Data Science PS

Yhills

Satellite imagery based property valuation.

Submitted by : Madhav Bansal | EE-2Y | 24115094

Overview

The main aim was to predict the property price using both structured property attributes and unstructured spatial context through a multimodal regression framework. Through spatial context we mean processing satellite images, for procuring the images by latitude and longitude coordinates we have used mapbox API.

Given structured property attributes likes covered area, number of rooms are important factors but one of the most important factor of property valuation is location of property. This satellite imagery tackles that problem by training the model on the basis of image features like road density, greenery, surrounding environment etc.

For the image modality, a pre-trained Convolutional Neural Network (CNN) is used as a feature extractor to transform satellite images into compact, high-level embeddings that encode neighborhood characteristics.

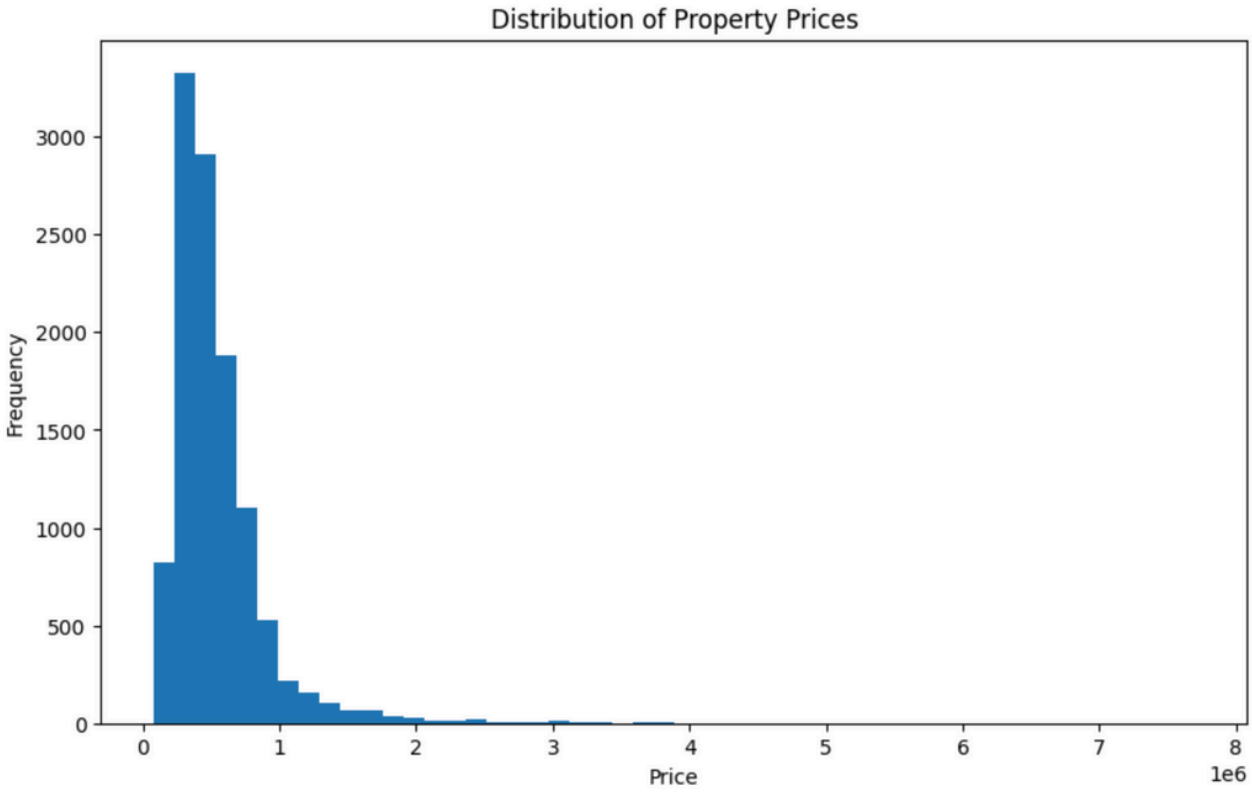
The extracted image embeddings and tabular features are then concatenated at the feature level (early fusion) to create a combined representation. This fused feature vector is used as input to a gradient-boosted regression model (XGBoost), which is well-suited for handling heterogeneous features and capturing non-linear relationships.

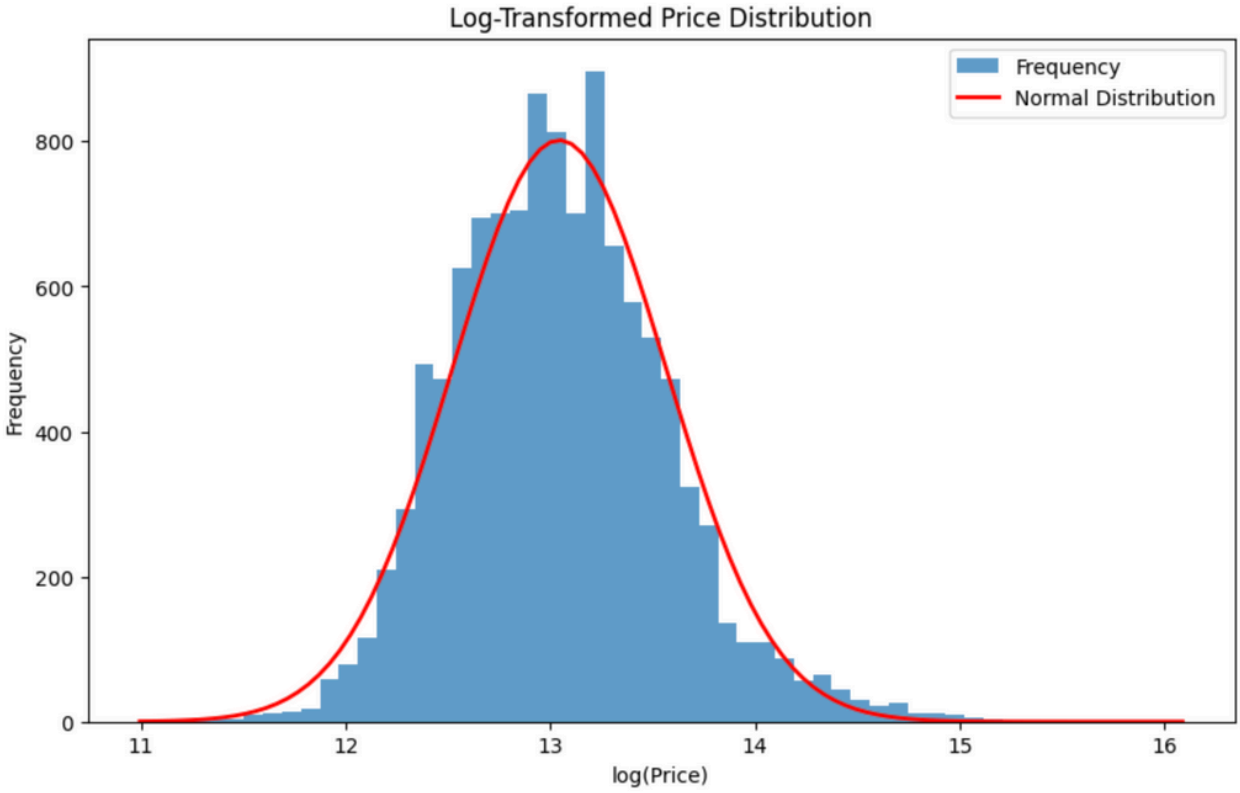
The model learns complex interactions between visual neighborhood context and structured property attributes to predict the target variable, property price.

Overall, this strategy allows the model to incorporate both observable characteristics and latent spatial factors, resulting in a more expressive and accurate prediction framework compared to traditional tabular-only approaches.

Exploratory Data Analysis

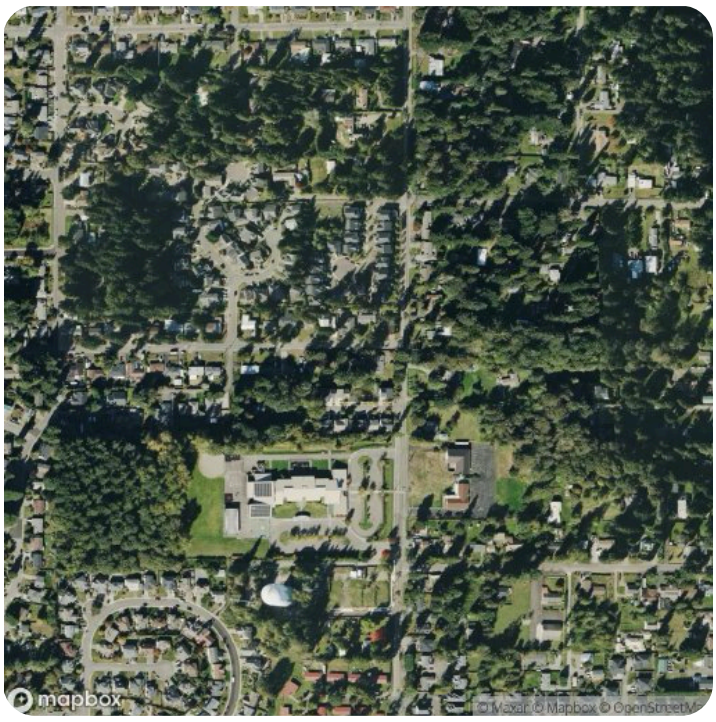
Below we have plotted the property price distribution, in first graph it is clearly observed that distribution is not normal this is because there are small number of properties with large values. To adjust this accordingly we have log transformed the price to get closer normal distribution graph. This increases the stability and improve predictions.

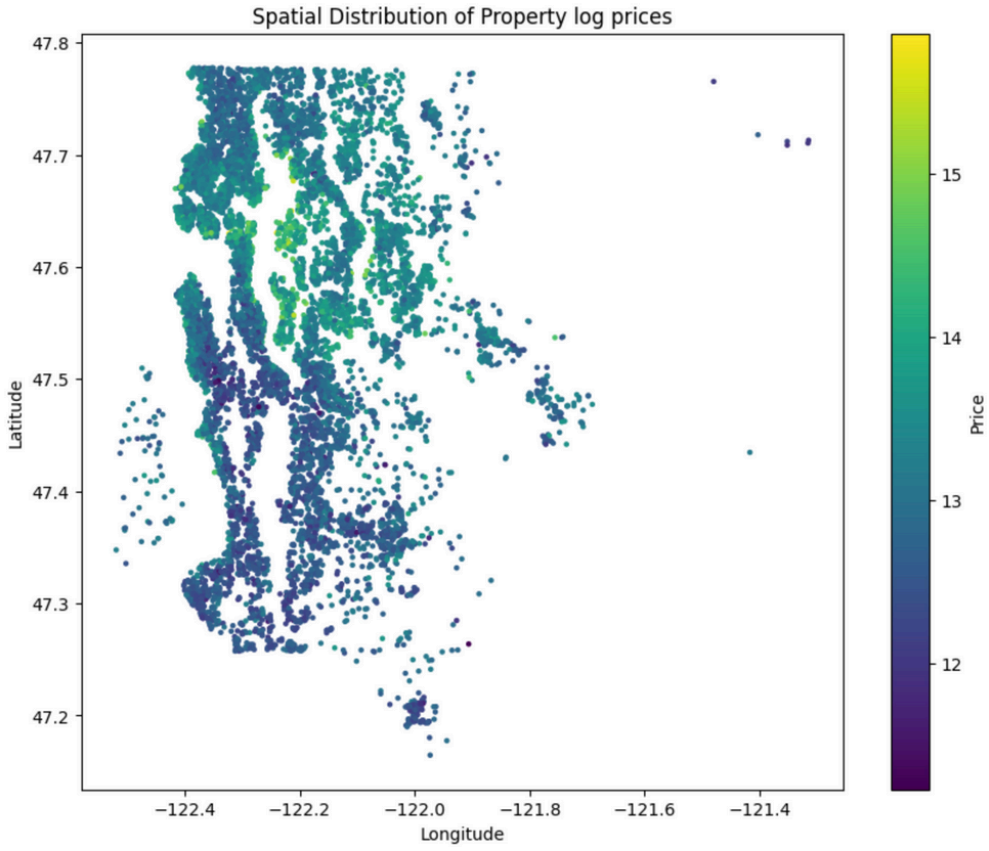




Exploratory Data Analysis

- We have used Mapbox API for satellite images.
- Zoom is 12x with size 256 to obtain clear image, for better feature extraction.
- Through image processing we tend to differentiate like High-priced properties tend to be located in well-connected and dense urban regions, while lower-priced properties often correspond to sparse or peripheral areas.





- The spatial distribution of log transfer prices shows clear geographic clustering, reinforcing the importance of location dependent features and spatial context.

Financial insights

While the multimodal model achieves strong predictive performance, understanding which visual characteristics influence property value is critical for interpretability and real-world relevance.

How visual features train the model ?

Satellite images are processed by a CNN, which learns latent visual features representing:

- vegetation cover
- road density
- building compactness
- urban vs. suburban structure

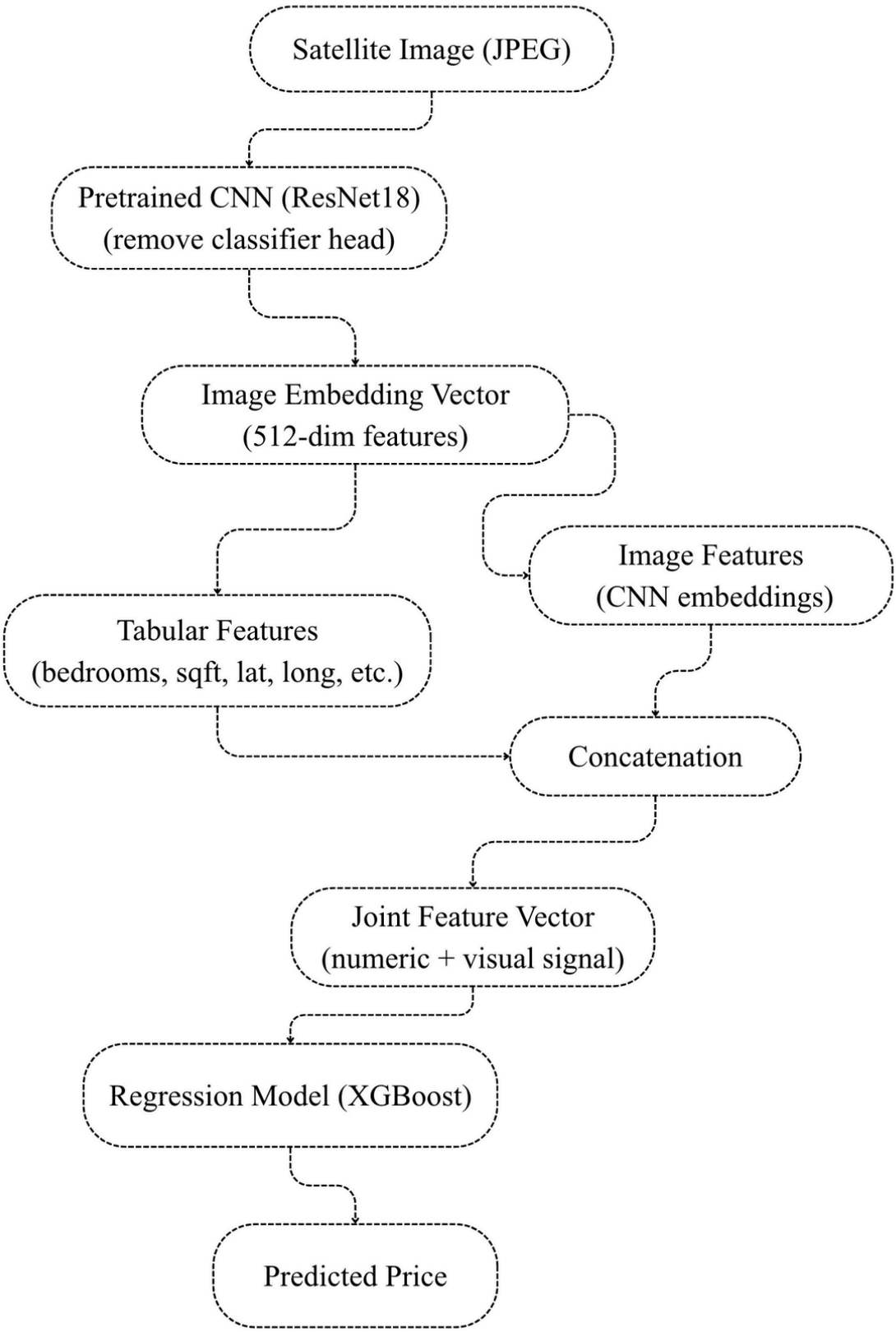
These features are not manually labeled, but emerge naturally from data.

Since CNN features are abstract, we analyze them indirectly using feature attribution and visual grouping.

1. Model performance is compared between a tabular-only baseline and a multimodal model incorporating satellite imagery. The observed improvement in predictive accuracy indicates that visual information contributes meaningful economic signal beyond structured data.
2. Neighborhoods with visible tree cover, parks, and open green spaces are consistently associated with higher predicted prices.
3. Areas exhibiting well-structured road layouts and clear connectivity tend to command higher prices.
4. Regions dominated by continuous concrete structures with limited open space are generally associated with lower predicted prices.

From a financial perspective, the model implicitly learns to price neighborhood quality alongside physical property attributes. Visual features such as greenery, spatial organization, and accessibility act as value multipliers or discounts on top of baseline property characteristics. Importantly, these visual signals often amplify or dampen the effect of traditional variables like area or location, highlighting strong cross-modal interactions.

Architecture Diagram



Results

Performance for Image + Tabular features

I split the training data into 70:30 split for training and validation dataset.

Then trained the data on 70 and validated it to rest dataset, metrics are as follows :-

R Square Score: 0.887

Mean Absolute Error: 68205.5

Performance for Tabular features

The identical R^2 observed for tabular and multimodal models indicates that CNN-derived image embeddings were largely non-informative for price prediction in this setup.

This is likely due to the use of ImageNet-pretrained features, which do not strongly encode satellite-specific land-use semantics, combined with the dominance of high-signal tabular variables (e.g., sqft, latitude/longitude).

Consequently, the model effectively ignored image features during training.

Things we tried:

We implemented a full multimodal regression pipeline combining:

- Satellite image embeddings extracted using a pre trained CNN
- Structured property attributes (size, rooms, location, age, etc.)
- A tree-based regressor (LightGBM) for final price prediction
- Used EfficientNet-B0 (pre trained on ImageNet) via timm.
- Removed the classification head and used global average pooled embeddings.
- Because CNN embeddings are high-dimensional, we applied:
  - Standard scaling
  - Incremental PCA (128 components) to manage memory and reduce noise

However for above model we obtained an average R square of 0.69 so we didn’t move forward with this.