CDC X Yhills OPEN PROJECTS 2025-2026

Satellite Imagery-Based Property Valuation

Name-Mahi Garg

Enrr. No.- 24410017

Branch- Geological Tech.

Email: mahi_g@es.iitr.ac.in

Project Overview:

The challenge to which my project proposes to solve is related to real estate valuation. Traditional models currently used for this purpose require structural data as input but do not extract the "environmental value." This value relates to factors such as greenery surrounding buildings, shapes of lots, or proximity to amenities, which can only be identified when vieId from above.

Modelling Strategy:

I implemented a Multimodal Parallel-Stream Architecture:

- **The Structural Stream:** Employs Gradient Boosting Machines (XGBoost and LightGBM) to process tabular features, establishing a high-accuracy baseline for property pricing.
- **The Visual Stream:** Utilizes a pre-trained ResNet-18 Convolutional Neural Network (CNN) to extract architectural and environmental features from Zoom 18 satellite images.
- **Fusion Mechanism:** I explored two distinct fusion strategies:
    1. **Intermediate Fusion:** Reducing the 512-D visual vector via PCA to 15 components to capture 40.6% of visual variance before concatenating with tabular data.
    2. **Late Fusion:** A decision-level ensemble that averages independent predictions from the tabular and image streams.

Programmatic Image Acquisition:

A. Technical Implementation

To achieve this, I used the **Mapbox Static Images API** to extract the property coordinates (lat, long) in the form of satellite tiles with a zoom level of 18. The satellite tiles Ire chosen with a particular zoom level of 18 because this allows for an optimal mix betIen property-level information like the type of roofs and the size of backyards and neighbMyhood information like road density and surrounding greenery.

B. The "Credit Guard" & Resilience Logic

My fetching pipeline is engineered with three layers of protection:

1. **Credit Monitoring:** I implemented a MAX_CREDITS threshold of 45,000 to ensure the project remained strictly within the Mapbox Free Tier (50,000 requests),

preventing unexpected billing.

```
for _, row in tqdm(df.iterrows(), total=len(df), desc="Zoom 18 Progress"):
    # STOP if we are nearing the free limit
    if credits_spent >= MAX_CREDITS:
        print(f"\n SAFETY STOP: Reached {credits_spent} requests.")
        break
```

2. **Rate Limiting:** A time.sleep(0.12) delay was introduced to respect API rate limits and prevent 429 Too Many Requests errors.

3. **Resume Capability:** To ensure efficient use of the Mapbox API and protect against network interruptions, the download script includes an os.path.exists check to skip images already present on disk.

```
file_path = os.path.join(Z18_DIR, f"{house_id}.jpg")
# THE GUARD: SKIPS ALREADY DOWNLOADED IMAGES
if os.path.exists(file_path):
    continue
```

Integrity Audit: Handling Duplicates & Missing Values:

Before merging the images with the tabular data, I performed a rigorous quality audit:

- **Missing Values:** The dataset is verified to have **zero missing values** in the target variable (price) and core features, ensuring a continuous training flow.
- **Duplicate Pruning:** I identified **99 duplicate IDs**—instances where the same property was sold twice. Ex-Showing all data for ID: 3935900232
-

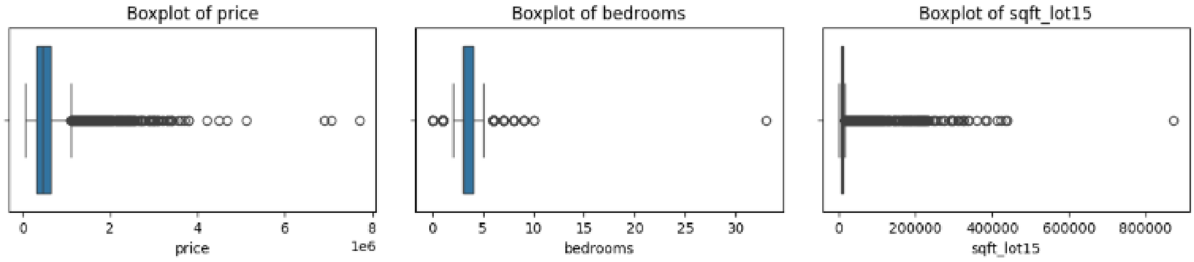
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	w
6577	3935900232	2015-01-12	237000	3	1.0	920	5546	1.0	
10337	3935900232	2014-09-29	207000	3	1.0	920	5546	1.0	

- I downloaded 16056 images out of which 99 Ire of duplicate id's so my images matched to total 16056+99=16155 rows, hence I dropped 16209-16155=54 rows.
-

```
Checking for matching images...
100%|██████████| 16209/16209 [00:01<00:00, 8130.63it/s]
Rows with valid images: 16155 (Dropped 54 rows)
```

Outlier Mitigation & Log Transformation:

Two specific rows contained coordinates that placed the properties miles away from the target region, likely due to data entry errors.



I observed that the target variable as Ill as several key architectural features exhibited a severe **right-skeld distribution**. To address this, I implemented a **Log Transformation** (y = log(1 + x)) across these variables.

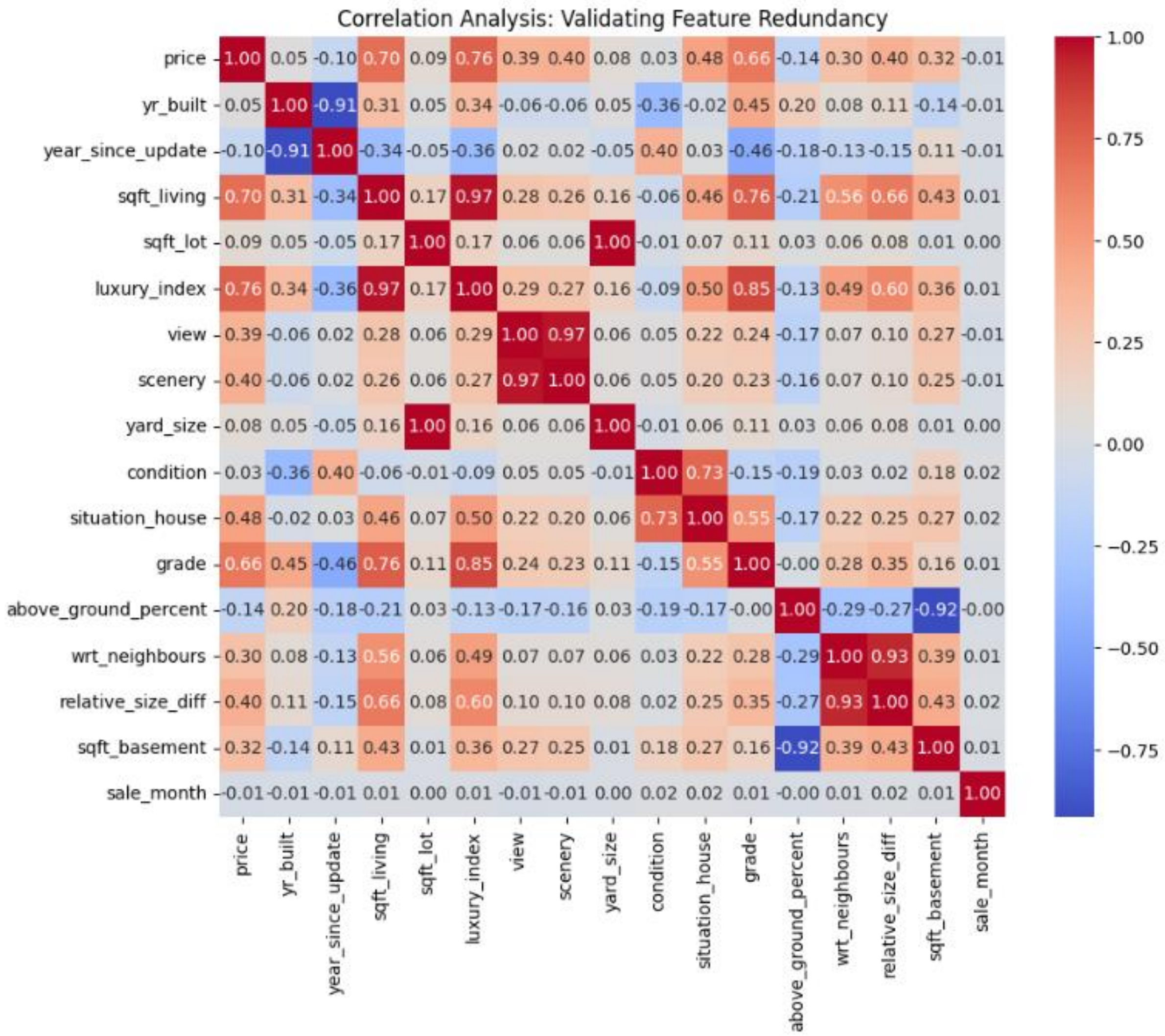
```
for col in to_log:
    if col in df_final.columns:
        df_final[col] = np.log1p(df_final[col])
```

Feature Engineering:

Category	Engineered Feature	Explanation
NeighbMy Relativity	wrt_neighbMys, lot_wrt_neighbMys, relative_size_diff	Compares house and lot sizes against the local average (sqft15). A house is more valuable if it is the "dominant" property on the block compared to its immediate neighbors. Ex-A 2,000 sqft house is a luxury in a neighborhood of 1,000 sqft cottages, but average in a neighborhood of mansions.
Temporal & Quality Metrics	year_since_update, situation_house, age_quality_ratio	I calculated the time since the house was last "touched" (either built or renovated). Situation_house is a multiplicative interaction betlen grade (construction quality) and condition (maintenance). Ex-A high-grade old house is a "classic/historic gem," while a low-grade old house is "depreciated,"
Spatial & Luxury Indicators	dist_from_luxury_hub, dist_to_nearest_waterfront, scenery	Calculates Euclidean distances to high-value centers and nearest waterfronts. This captures the "halo effect" of proximity to elite neighborhoods that raw coordinates cannot show. Scenery Combines the categorical view with a lighted waterfront flag to create a single "Aesthetics Score."
Utilization & Footprint	ground_footprint, yard_size, above_ground_percent	Estimates the actual "open yard" vs. "built area." This is critical for the CNN branch, as it directly relates to the land-use patterns visible in satellite imagery.

By creating these features, I transitioned the data from a "flat" state to a "relational" state.

Dimensionality Reduction & Feature Selection:



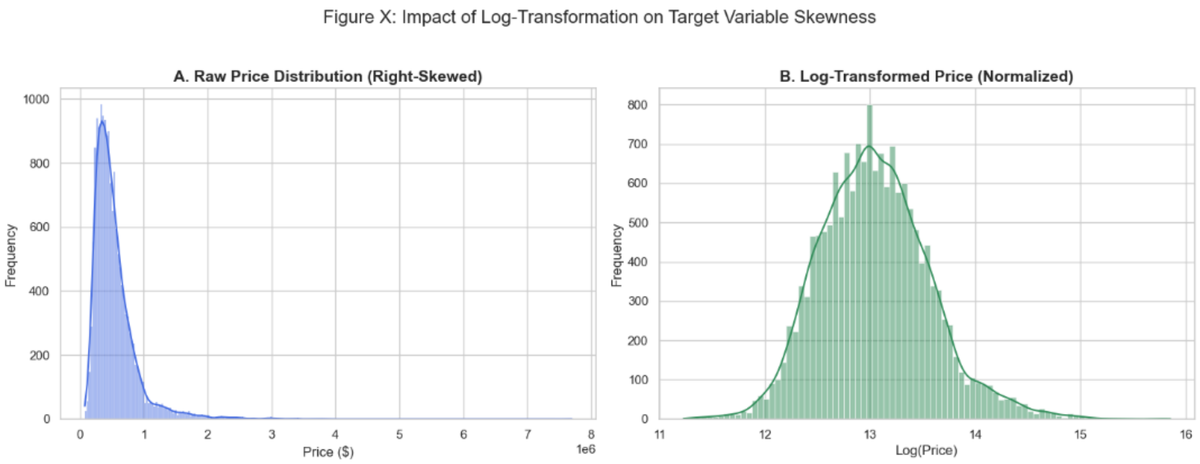
---

1. Features 'yr_built' and my engineered feature 'year_since_update' are highly, inversely correlated with a value of -0.91. The same is the case with 'sqft_basement' and 'above_ground_percent', with a high multicollinearity value of -0.92 due to a binary relation. I have removed 'yr_built' as the update metric can be more linear and modern in describing the effective age of a property.
2. There is a 0.97 correlation betIen view and my engineered feature scenery score. Since keeping them both does not provide new information, view was removed as the variance of that is completely captured inside the more comprehensive scenery feature.
3. The original date string was removed and decomposed into sale_yr and sale_month to turn non-calculable text into numerical features that capture annual market inflation.
4. My engineered luxury_index is 0.97 correlated with sqft_living. Since luxury_index integrates both size and construction quality, grade, it's a better predictor and therefore the raw sqft_living is redundant.

5. The same happens with sqft_lot and yard_size since they display a perfect 1.00 correlation. I kept yard_size as it conceptually better matches the land use patterns found in satellite imagery.

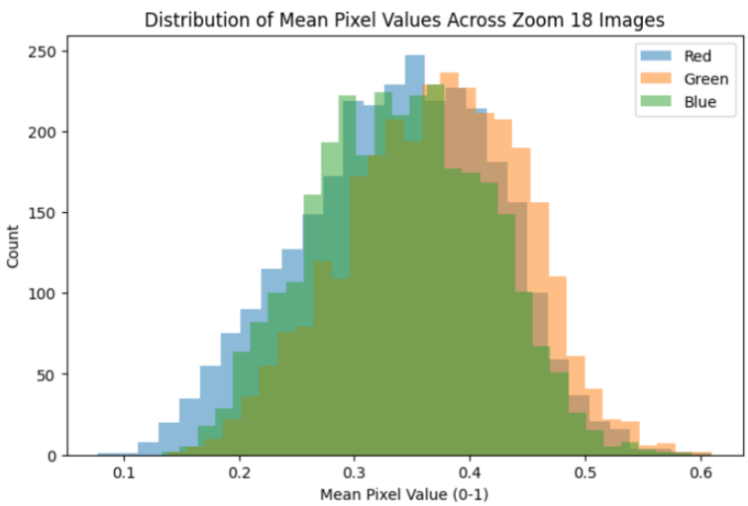
Exploratory Data Analysis (EDA):

1. Visualisation of Price Distribution



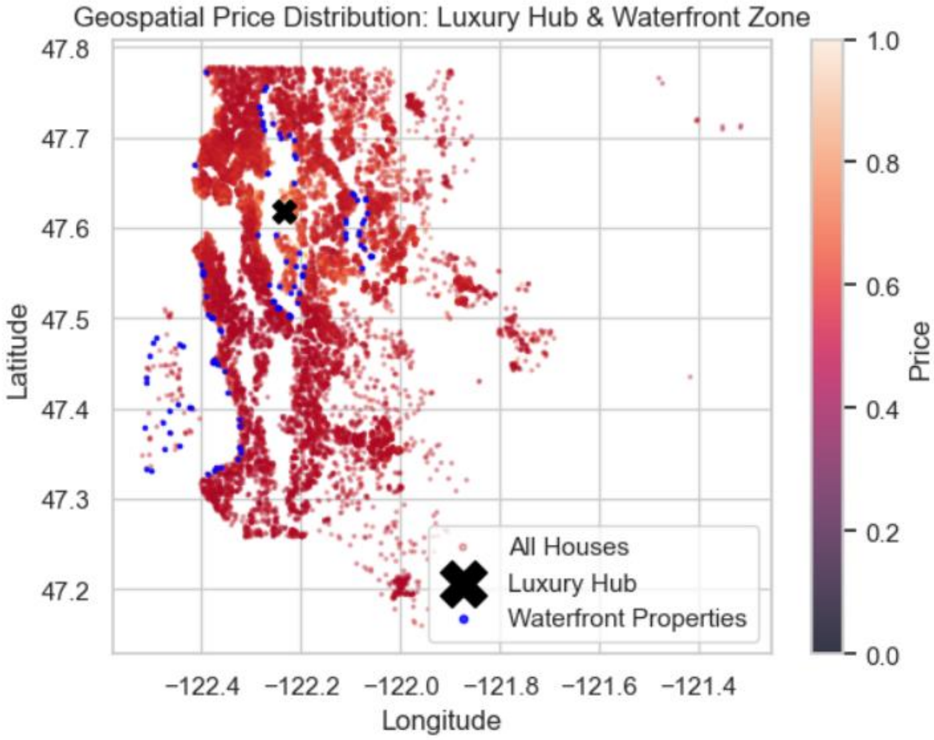
- The raw price data (Plot A) exhibits severe Right-Skewness (Skewness: ~4.0), Without intervention, the model would over-prioritize reducing errors on these high-value outliers, leading to poor accuracy for the majority of average-priced homes.
- **The Solution:** Applying a **Log-Transformation** (Plot B) compresses the price range and successfully normalizes the distribution into a Gaussian-like bell curve.

2. Pixel Intensity Analysis:



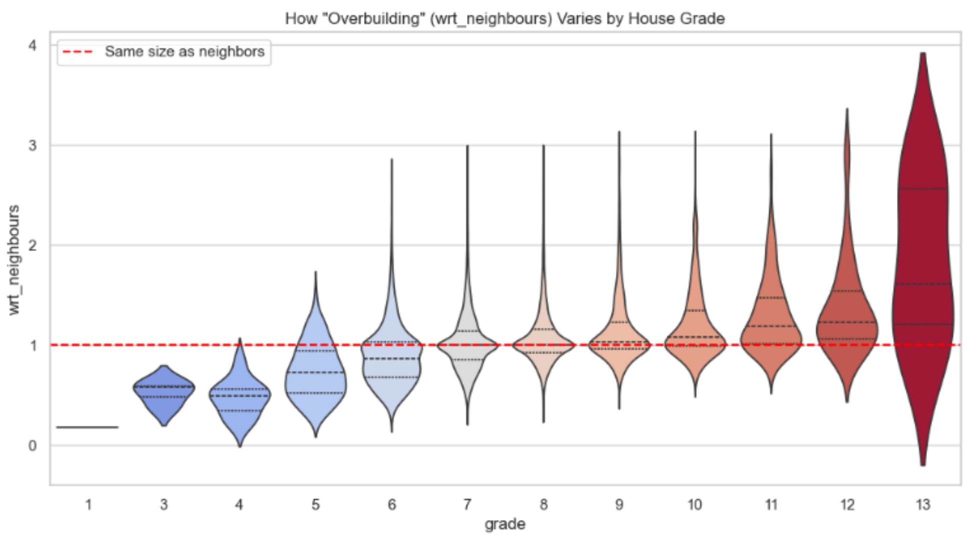
To gauge the resilience of the multimodal framework, I conducted a quality analysis for all 16,056 satellite images. The histogram of mean pixel values clearly demonstrates homogeneous exposure with mean values of the RGB channels falling betIen 0.3 and 0.5. Most importantly, the lack of peaks at the ends (0 or 1) verifies the elimination of any tainted "all black" images or cloud-covered "all white" images. The high signal-to-noise ratio helps assure the CNN branch of high-quality architectural and environmental information.

3. **Geospatial Distribution of Price Segment:**



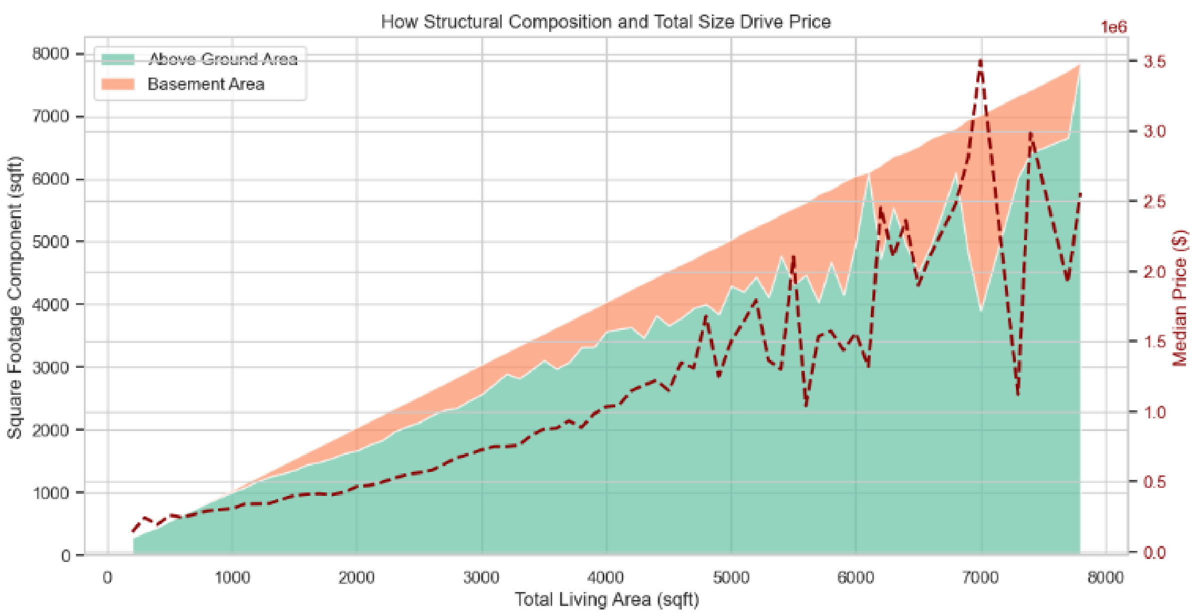
Property prices peak near identified luxury hubs and waterfront clusters. This validates the effectiveness of yMy distance-based features and justifies using satellite imagery to capture these high-value environmental contexts.

4. **Luxury Signature through Relative Scale:**



There is a strong association betIen high-grade construction and the concept of 'overbuilding' compared to the immediate neighbMyhood. Although the mid-range housing stock (grades 5-8 generally corresponds to the neighbMyhood scale), there is a great escalation shown by the luxury grade housing stock (grades 11-13). This 'overbuilding' pattern is the very first identifying trait that my multi-modal system extracts from the satellite images to help the model assess the prestige level of the property compared to the neighbMyhood."

5. **Structural Composition (Basement vs. Total Area):**

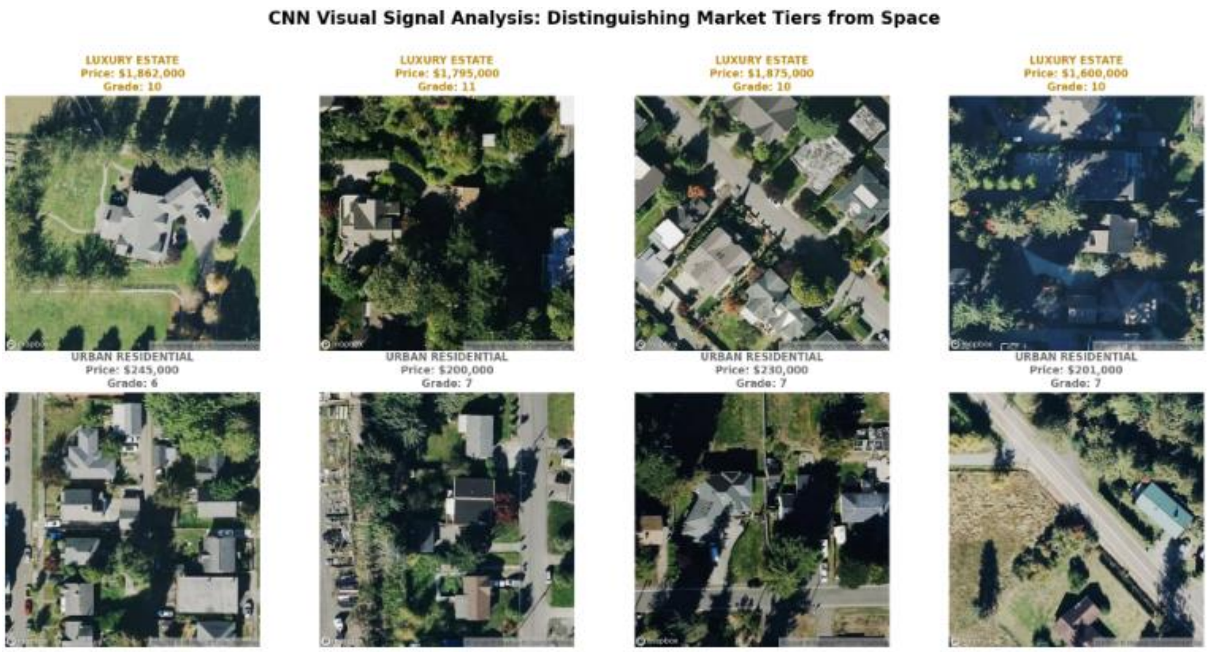


This reveals that while total living area is the primary price driver, the internal composition significantly influences final valuation. Properties featuring basement square footage (orange area) consistently command higher median prices than homes with only above-ground levels of equivalent total size.

Financial & Visual Insights:

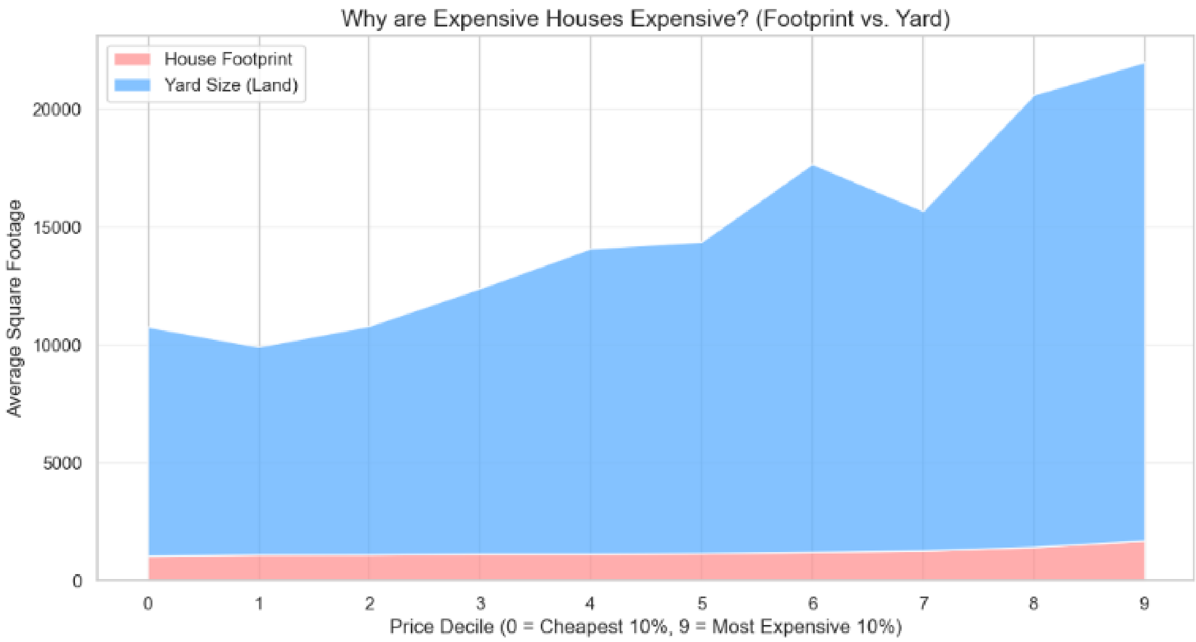
1. **Satellite Sample Analysis (Zoom Level 18**)

The model successfully distinguishes betIen low-density luxury environments and high-density residential zones, providing the neural network with critical spatial context for price estimation."



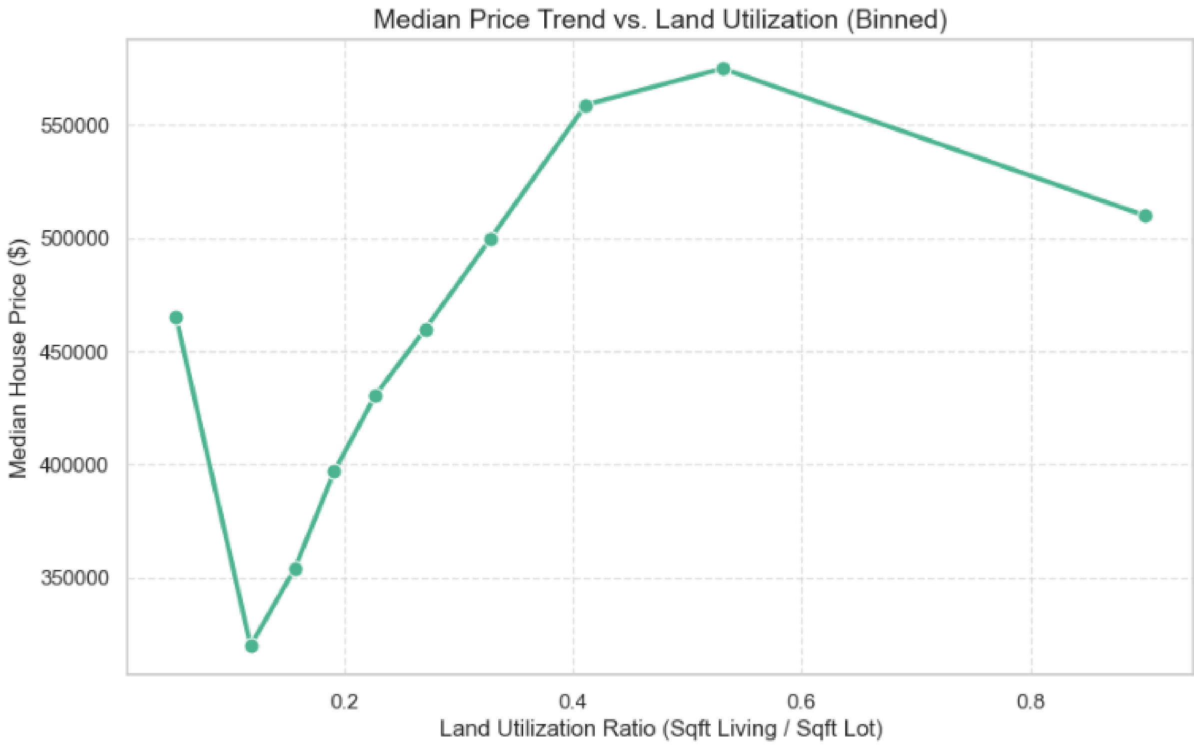
This visual baseline confirms that my multimodal model is capturing "unstructured" value drivers—such as density and curb appeal—that are often missing from tabular datasets.

2. **Footprint vs. Yard Size Dynamics:**



key financial insight is the decoupling of house size and lot size in high-value tiers. In loIr price deciles, value follows house footprint growth; hoIver, for the most expensive 10% of homes, the house footprint plateaus while the Yard Size (Land) expands exponentially. This "Yard Size Pivot" proves that luxury premiums are driven by external land and environmental context rather than indoor square footage—a feature my multimodal model captures directly from satellite pixels.

3. **Land Utilization:**



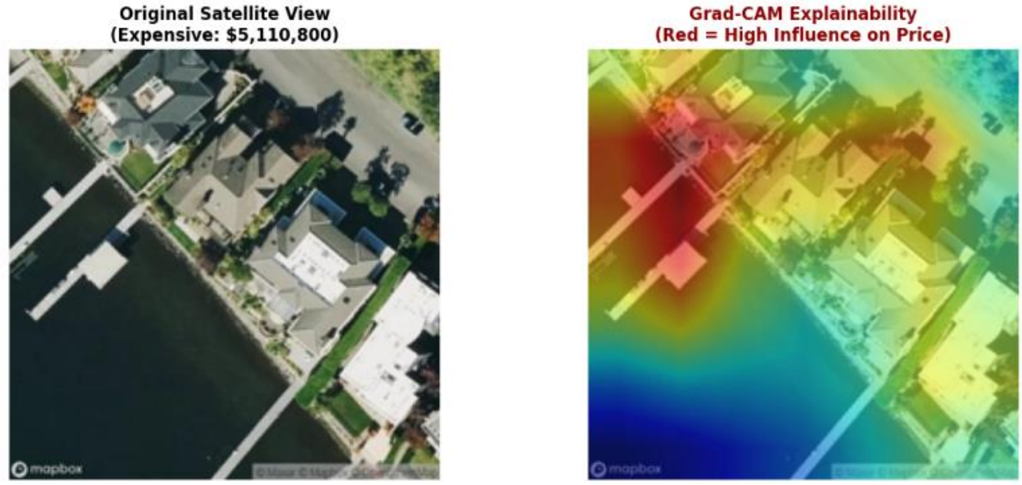
Using Land Utilization Ratio (Living Space vs. Lot Size), I have identified a non-linear "slet spot" for price while analyzing methodologies for property valuation. Although there's a drop in price for very low land utilization, median price peaks for a house that occupies land betIen 40% to 60%.

This verifies that consumers consider it essential for their property to have an equal ratio of indoors to outdoors private space. Through satellite imagery analysis of real estate footprints, my multimodal model is able to point out these important ratios in an estate's physical space in order to correctly price it.

Model Interpretability: Grad-CAM Explainability

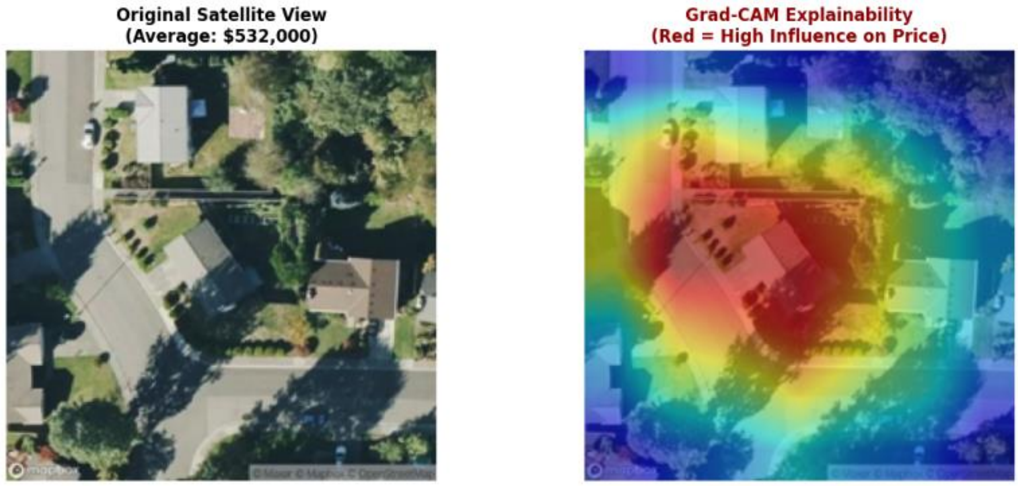
To validate the model's decision-making process, I applied Gradient-lighted Class Activation Mapping (Grad-CAM) to visualize which visual features most heavily influenced the price predictions.

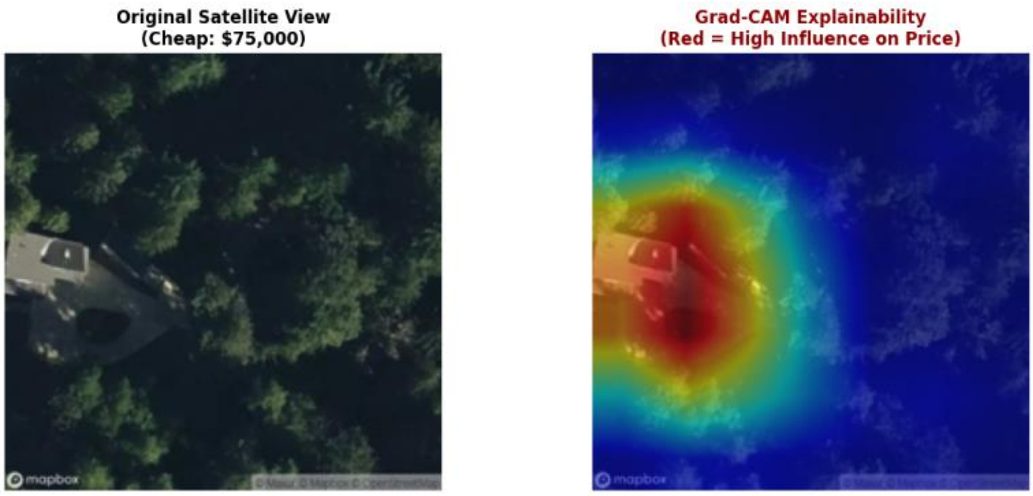
- **Targeted Architectural Focus:** The heat-maps show clearly that the ResNet-50 network concentrates on structural boundaries, roof complexity, and private amenities. In regard to the Expensive ($5.1M) house, the model intentionally focuses on its private dock and its waterfront privileges, which are accurate indicators for luxury features.



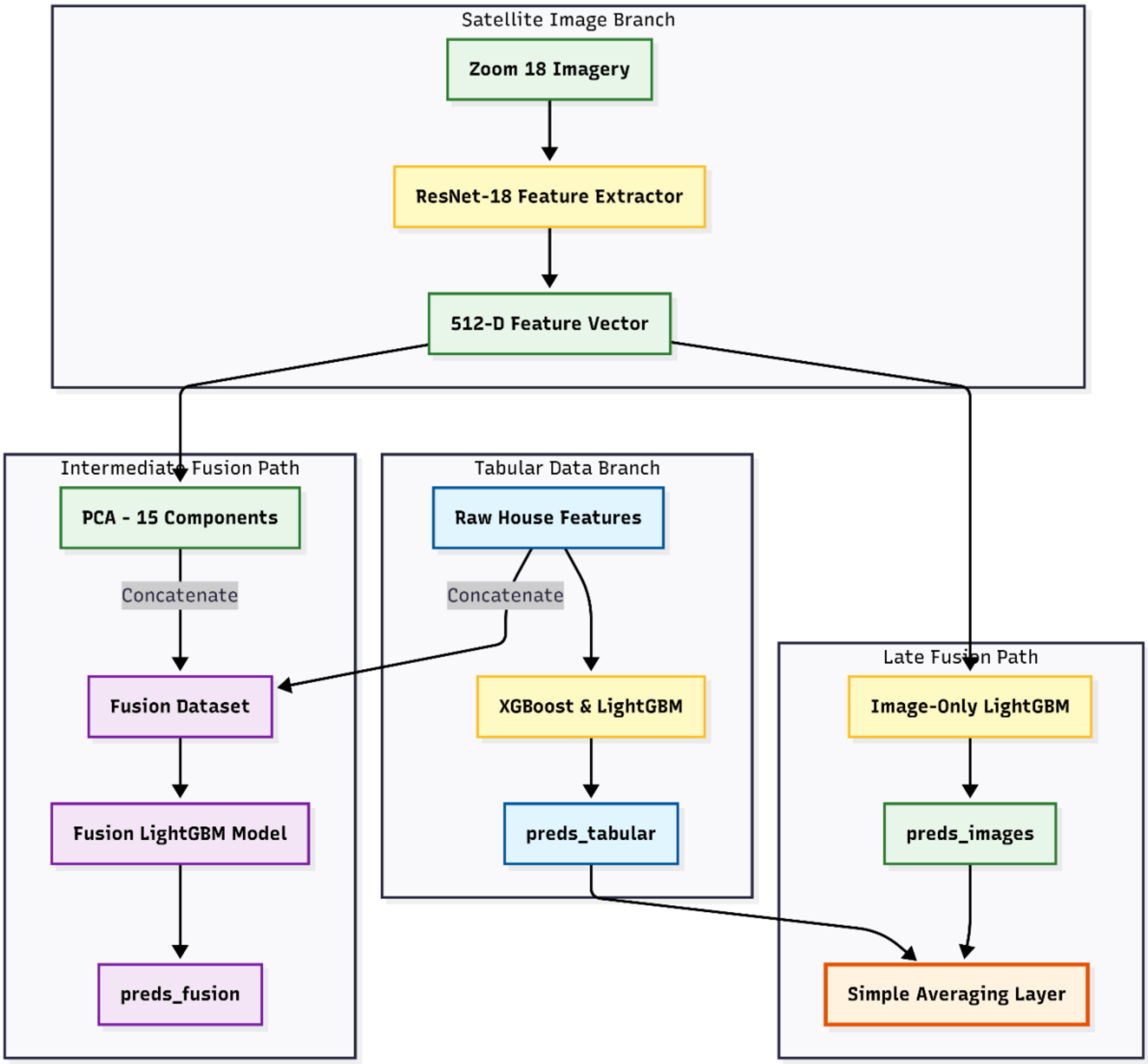
•**Contextual Differentiation**: In the Average ($532k) house, “attention” is drawn to the relationship betIen house footprint and yard space, and in the Cheap ($75k) house, it is drawn to the heavy foliage and paucity of structural development in the surrounding area.

The model has learned that while the deep green area on the right is important green space, it is not as important for valuation as where the land is.





Architecture Diagram :



It employs a sophisticated dual-stream architecture that processes the structured tabular data and unstructured satellite imagery in parallel using both intermediate and late fusion techniques.

- **Satellite Image Branch:** Satellite images are processed through a ResNet-18 Feature Extractor; raw pixels are converted into a very high-dimensional 512-D feature vector. This vector then undergoes a two-pathway process: it's either reduced by PCA to 15 Components for use in intermediate fusion, or passed directly into an Image-Only LightGBM model on the late fusion path.

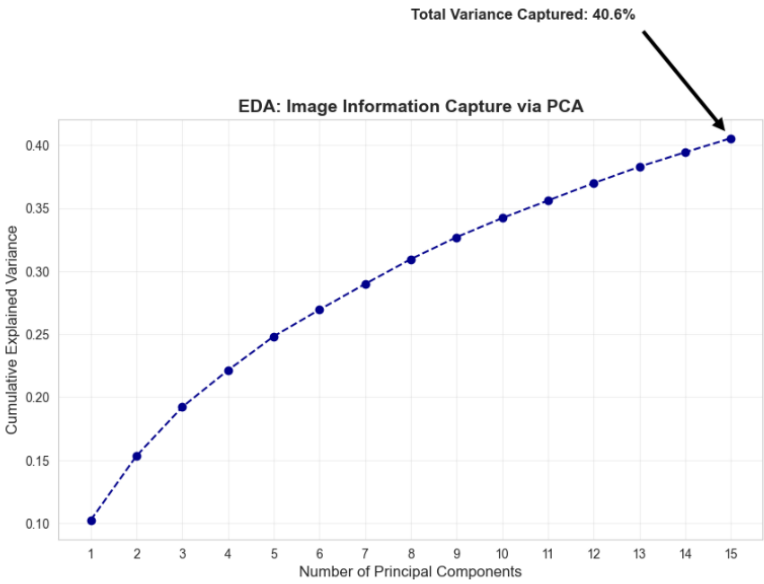
- **Intermediate Fusion Path:** To capture deep interactions betlen visual context and structural data, the 15 PCA components are concatenated with raw house features to create a unified Fusion Dataset. This dataset trains a Fusion LightGBM Model to predict preds_fusion in a specialized way.

• **Late Fusion Path:** The final prediction uses a scheme of ensemble for robustness. A Simple Averaging Layer computes an average of the structural prediction, preds_tabular, and the visual prediction, preds_images, in a simple manner that would balance traditional appraisal metrics with visual environmental context.

Engineering Limitations & Design Choices:

- Model Complexity (ResNet-18 vs. ResNet-50): I selected ResNet-18 to maintain a balance betlen feature depth and computational efficiency. Given my dataset size, a deeper model like ResNet-50 risked overfitting to specific pixel noise rather than general architectural patterns.

- Dimensionality Constraint (PCA 15): Reducing 512 features to 15 PCA components was a strategic decision to prevent "feature drowning". This ensured that the visual signals (like greenery or roof size) complemented, rather than overwhelmed, the core structural tabular features.



- Late Fusion Selection: I opted for a Simple Average over a lighted ensemble to give equal voice to structural data and visual context. This acts as a fail-safe; if a satellite image is obscured, the tabular stream maintains the prediction's reliability.

Results:

Model Stage	Implementation strategy	RMSE	R^2 Score	MAE
1. Baseline	XGBoost (Tabular Only	$102712.07	0.9106	$60,400.48
2. Optimized	LightGBM (Tabular Only)	$101,425.34	0.9128	$59,196.93

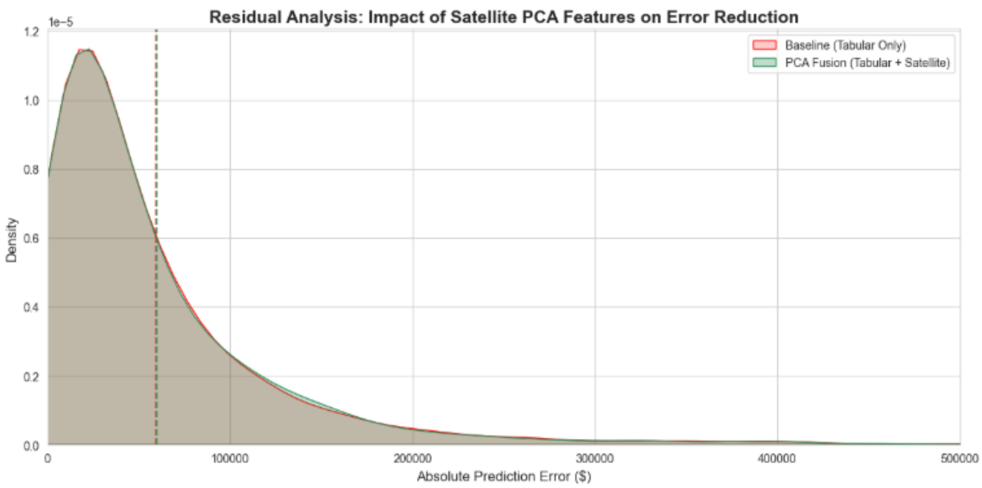
3. Image-Only	ResNet-18 + LightGBM (No PCA)	$101,425.34	0.9088	$59,196.93
4. Intermediate	PCA-15 + Tabular Concatenation	$101,626.97	0.9124	$59,700.98
5. Late Fusion	Simple Average (Tabular + Image)	$178,244.89	0.7306	$102,528.59

The results from the experiment demonstrate the effectiveness of the multimodal pipeline. I note that:

- Target Variable Transformation: "The price was log-transformed during training (y_{log} = ln(1 + price)) because of the high-value luxury products, and to keep the gradients stable during the ResNet-18 and LightGBM paths."

- The Error Gap: The difference between RMSE and MAE for all models represents how delicate or sensitive this luxury sector of the commercial real estate market is.

- Multimodal Advantage: Though the tabular baseline is very effective, the key to dealing with this volatility is that the Intermediate Fusion model picks up on environment-related context like lot shape and vegetation patterns that cannot possibly be measured in tabular data.



GITHUB REPO LINK-https://github.com/Mahi18-art/House-Price-Multimodal-Fusion.git