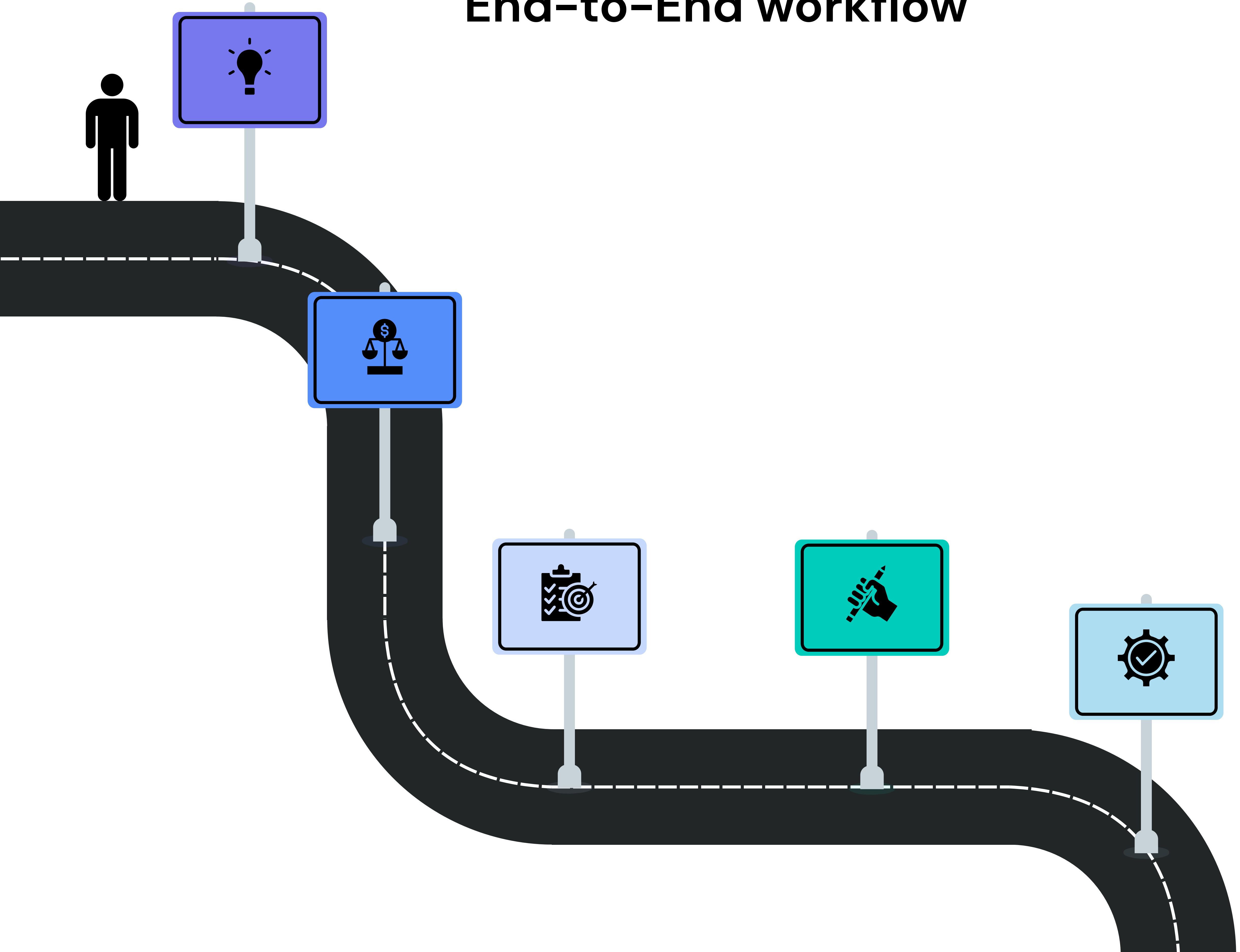
Open Project 2025-26 || Data Science PS

Satellite Imagery-Based Property Valuation

CDC X Yhills

Submitted By-
Rishita Rathi (ECONOMICS)
3rd Year
23322026

End-to-End workflow



01 Problem & Approach

Formulate the house price prediction task and decide on a multimodal approach using tabular data and satellite imagery.

02 Exploratory Data Analysis

Visualize price distributions, geographic variations, and sample satellite images to understand key drivers.

03 Feature & Insight Extraction

Extract tabular features and semantic visual cues (e.g., greenery vs. concrete) influencing property value.

04 Model Architecture Design

Building the multimodal learning pipeline using satellite images and tabular data.

05 Evaluation & Results

Compare model performance and interpret results

Overview

The Problem:

- Traditional models rely only on then static numerical inputs like square footage and bedroom counts, which suffer from **dimensional blindness** by ignoring the surrounding environmental quality that drives value.

The Solution:

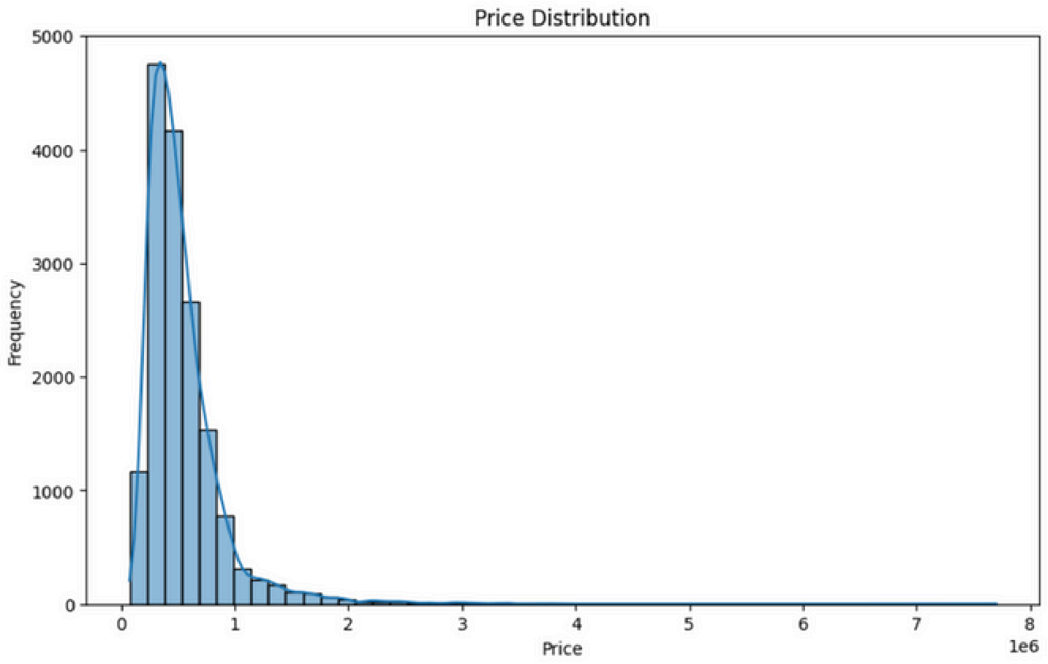
- Developing a **Multimodal Regression Pipeline** that combines historical housing data with programmatically acquired satellite imagery.
- This system integrates environmental context such as vegetation density into a **single predictive engine**

Core Objectives

- **Data Fusion:** Build a single brain architecture (CNN + MLP) capable of learning from pixels and tabular numbers simultaneously.
- **Feature Engineering:** Programmatically extract semantic features Green and Concrete scores to quantify neighborhood characteristics.
- **Interpretability:** Use Grad-CAM to visualize Price Attention Regions, proving why the model values certain visual features over others.
- **Explainable Real Estate AI:** By utilizing PCA and K-Means clustering, the system translates 512-dimensional visual vectors into four human interpretable market segments: Urban, Suburban, Rural, and Luxury.

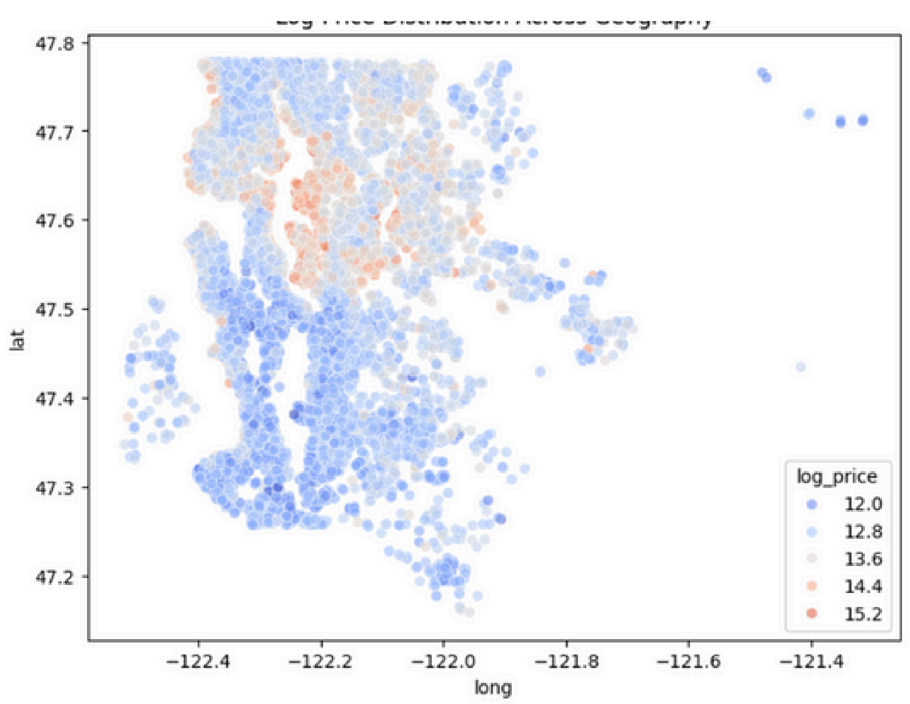
EDA

Price Distribution



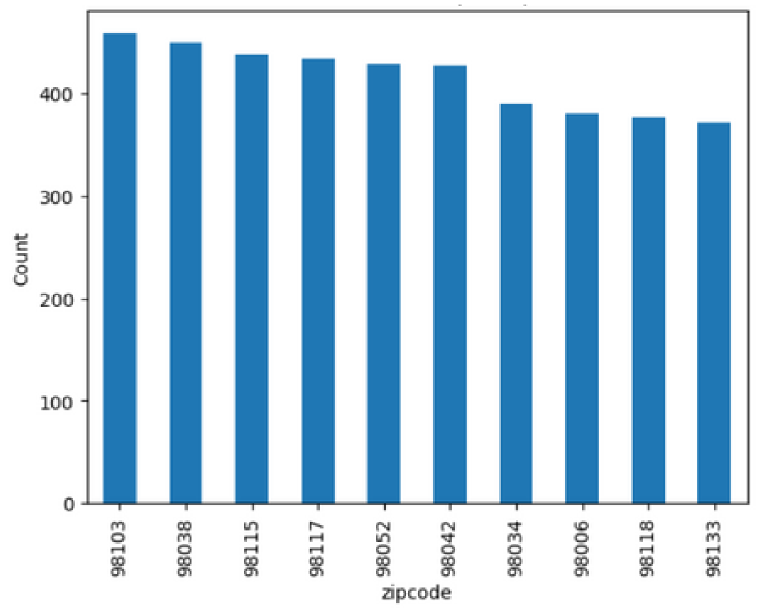
- The house price distribution is **highly right-skewed**, with a long tail toward expensive properties.
- Most houses are concentrated in the lower–mid price range, indicating a **majority of moderately priced homes.**
- A small number of luxury properties exist at very high prices, **acting as outliers.**

Log-Price Distribution Across Geography

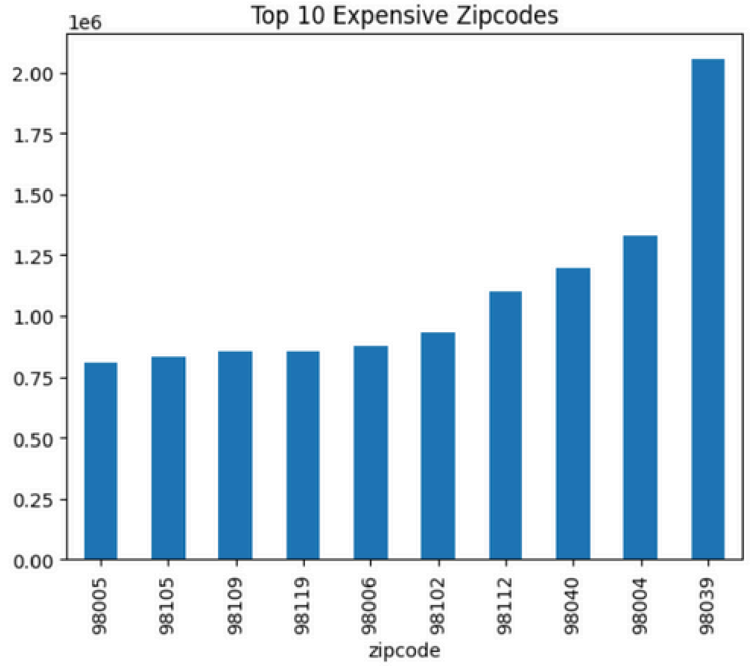


- House prices vary clearly across locations and **form distinct spatial clusters.**
- Higher-priced houses appear in specific areas, not evenly across the region.
- Lower-priced houses are more common in outer locations.
- **This shows that location (latitude and longitude) strongly influences prices.**

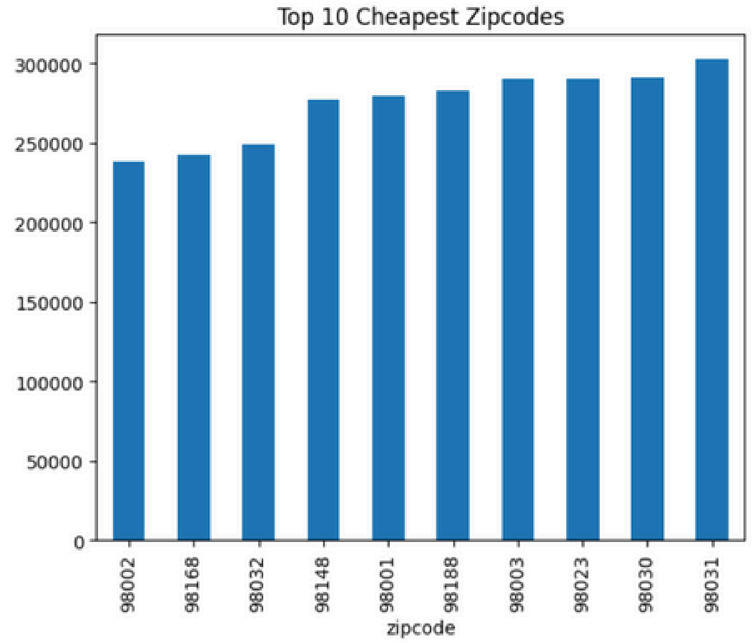
Number of Houses Per Zipcode



Top 10 Expensive Zipcode

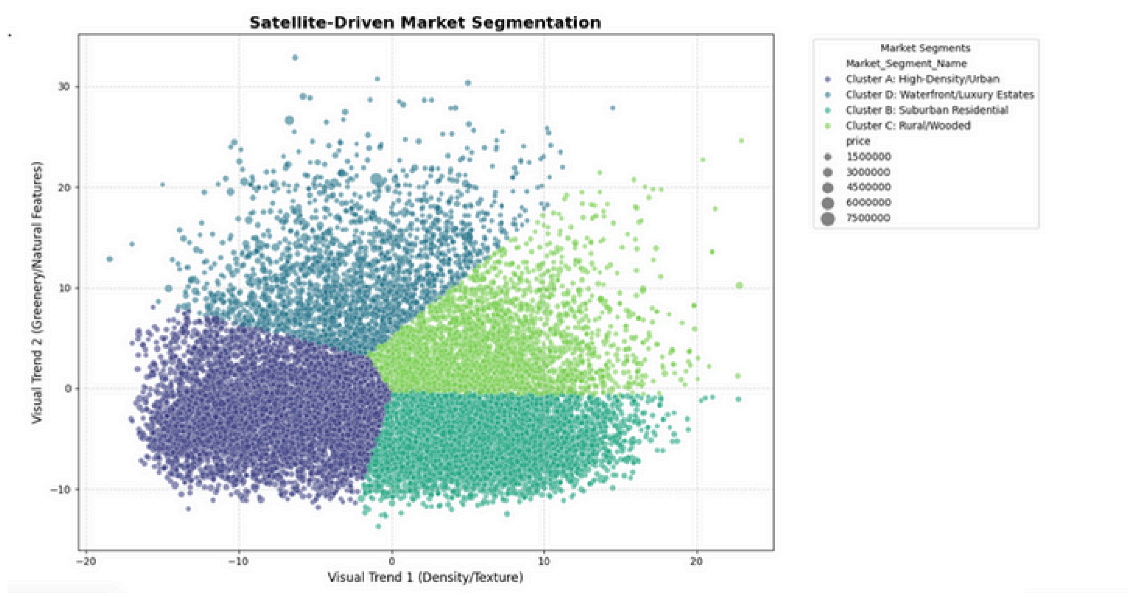


Top 10 Cheapest Zipcodes

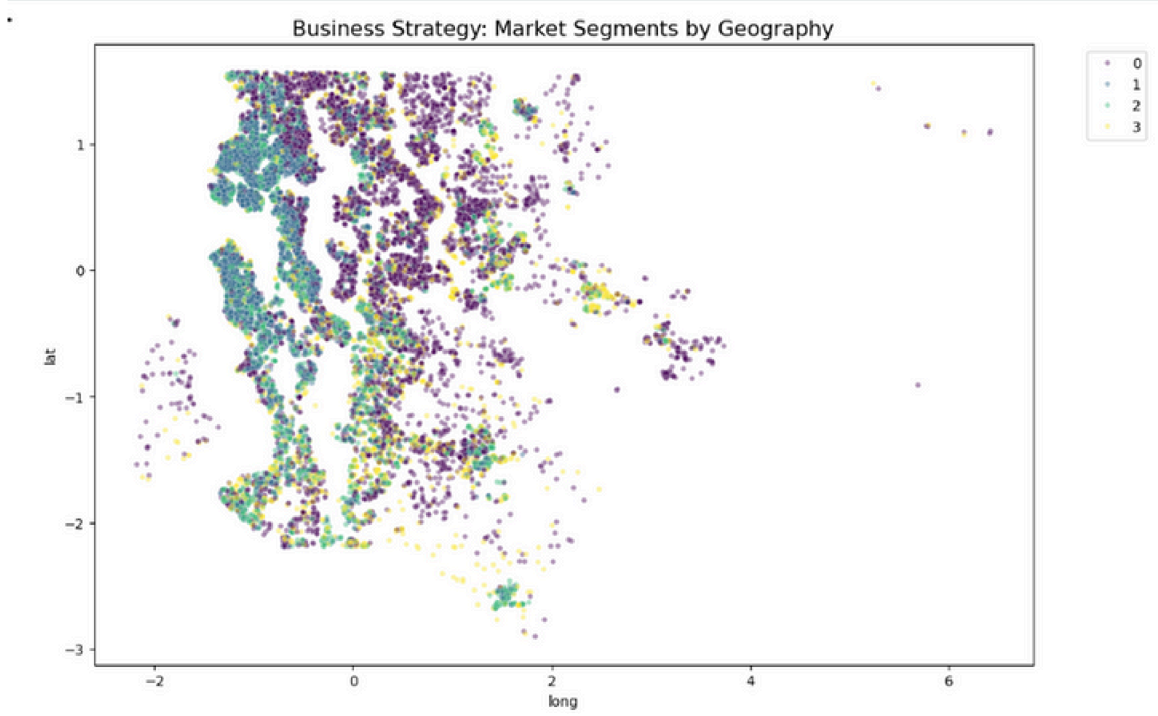


- Strong geographic **price variation exists across zipcodes**, indicating that location plays a crucial role in house pricing.

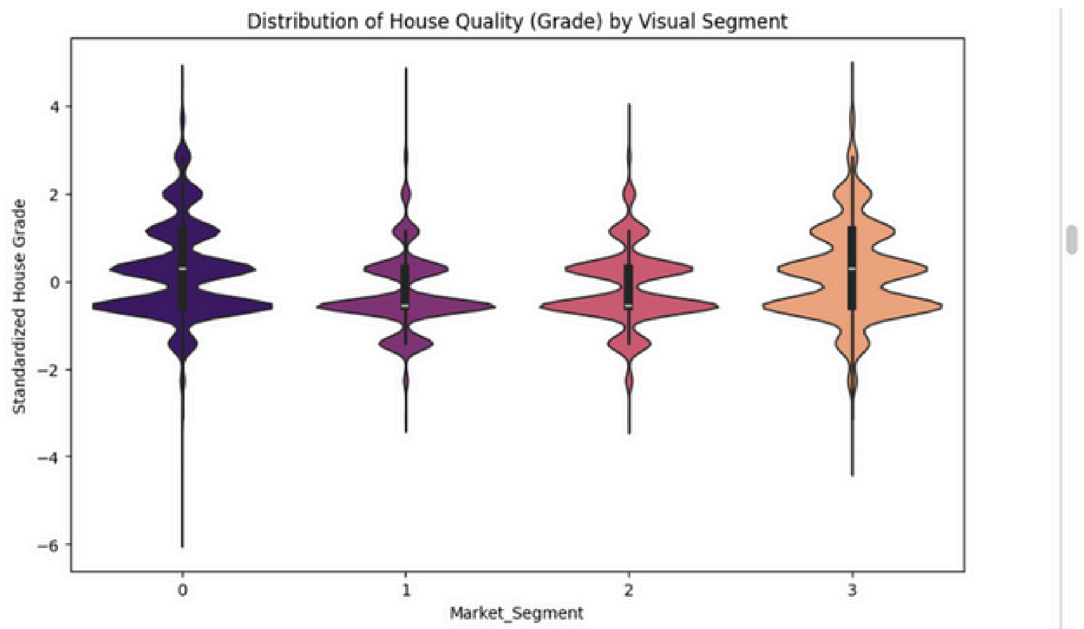
EDA & Market Segmentation



- **Cluster A:** High-Density/Urban: Characterized by low greenery and high structural density. These are likely city centers or older urban neighborhoods.
- **Cluster D:** Waterfront/Luxury Estates: These points have the highest Greenery scores and show the largest bubble sizes indicating that lush visual features are a top predictor for high-value luxury properties.
- **Cluster B:** Suburban Residential: Moderate density with some open space but lower natural vegetation represents standard residential housing tracts.
- **Cluster C:** Rural/Wooded: Low structural density and high greenery. These properties have lots of land and trees but have lower prices than Cluster D due to distance from urban amenities.



- The visualization helps identify geographical concentrations of specific market segments and their associated profitability.
- For example, the dense purple areas on the left indicate high-density city regions, while the green points, found near the edges, represent waterfront luxury estates.
- **The size variation within these clusters allows for quick identification of the most profitable locations within each segment.**



- The wider sections indicate a higher concentration of houses at that specific grade, while narrower sections or the tails show fewer houses.
- Market Segment 0: Shows a wide range of house grades, centered around the standardized mean of 0, indicating high variability in quality within this segment.
- Market Segment 1 & 2: These segments have narrower distributions, suggesting more consistent house quality within each group. Segment 1 is centered near the mean 0, while Segment 2 is centered slightly below the mean.
- Market Segment 3: It has the highest average house grade, centered above the mean, and exhibits a wide distribution, similar to Segment 0.

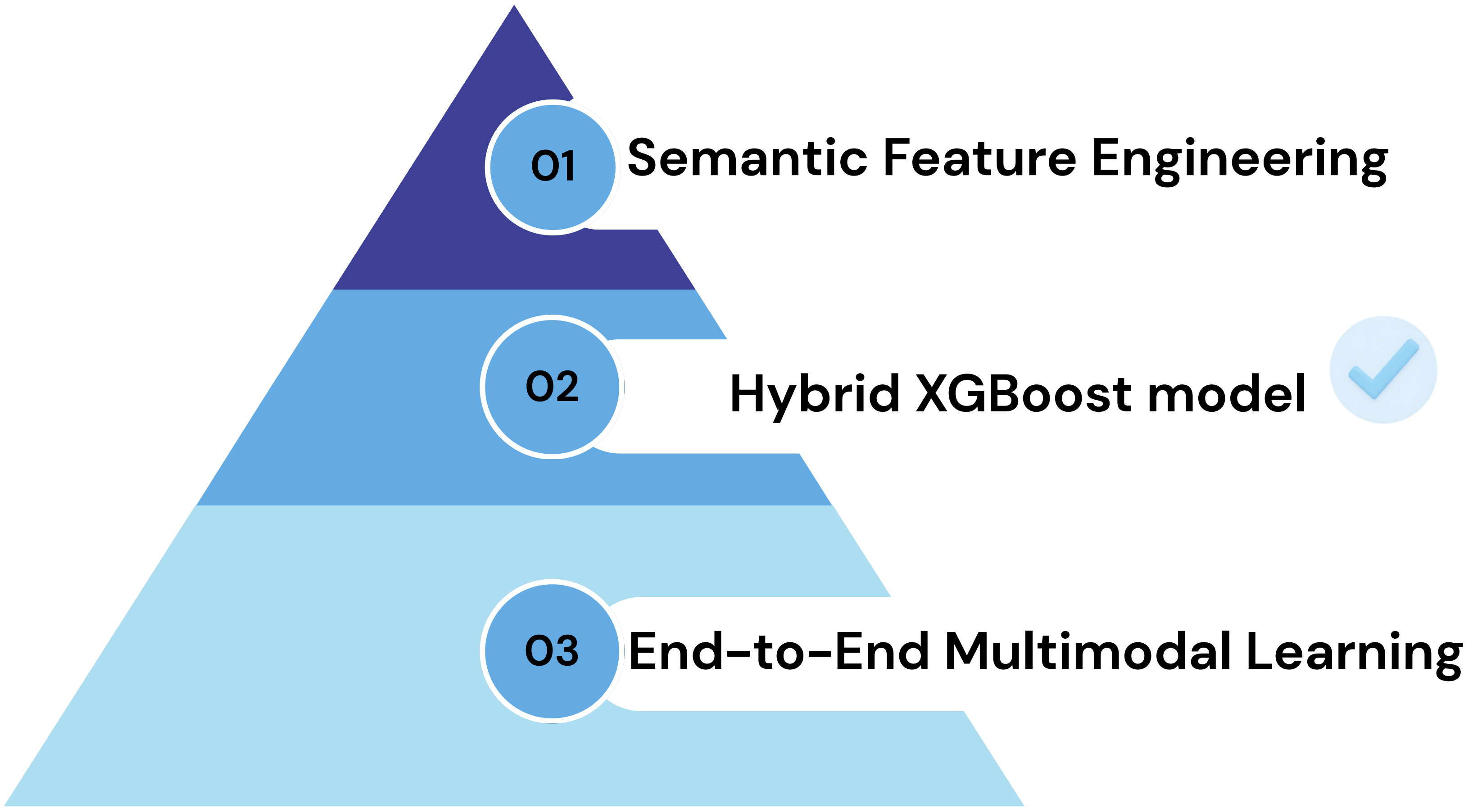
Model 1: Tabular Data Only

Random Forest
Random Forest - (RMSE): 0.1788
Random Forest - (R2): 0.8841

XGboost
XGboost- (RMSE): 0.1745
XGboost-(R2): 0.8897
Choosen Model

Gradient Boosting
Gradient Boosting -(RMSE): 0.1819
Gradient Boosting - (R2): 0.8800

Model 2: Tabular + Satellite Images.



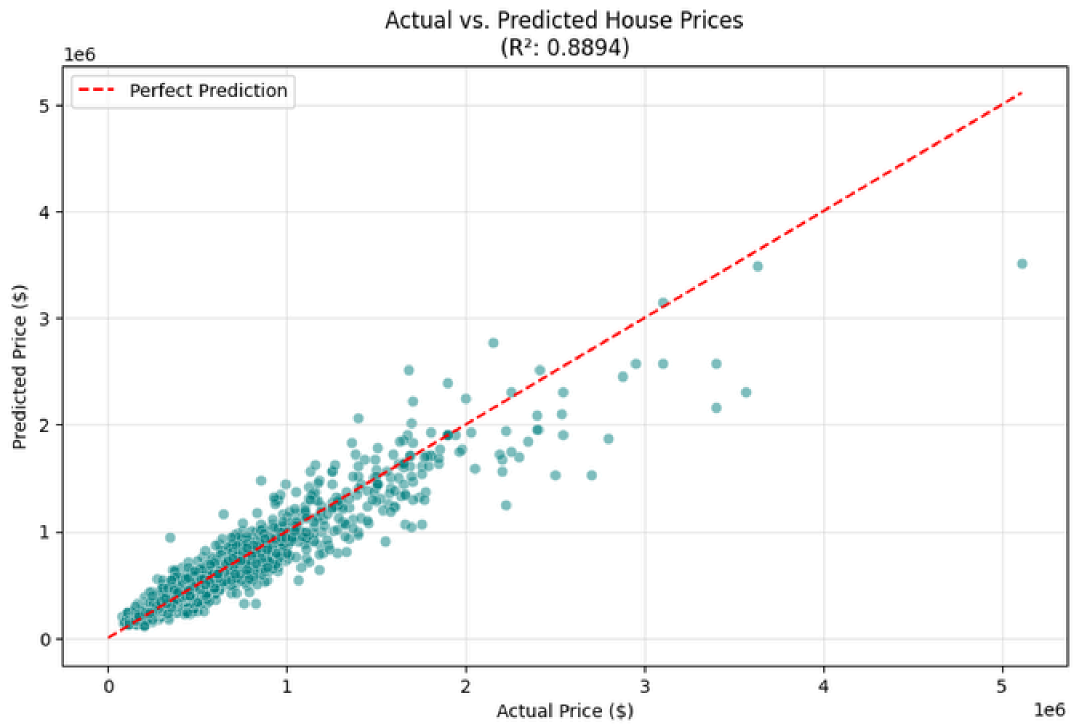
- Satellite images processed using a pretrained CNN model like ResNet to turn satellite pixels into deep feature embeddings.
- Combined tabular features with CNN image embeddings
- Built a neural network with CNN + MLP fusion
- Enables Grad-CAM visual explanations on satellite images

Phase 1 – Semantic Feature Engineering

- Extract environmental intelligence from satellite imagery to improve the tabular model.

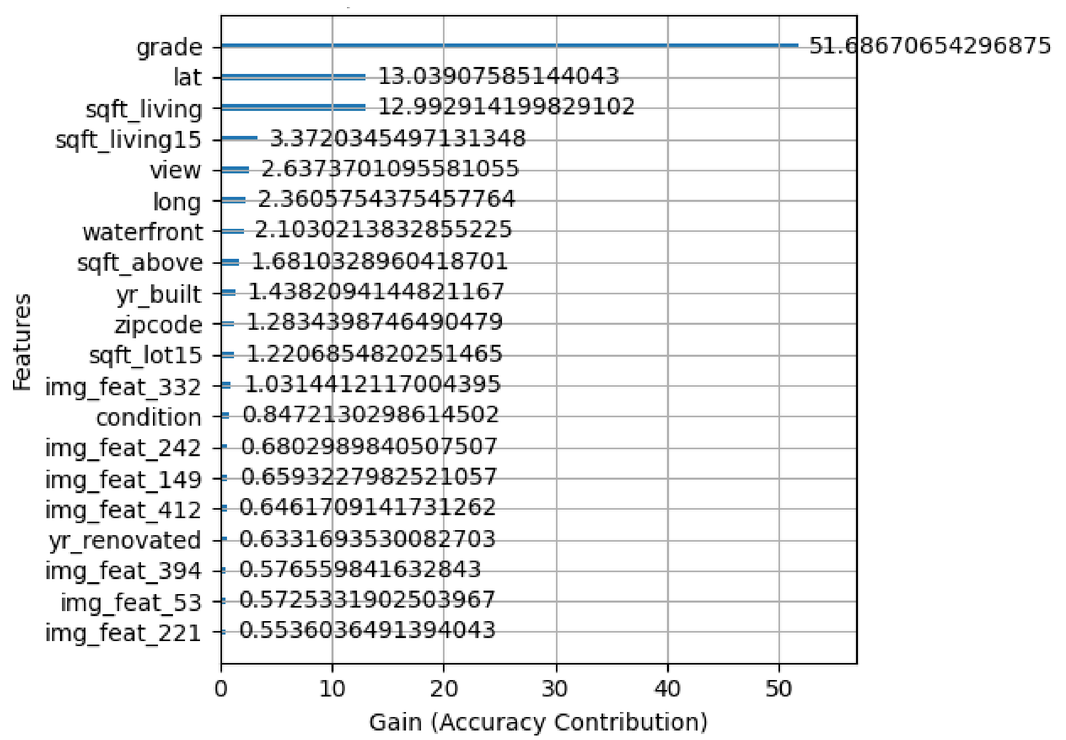
- Process: Downloaded satellite images using **Mapbox API for every property coordinates.**

- Applied Semantic Segmentation to **calculate Green Score (vegetation density) and Concrete Score (urban density)**.

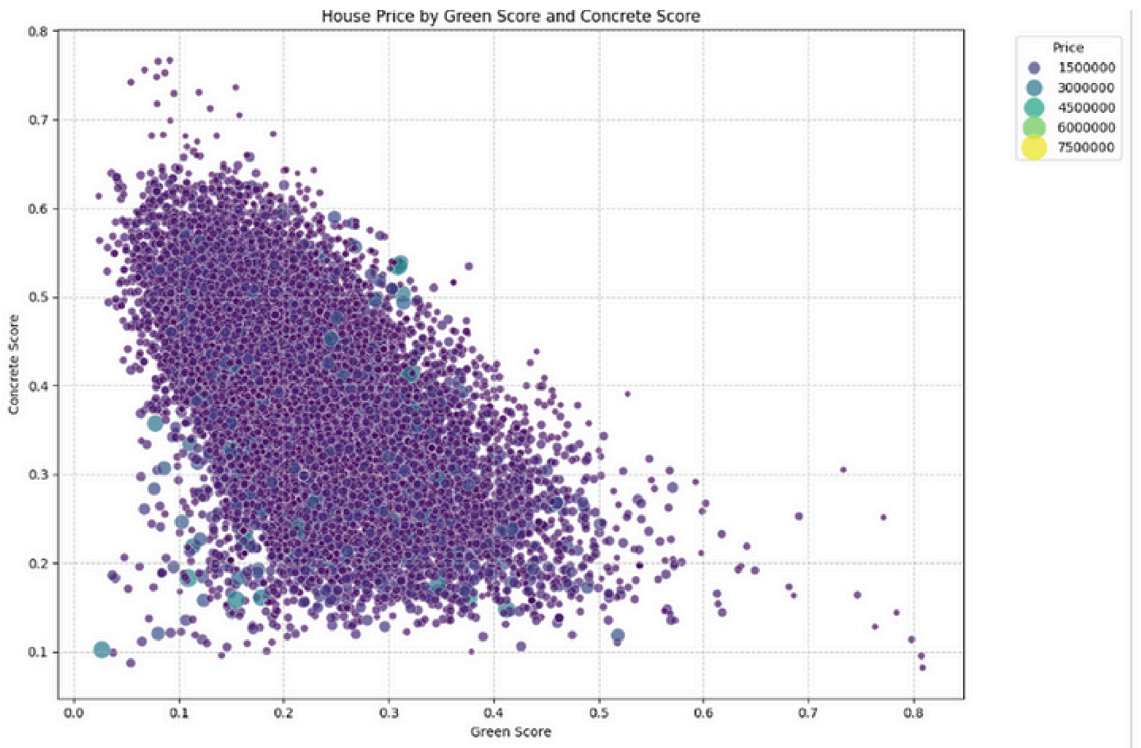


- **R-square: 0.8894**
- **MSE: 117,097.25**

Top 10 influential Parameters



House Price by Green and Concrete Score



- We utilized color-based semantic features as an interpretable baseline to quantify environmental context.

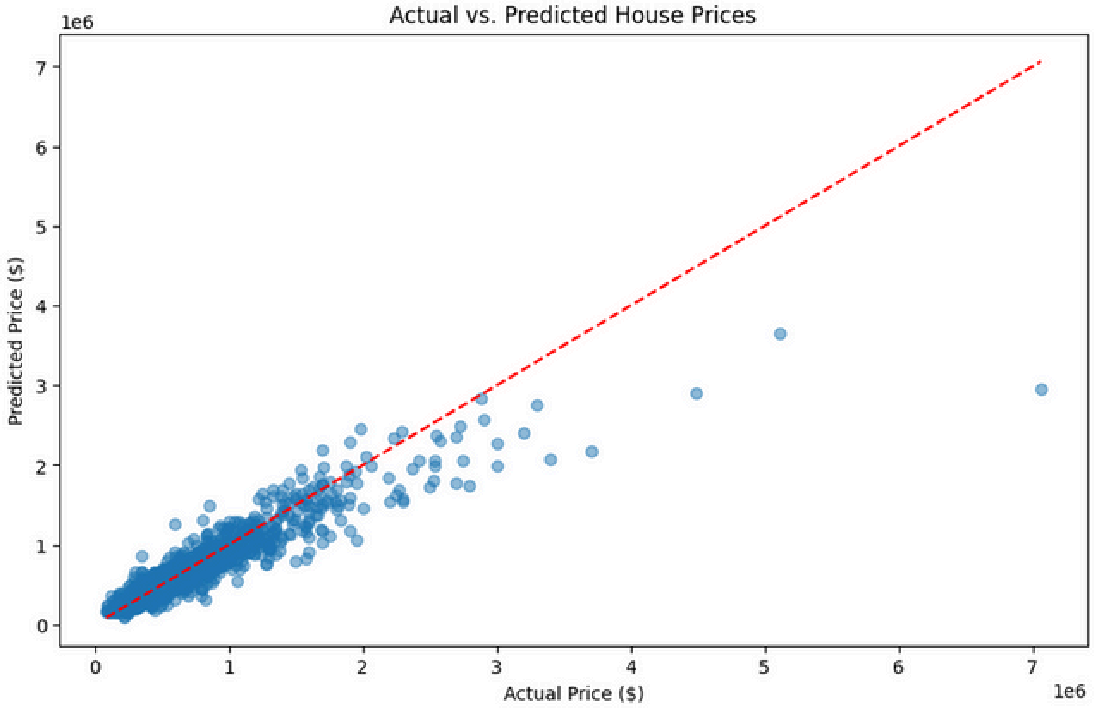
- While these scalar features alone **showed moderate predictive power,** they successfully **validated the hypothesis** that neighborhood greenery and urban density are active price drivers.

Motivated the transition to a CNN-based multimodal architecture, which moves beyond simple ratios to capture the spatial layout and complex visual interactions

Phase 2 – Hybrid XGBoost model

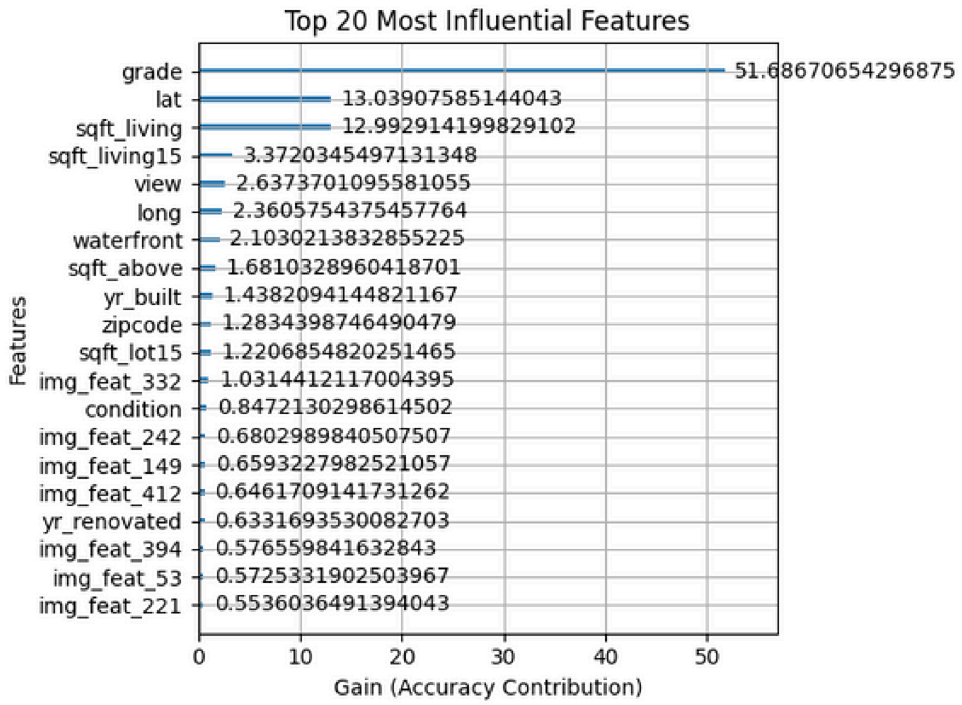
- Extract environmental intelligence from satellite imagery to improve the tabular model.

- Process: **Passed images through a Pre-trained ResNet-18.**
- Extracted the final 512-dimensional vector (Embedding) for each house.
- **Appended these visual vectors to our tabular features.**

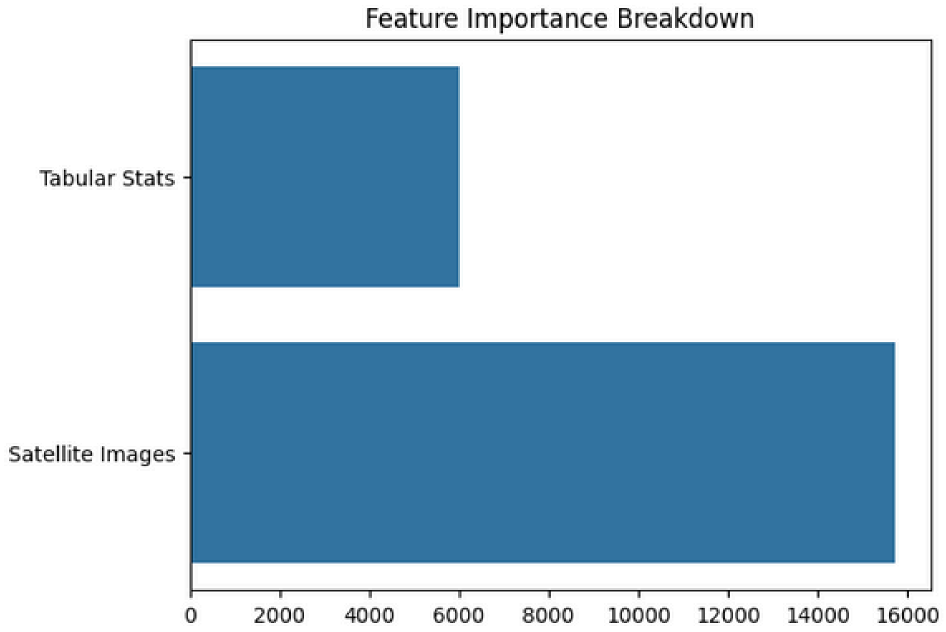


- **R2: 0.8904**
- **MSE: 0.0313**

Top 20 influential Parameters



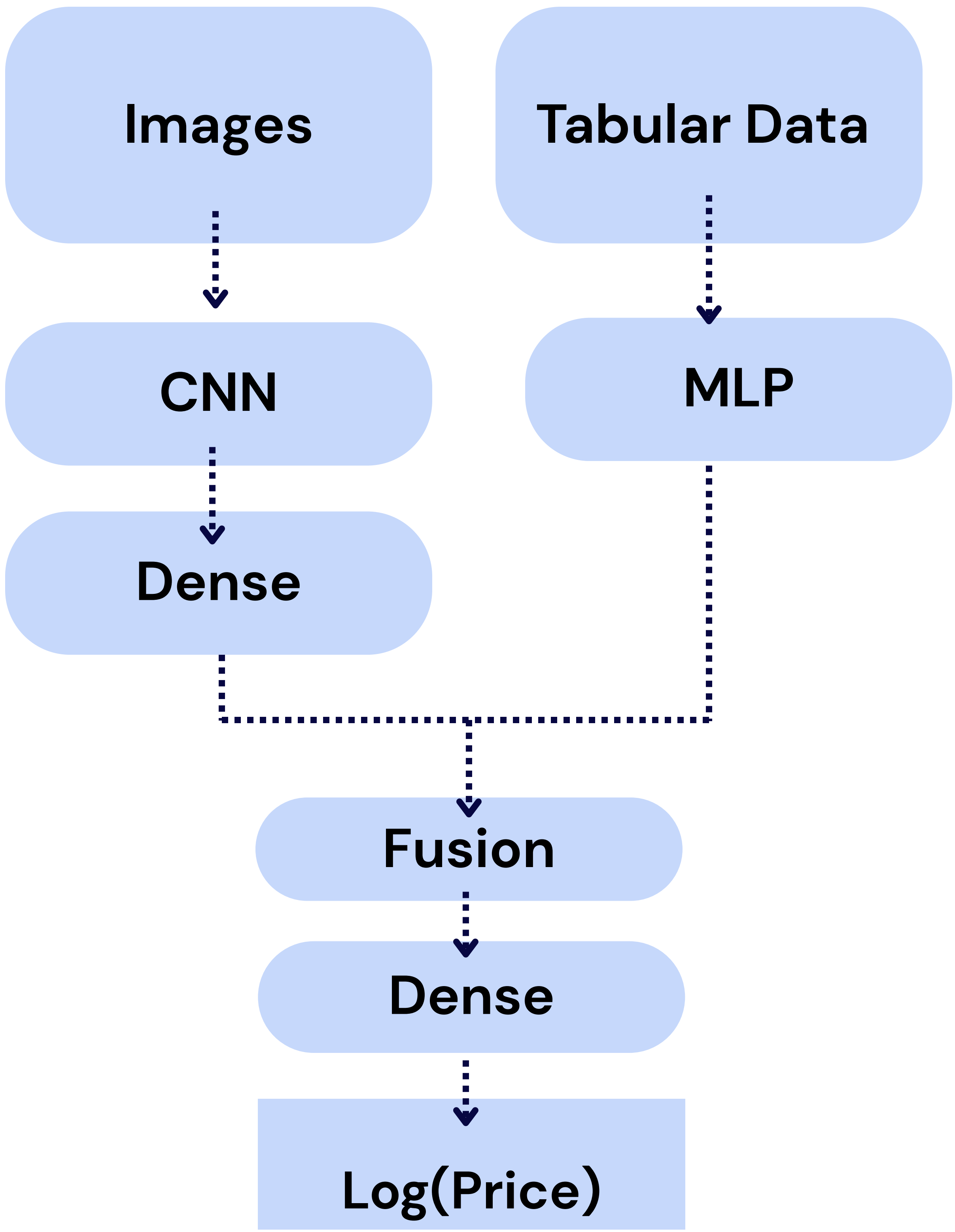
Feature Importance Breakdown



- Our feature importance breakdown reveals that **Satellite Images contribute 72.36% of the model's predictive power, compared to only 27.64% from traditional Tabular Stats.**
- **This confirms that the visual layout and environmental quality of a property are stronger predictors of price than standard metrics like square footage or bedroom count.**

Phase 3 – End-to-End Multimodal Learning

- Build a single brain that learns from pixels and numbers simultaneously.



Stage 1: Head Training (Frozen CNN)

- The ResNet-18 backbone is initially frozen to prevent the pretrained weights from being destroyed by large initial gradients.
- Only the MLP branch and the final Fusion layers are trained to align numeric data with existing visual embeddings.

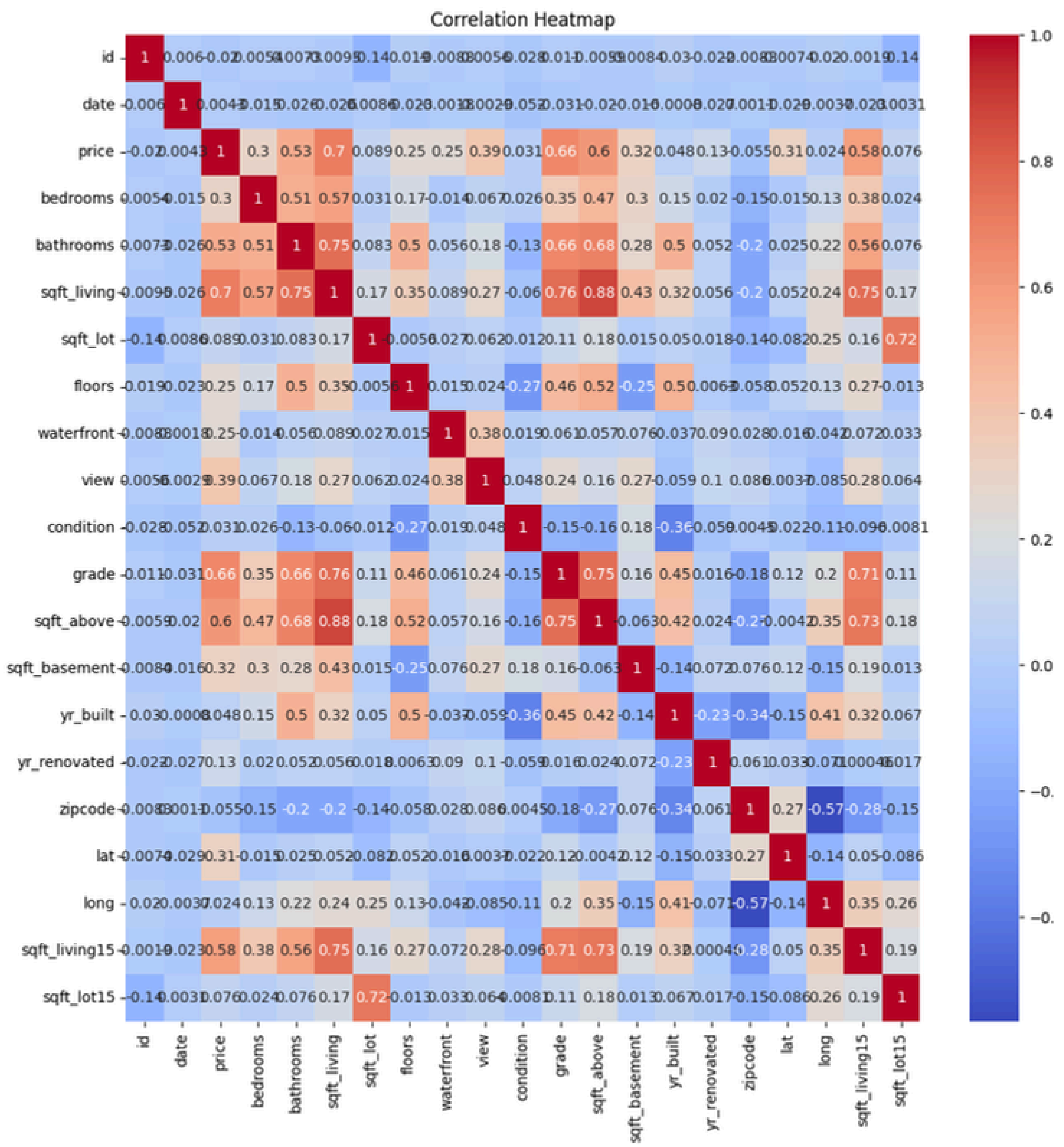
Stage 2: Fine-Tuning (Unfrozen CNN)

- The entire network is unfrozen, allowing the CNN to fine-tune specifically for property valuation tasks.
- Differential Learning Rates: A smaller learning rate is used for the CNN to preserve its feature extraction capabilities while a larger rate is used for the regression head.

Results:

```
Multimodal R2: 0.4391214062707619
Multimodal RMSE: 292136.7339483974
```

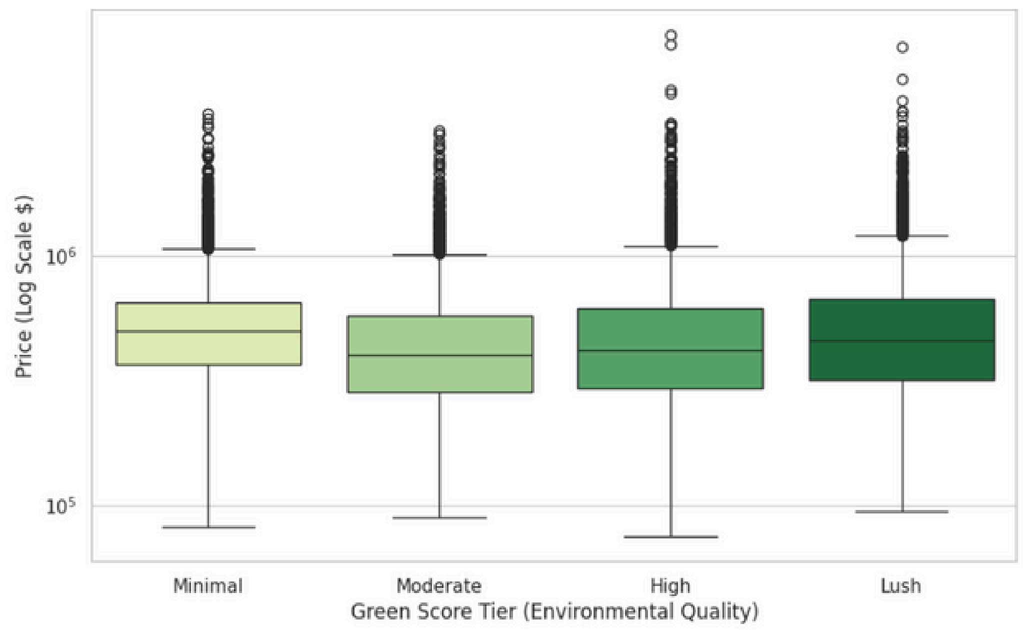
Visual Insights based on Modeling



Parameters Affecting Price Most

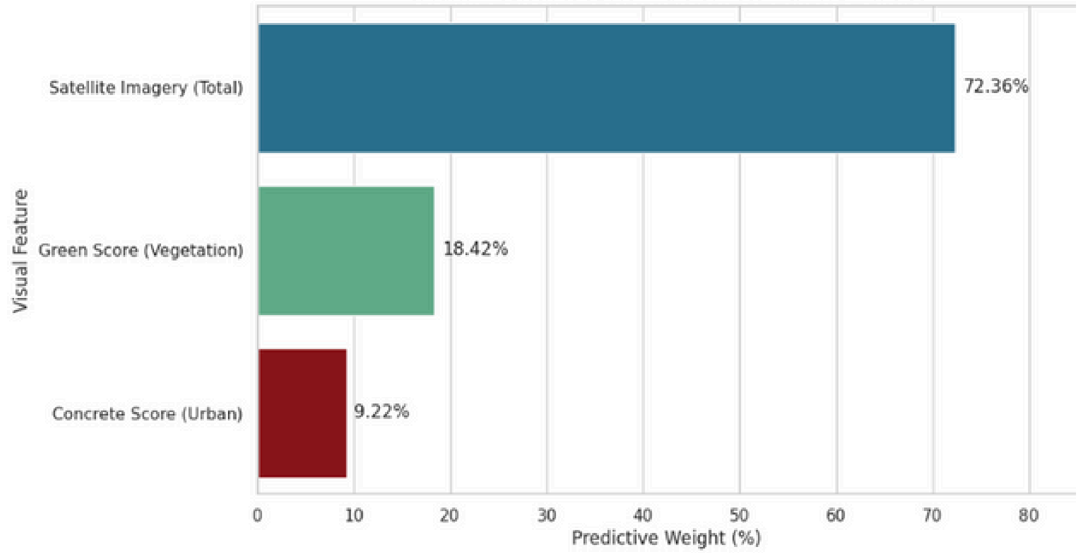
```
sqft_living      0.700933
grade            0.664266
sqft_above       0.602648
sqft_living15    0.581781
bathrooms        0.525487
view             0.390534
```

Economic Impact of Vegetation Density



- Properties located in greener environments exhibit higher median prices, though significant overlap exists across tiers.
- This indicates that vegetation contributes positively to valuation but interacts with other factors such as location and urban density, motivating multimodal modeling approaches.

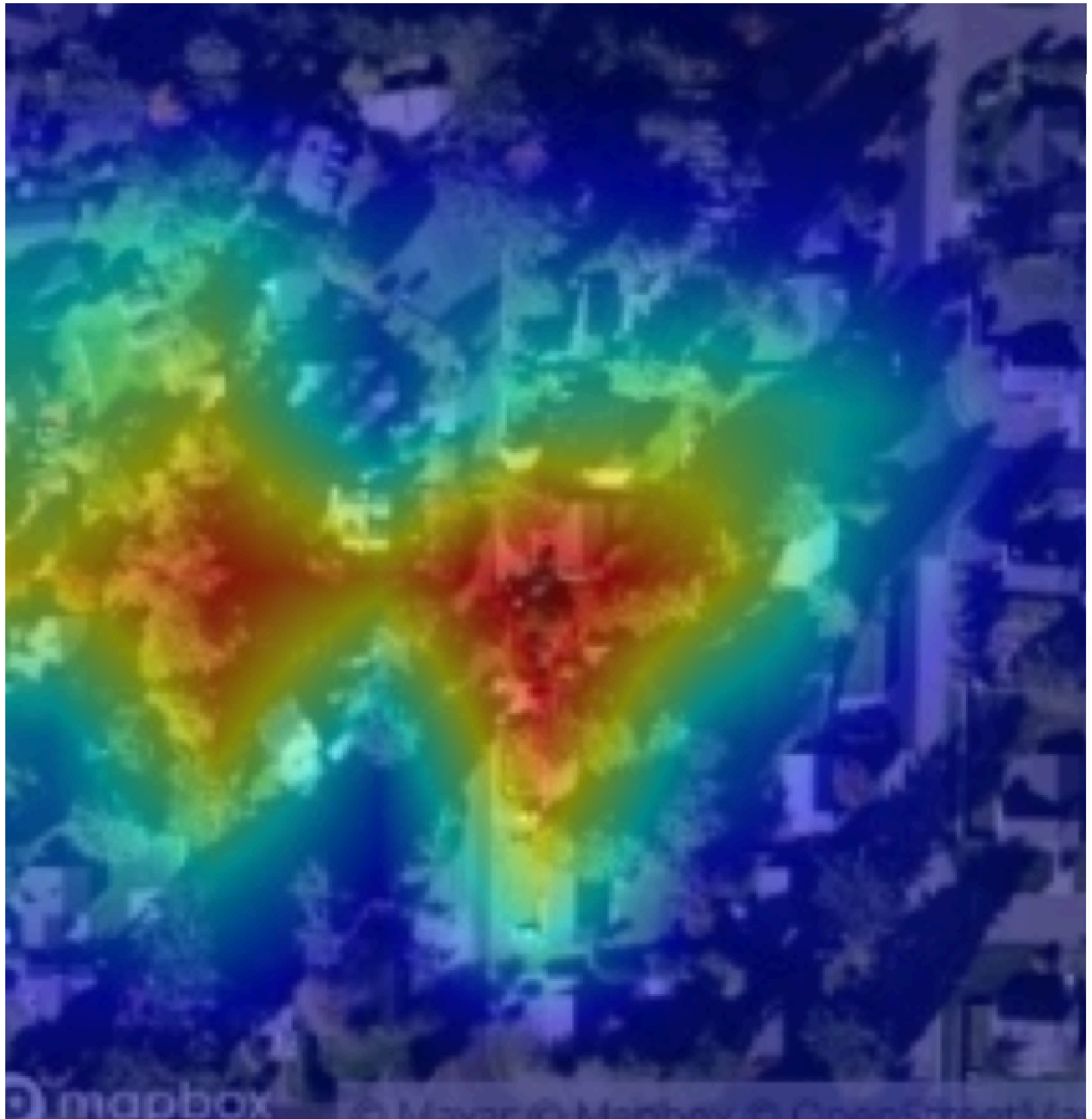
Visual Feature Contribution to Market Value



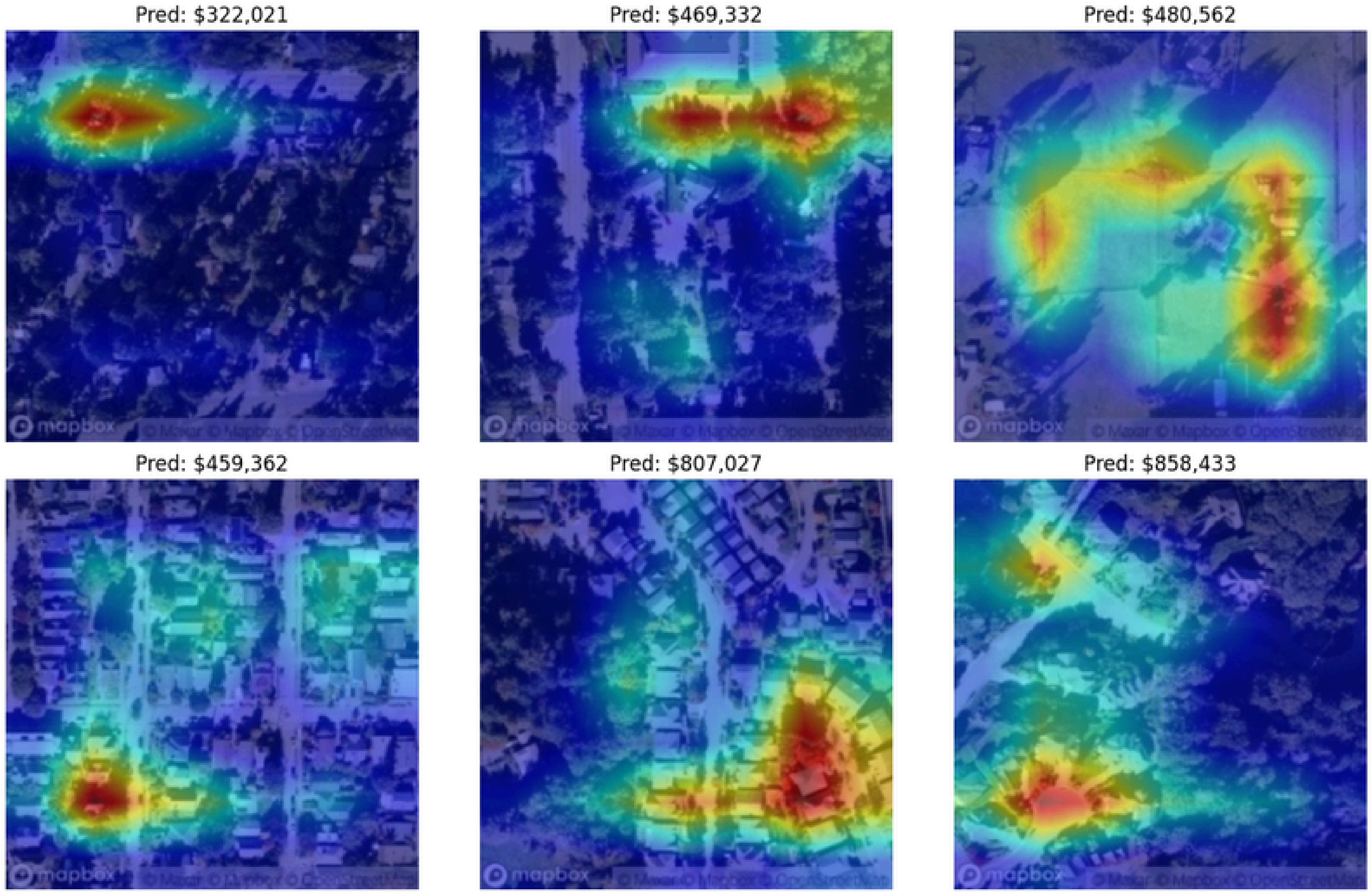
- Raw satellite imagery contributes substantially more information than hand-crafted semantic features.
- Green and concrete scores capture only a fraction of the visual signal.
- CNNs extract richer, higher-dimensional visual patterns than scalar features.

Visualized Visual Multimodal Price Prediction

- Heatmaps show high attention (Red/Yellow) on tree canopies and structural footprints.



Low Priced | Medium Priced | Expensively Priced



Model Comparison & Performance Metrics

Tabular Only

VS

Model Algorithm and Performance metrics

- XGBoost , R-square- 0.88, MSE: 0.1745

Capturing Diminishing Returns

- Often assumes a linear relationship between features and price.

Transition from "What" to "Where

- Knows what features a house has (e.g. 3 bathrooms).
- Relies on "scalar" proxies for the environment, such as a single numeric Green Score.

Information Density

- Relies on "scalar" proxies for the environment, such as a single numeric Green Score.

01

02

03

04

Tabular + satellite

Model Algorithm and Performance metrics

- Hybrid XGBoost + ResNet , R-square- 0.89,
- MSE: 0.0313

Capturing Diminishing Returns

- Better handles saturation effect", mid-range greenery (0.2–0.4) often correlates with higher suburban premiums than extreme forestation.

Transition from "What" to "Where

- Understands where the house is positioned in its environment, allowing the model to perform a digital appraisal.

Information Density

- Captures high-dimensional visual features like neighborhood quality and architectural style, that numbers alone cannot represent.

The Road Ahead

What More Can Be Done?

- **Higher Resolution Imagery:** Moving from standard satellite tiles to high-resolution aerial photography would allow the model to detect smaller value-driving features like swimming pools, roof quality, or solar panels.
- **Integrating Time-Series imagery:** could help the model understand neighborhood trajectory, such as recent construction or maturing greenery, which significantly impacts long-term investment value.
- **Robustness to Seasonality:** Training with images from different seasons would ensure the "Green Premium" is calculated based on permanent vegetation rather than temporary seasonal blooms.
- **Expanding Geographic Diversity:** The current model is specialized in specific clusters; training on diverse urban, suburban, and rural biomes would improve the current validation metrics.

How This Model Helps the Company

- **Captures Hidden Value:** Traditional models rely on the "bones" of a house, but this model can now quantify the "Lush" greenery premium, identifying properties that command higher median prices due to their environmental quality.
- The firm **no longer needs manual inspectors** to rate neighborhood quality; the CNN Branch (ResNet-18) automatically extracts these features from 224x224 satellite tiles.
- **Investment Risk Mitigation:** Using Grad-CAM, the company can provide clients with visual proof of a valuation, highlighting the specific tree canopies or lot characteristics that drove the price.
- **Market Trend Identification:** Analyzing the interaction between land use (concrete) and vegetation allows the firm to identify emerging Luxury Zones before they are reflected in standard market reports.

Thank you!