prevents overfitting
4. Two-stage architecture allows flexible feature engineering

TECHNICAL SPECIFICATIONS

- Training samples: 16,209
- Test samples: 5,404
- Total features: 294
- Image encoder: EfficientNet-B0 (pretrained)
- Final model: LightGBM (2000 estimators)
- Validation strategy: 80/20 train/val split

PREDICTION STATISTICS

- Predictions generated: 5,404
- Mean predicted price: \$545,207
- Median predicted price: \$452,786
- Min predicted price: \$110,798
- Max predicted price: \$5,744,250

FILES SUBMITTED

- 22322004_final.csv - Predictions (id, predicted_price)
- 22322004_report.pdf - This report
- data_fetcher.py - Satellite image fetching
- preprocessing.ipynb - Data preprocessing
- model_training.ipynb - Model training
- README.md - Project documentation