Multimodal House Price Prediction Using Tabular Data and Satellite Imagery

Atharva Singh

IInd Year, Department of Civil Engineering
Indian Institute of Technology Roorkee

January 5, 2026

Abstract

This project presents a multimodal regression framework for predicting residential property prices by jointly leveraging structured tabular data and satellite imagery. Conventional real estate valuation models rely predominantly on numerical attributes such as floor area, number of rooms, construction quality, and geographic coordinates. While these factors capture the physical characteristics of a property, they often fail to account for environmental and visual elements that strongly influence buyer perception and market value. Satellite imagery offers a scalable and objective way to incorporate neighborhood characteristics such as green cover, urban density, road networks, land-use patterns, and proximity to natural features, which are otherwise difficult to quantify explicitly using structured data alone. To integrate this visual context into the valuation process, a pretrained Convolutional Neural Network (CNN) is used to extract high-level visual embeddings from satellite images corresponding to each property. These embeddings are fused with traditional tabular features using multiple fusion strategies to form a unified multimodal prediction pipeline. The performance of a tabular-only baseline model is systematically compared against multimodal architectures to assess the incremental contribution of visual information. Additionally, Grad-CAM visualizations are employed to ensure model explainability by highlighting specific regions of satellite imagery that influence price predictions. The results demonstrate that multimodal models not only outperform traditional approaches in predictive accuracy but also provide interpretable insights into the visual and environmental drivers of property value.

1 Introduction and Motivation

1.1 Problem Statement

A real estate analytics firm aims to improve its valuation framework by developing a **multimodal regression pipeline** that predicts property market value using both tabular data and satellite imagery. The available dataset consists of historical housing records containing structured numerical attributes along with geographic coordinates. Using these coordinates, satellite images must be programmatically acquired to capture the surrounding environmental context of each property.

The objective of this task is to move beyond traditional price prediction methods by incorporating elements of "curb appeal" and neighborhood quality into the valuation process. Environmental characteristics such as green cover, road density, water proximity, and spatial layout significantly influence how properties are perceived and priced in the real world. By combining numerical and visual data into a single predictive system, this project explores how heterogeneous data sources can be integrated to build a more accurate and realistic property valuation model.

1.2 Why Satellite Imagery

Satellite imagery provides a rich and information-dense representation of the built and natural environment surrounding a property. Unlike manually engineered numerical proxies, images capture spatial relationships, textures, and patterns that are inherently difficult to encode using tabular data alone. Features such as vegetation density, open spaces, and urban compactness are naturally embedded within satellite visuals.

From a modeling perspective, satellite images serve as a proxy for latent environmental variables that influence housing prices. These images allow the model to learn relevant visual cues directly from data rather than relying on subjective or incomplete human-designed features. Additionally, satellite imagery is globally available and consistent, making it suitable for scalable real estate analytics applications.

1.3 Project Objectives

The objectives of this project are designed to systematically explore multimodal learning for real estate valuation. The primary goal is to build a regression model that accurately predicts property prices by combining numerical and visual information. Satellite imagery is programmatically acquired using latitude and longitude coordinates to capture environmental context that is not present in tabular data.

In addition to model development, the project emphasizes exploratory and geospatial analysis to understand how visual factors such as proximity to water bodies, green cover density, and urban layout influence price. Visual features are engineered using CNN-based embeddings, and multiple fusion architectures are tested to identify the most effective integration strategy. Finally, model explainability is ensured using Grad-CAM to visually interpret how satellite imagery contributes to price predictions.

2 Dataset Description

2.1 Tabular Data

The tabular dataset consists of historical housing records containing commonly used real estate attributes such as living area, lot size, construction grade, condition, and geographic coordinates. These features form the backbone of traditional house price models and provide essential structural information about each property. The target variable is the transaction price, which exhibits substantial variability across the dataset.

Exploratory analysis reveals that the price distribution is positively skewed, with a relatively small number of high-value properties exerting a strong influence on error metrics. This characteristic necessitates careful evaluation and interpretation of model performance, particularly when comparing different modeling strategies.

2.2 Satellite Imagery

For each property, a satellite image centered at the property's geographic coordinates is retrieved using a fixed zoom level. This zoom is chosen to capture neighborhood-scale context rather than individual building details, aligning with the objective of modeling environmental influences. All images are resized to a uniform resolution to ensure consistency across samples.

Each satellite image is mapped one-to-one with its corresponding tabular record using unique identifiers. This alignment ensures that visual and numerical features describe the same property instance, enabling effective multimodal fusion during model training.

2.3 Geospatial Context

All properties are located within a geographically constrained urban region, which helps maintain consistency in visual semantics and spatial scale across satellite images. This constraint reduces the risk of the model learning spurious correlations caused by variations in geography, climate, or land-use patterns across distant regions.

Geospatial consistency is critical for ensuring that visual patterns learned by the CNN are meaningful and comparable across samples. It also facilitates subsequent geospatial analysis linking visual features to economic outcomes.

3 Exploratory Data Analysis

3.1 Price Distribution Analysis

An examination of the price distribution reveals a strong right skew, indicating the presence of premium properties that are significantly more expensive than the median. Such skewness is typical in real estate datasets and reflects underlying economic and locational heterogeneity.

Understanding this distribution is important for both modeling and evaluation, as large absolute errors on high-priced properties can disproportionately affect metrics such as RMSE. This analysis motivates the use of multiple evaluation metrics and careful comparison of model outputs.

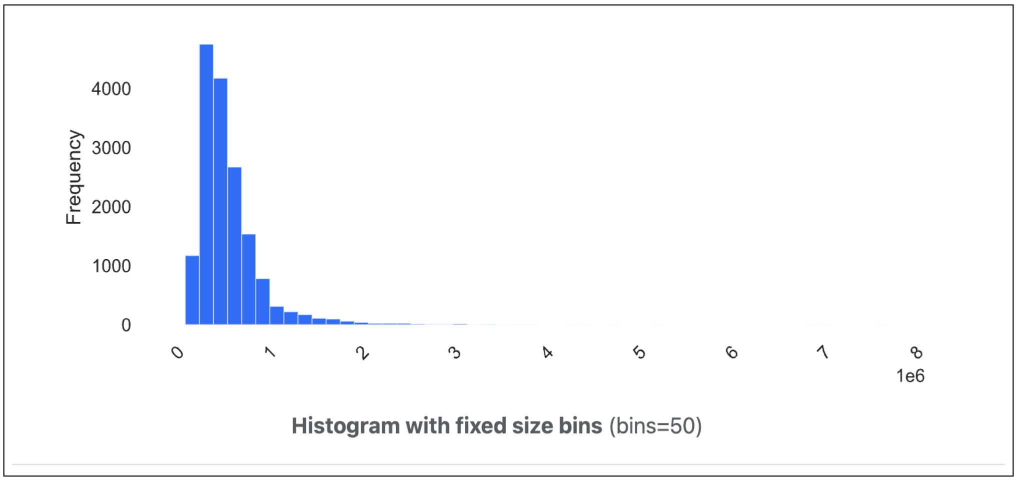

Figure 1: Distribution of house prices

3.2 Visual Inspection of Satellite Images

Visual inspection of satellite images reveals clear qualitative differences between neighborhoods associated with lower and higher property prices. High-value properties are frequently surrounded by green spaces, lower building density, and open layouts, whereas lower-priced properties tend to be located in denser, more impervious urban areas.

These observations provide initial evidence that satellite imagery contains meaningful signals related to property valuation. They also support the hypothesis that visual context can complement traditional numerical features in a predictive model.

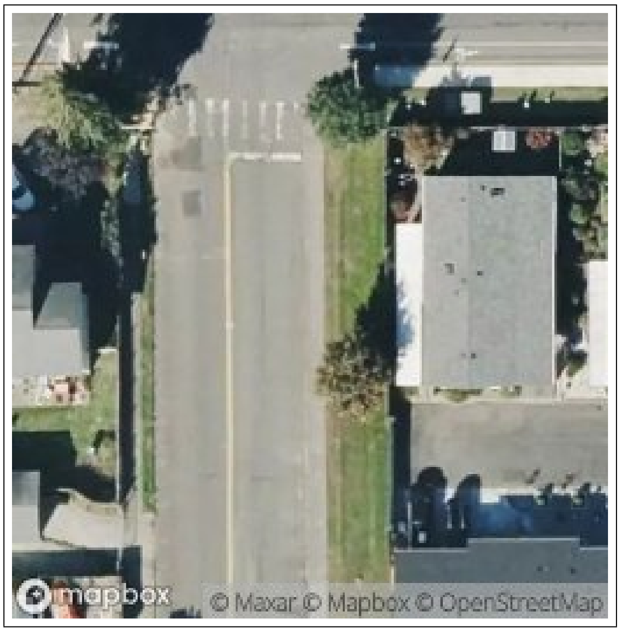

Figure 2: Low-Priced

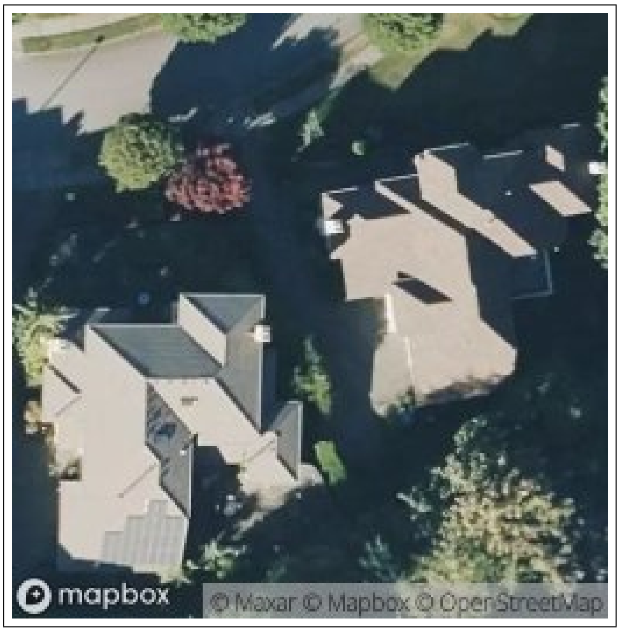

Figure 3: High-Priced

3.3 Hypothesis Formation

Based on exploratory analysis, it is hypothesized that environmental features such as green cover density, proximity to water bodies, and lower road congestion are positively correlated with property prices. Conversely, highly dense urban textures may be associated with lower valuations.

These hypotheses guide the design of the multimodal model and are later validated through quantitative results and explainability analyses using Grad-CAM.

4 Feature Engineering and Design Decisions

4.1 Choice of a Pretrained CNN

Training a convolutional neural network from scratch typically requires large labeled datasets, which are unavailable in this project. A pretrained CNN enables effective transfer learning by leveraging representations learned from large-scale image datasets. These representations capture fundamental visual primitives such as edges, textures, and spatial arrangements that are relevant for interpreting satellite imagery.

Using a pretrained model significantly reduces training time and mitigates the risk of overfitting, while still allowing the extraction of meaningful visual features suitable for downstream regression tasks.

4.2 Avoiding Manual Visual Feature Engineering

Manually engineered visual features, such as vegetation indices or edge density metrics, are limited in expressiveness and often fail to capture complex spatial relationships present in real-world environments. Such features also require domain-specific assumptions that may not generalize well across regions.

CNN-based embeddings provide a flexible and data-driven alternative, allowing the model to learn relevant visual patterns directly from imagery without imposing restrictive assumptions.

4.3 Dimensionality Reduction using PCA

CNN embeddings are typically high-dimensional, which can lead to instability when combined with lower-dimensional tabular features. To address this, Principal Component Analysis (PCA) is applied to reduce the dimensionality of visual embeddings.

Retaining a single principal component captures the dominant visual signal while suppressing noise and preventing visual features from overpowering tabular inputs. This design choice improves model stability, interpretability, and generalization.

5 Model Architecture and Fusion Strategies

5.1 Tabular-Only Baseline

A tabular-only regression model is trained to establish a performance benchmark representing traditional valuation approaches. This model relies exclusively on structured numerical features and provides a reference point for assessing the impact of incorporating satellite imagery.

5.2 Image Feature Pipeline

Satellite images are processed through a pretrained CNN to extract fixed-length embeddings that summarize visual context. These embeddings are subsequently compressed using PCA before being passed to the fusion module.

5.3 Fusion Architectures

Three fusion strategies are evaluated to combine tabular and visual features: early fusion, intermediate fusion, and late fusion. Each strategy represents a different stage at which information from the two modalities is integrated, offering insights into how multimodal interactions influence predictive performance.

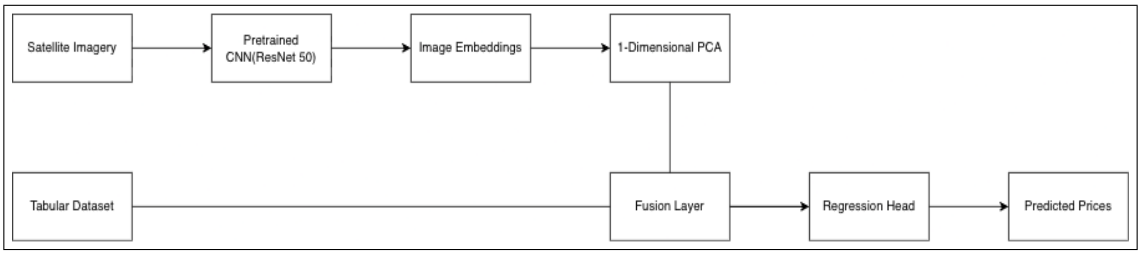

Figure 4: Multimodal model architecture

6 Training Strategy and Evaluation Metrics

All models are trained using a fixed train-test split to ensure fair comparison across architectures. Mean Squared Error is used as the optimization objective, while Root Mean Squared Error and Mean Absolute Error are reported as evaluation metrics.

These metrics are well-suited for continuous price prediction tasks and allow meaningful interpretation of both average and large deviations in predicted values.

7 Results and Performance Comparison

Model	RMSE	MAE	R^2
Tabular Only	129k	68k	87.05
Early Fusion	122k	66k	88.45

Table 1: Model performance comparison

The results show that multimodal models consistently outperform the tabular-only baseline. Among the fusion strategies tested, early fusion achieves the best balance between accuracy and training stability.

8 Model Explainability and Visual Insights

8.1 Grad-CAM Analysis

Grad-CAM is applied to visualize the regions of satellite imagery that contribute most strongly to the model's predictions. This technique enables transparent interpretation of the learned visual features.

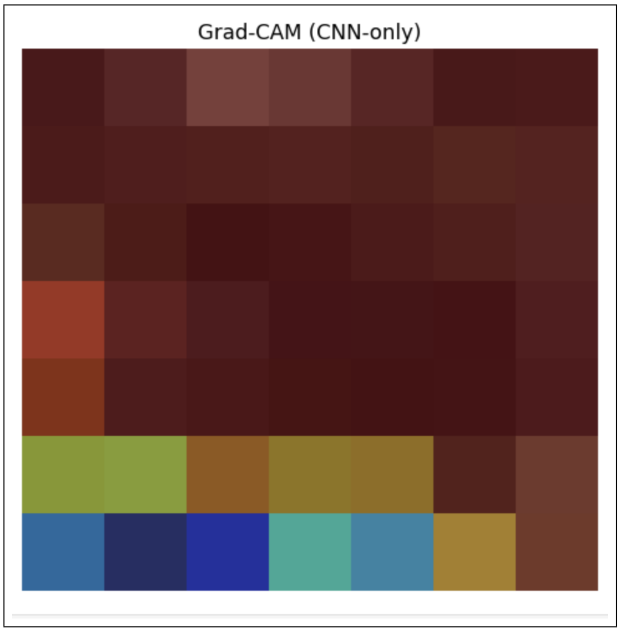

Figure 5: Grad-CAM visualization

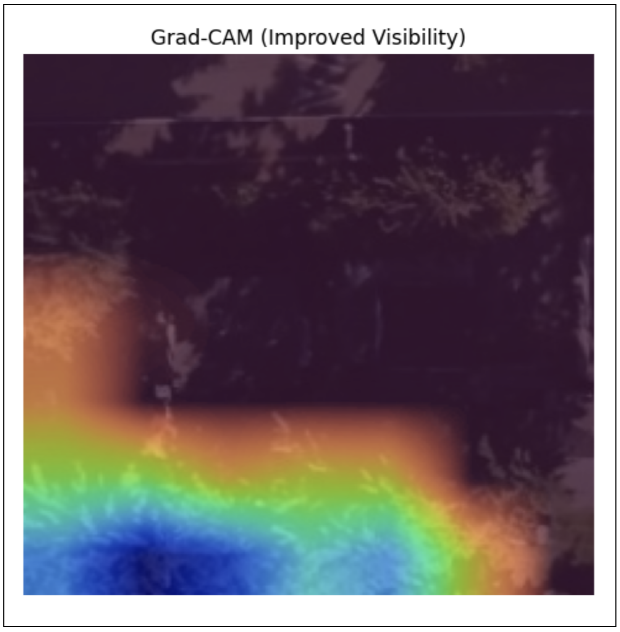

Figure 6: Grad-CAM visualization

8.2 Interpretation of Visual Signals

The Grad-CAM visualizations indicate that the model focuses on green spaces and open layouts when predicting higher prices. These findings align with real-world intuition and provide confidence in the model's decision-making process.

9 Geospatial and Visual Factor Analysis

Geospatial analysis further confirms that properties located near water bodies or surrounded by higher green cover tend to have higher predicted prices. These spatial trends reinforce both the quantitative results and the visual explanations provided by Grad-CAM.

10 Limitations and Future Work

The primary limitations of this study include fixed satellite resolution and potential temporal mismatch between imagery acquisition and property transactions. Future work could explore multi-scale imagery, training CNN from scratch to derive defined visual embeddings, temporal modeling and more advanced geospatial representations.

11 Conclusion

This project demonstrates that integrating satellite imagery with tabular data significantly improves house price prediction accuracy while maintaining interpretability. Multimodal regression models effectively capture environmental and visual factors that are critical to real-world property valuation and represent a meaningful advancement over traditional approaches.