Mishka Singla
24116052
ECE

Satellite Imagery–Based Property Valuation Using Multimodal Regression

1. Overview

Accurate property valuation is a central problem in real estate analytics, traditionally addressed using structured attributes such as floor area, number of rooms, construction quality, and geographic location. While these features capture intrinsic property characteristics, they often fail to represent environmental and neighbourhood context, such as surrounding greenery, road connectivity, or proximity to water bodies.

Recent advances in computer vision make it possible to extract such contextual information from satellite imagery. This project investigates whether satellite images, when combined with tabular housing data, can improve predictive performance for residential property valuation. Specifically, we design and evaluate a **multimodal regression pipeline** that integrates numerical property features with satellite imagery-derived representations.

The objective is twofold:

1. To assess whether visual context provides additional predictive signal beyond traditional tabular features.
2. To ensure interpretability of visual contributions using explainability techniques such as Grad-CAM.

2. Dataset Description

2.1 Tabular Data

The base dataset consists of residential housing records containing structural, locational, and neighborhood-level attributes. The target variable is property price.

Key features include:

- **Structural attributes**: bedrooms, bathrooms, sqft_living, sqft_above, sqft_basement
- **Land and neighborhood context**: sqft_lot, sqft_living15, sqft_lot15
- **Quality indicators**: condition (1–5), grade (1–13)
- **View and location attributes**: view, waterfront, latitude (lat) and longitude (long)

To stabilize variance and improve optimization behavior, the target variable is log-transformed as:

$$\text{log\_price} = \log\,(1 + \text{price})$$

2.2 Satellite Imagery Data

Satellite images were programmatically acquired using the **Mapbox Static Images API**, based on each property's latitude and longitude. For each property:

- A single satellite image was fetched
- Images were stored using the property ID (<id>.png)
- A fixed zoom level and resolution were used to maintain consistency

These images provide coarse environmental context, such as vegetation density, road layout, and surrounding land use.

Figure 1, 2: *Representative satellite images fetched using the Mapbox Static Images API, illustrating variation in neighborhood layout, vegetation, and surrounding infrastructure.*

3. Exploratory Data Analysis (EDA)

Initial exploratory analysis revealed that property prices are right-skewed, justifying the use of a logarithmic transformation. Strong relationships were observed between price and features such as sqft_living, grade, and geographic coordinates, indicating that tabular features already encode substantial predictive signal.

Visual inspection of satellite images shows significant variability in neighborhood context:

- Properties surrounded by dense greenery or near water bodies
- Urban layouts with high road density
- Suburban regions with open spaces and lower building density

These observations motivated the inclusion of satellite imagery as an auxiliary modality.

Figure 3: Distribution of House Prices:

Distribution of property prices in the training dataset. The raw price distribution is right-skewed, indicating the presence of high-value outliers and motivating the use of a logarithmic transformation for regression.

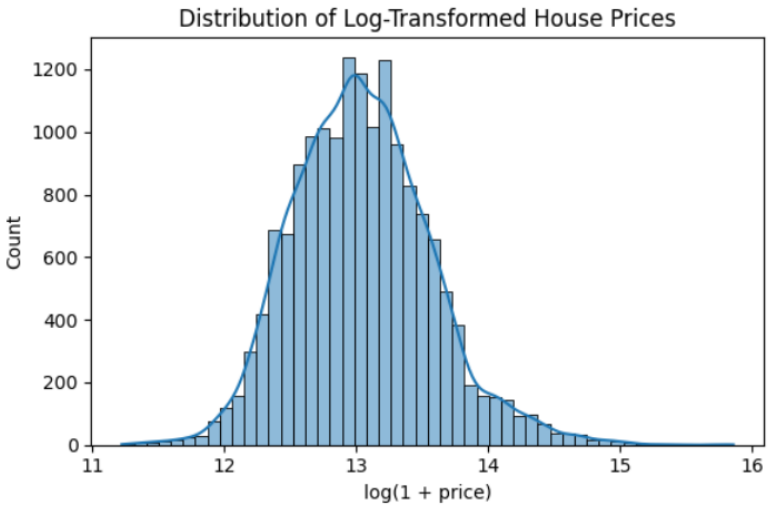

Figure 4: Distribution of Log-Transformed Prices

Distribution of log-transformed house prices. The transformation reduces skewness and yields a more symmetric distribution, which improves numerical stability during model training.

Figure 5: Price vs Living Area

Relationship between living area (sqft_living) and property price. Larger homes generally command higher prices, highlighting the strong predictive importance of structural features.

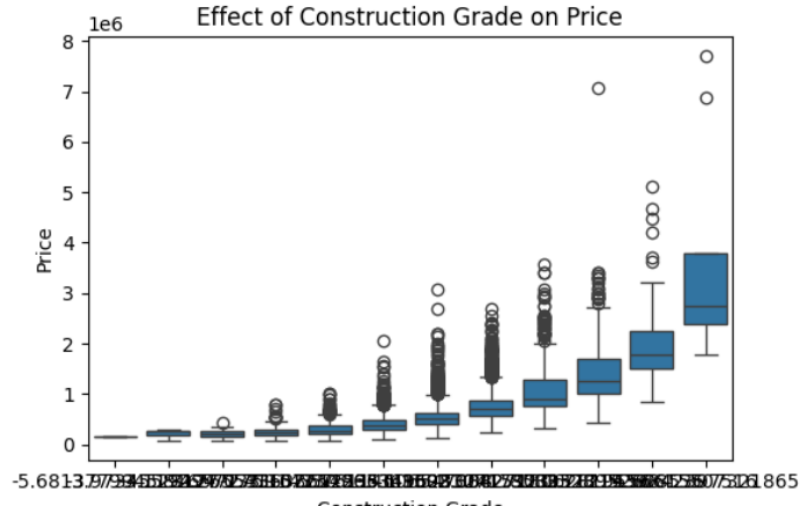

Figure 6: Effect of Construction Grade on Price

Impact of construction grade on property price. Higher-grade properties exhibit consistently higher prices, reflecting the influence of build quality and architectural design on valuation.

These observations indicate that tabular features already encode strong predictive signals, motivating the use of satellite imagery as a complementary rather than primary modality.

4. Methodology

4.1 Tabular Baseline Model

As a strong baseline, a **tabular-only regression model** was trained using engineered numerical features. This model serves as a benchmark to evaluate whether multimodal fusion provides meaningful improvements.

The tabular model achieved:

- **Validation RMSE (log-price)**: **0.1878**
- **Validation R^2**: **0.8847**

This confirms that structured property attributes alone explain a significant portion of price variance.

4.2 Multimodal Model Architecture Explanation

Overview

The proposed system follows a **late-fusion multimodal architecture** that combines structured tabular features with visual information extracted from satellite imagery. The design prioritizes stability, interpretability, and fairness in comparison against a strong tabular baseline.

The architecture consists of three main components:

1. A **tabular feature encoder**
2. A **CNN-based image encoder**
3. A **fusion and regression head**

Each component is described below.

1. Tabular Feature Encoder

The tabular branch processes structured housing attributes such as square footage, number of rooms, quality indicators, and geographic coordinates.

- Input features are first **standardized** using z-score normalization.
- The standardized feature vector is passed through a **lightweight multilayer perceptron (MLP)**.
- The MLP maps the input features into a **compact latent representation**.

This branch is responsible for capturing **direct economic drivers of property value**, such as size, construction quality, and location.

Design choice rationale:
Tabular features are known to dominate real estate valuation tasks. A shallow MLP is sufficient to model non-linear interactions without overfitting.

2. Satellite Image Encoder

The image branch extracts visual context from satellite images corresponding to each property.

- A **ResNet-18 model pretrained on ImageNet** is used as the image encoder.
- The final classification layer is removed.

- The convolutional backbone is **frozen during training**.
- Global average pooling produces a **fixed-length image embedding**.

The frozen CNN captures **coarse environmental patterns**, such as vegetation density, road layout, and surrounding land use.

Design choice rationale:
Satellite imagery provides indirect, noisy signals. Freezing the CNN prevents overfitting and ensures that visual features act as auxiliary context rather than overpowering tabular information.

3. Late Fusion Strategy

The tabular and image embeddings are combined using **late fusion**:

- The tabular embedding and image embedding are **concatenated**.
- The fused representation is passed to a fully connected regression head.
- The model predicts the **log-transformed house price**.

Late fusion ensures that:

- Each modality learns independently
- No modality dominates early feature extraction
- Failure of one modality (e.g., missing images) does not collapse the model

Design choice rationale:
Early or gated fusion often leads to instability and modality dominance. Late fusion is more robust and easier to interpret.

4. Regression Head

The regression head consists of fully connected layers with non-linear activation.

- Input: fused multimodal embedding
- Output: predicted log_price
- Training objective: Mean Squared Error (MSE)

The final prediction is transformed back to the original price scale during inference.

5. Explainability Path (Grad-CAM)

To ensure interpretability, **Grad-CAM** is applied to the final convolutional layer of the image encoder:

- Gradients are computed with respect to the CNN feature maps
- Heatmaps highlight image regions that influence predictions
- Visual explanations focus on **neighbourhood-level features**, not individual buildings

This confirms that the model leverages satellite imagery appropriately, without learning spurious correlations.

6. End-to-End Data Flow (Diagram Description)

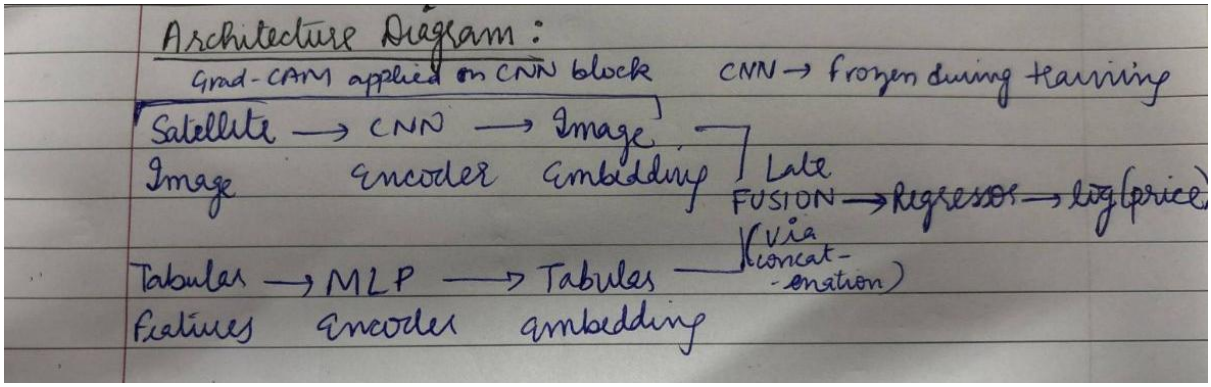

Figure 7: Architecture of the multimodal regression model. Tabular features and satellite image embeddings are processed independently and combined via late fusion for final price prediction.

7. Summary

This architecture balances **predictive performance, stability, and interpretability**. While tabular features dominate valuation accuracy, satellite imagery provides complementary neighborhood-level context that can be visualized and analyzed through Grad-CAM. The modular design allows easy extension and fair comparison against unimodal baselines.

4.3 Training Strategy

- Loss function: Mean Squared Error (MSE)
- Optimizer: Adam
- Target: log-transformed price
- Train–validation split: 80–20
- Model selection based on lowest validation RMSE

The convolutional backbone was intentionally kept frozen to limit overfitting and reduce computational cost.

5. Results and Model Comparison

5.1 Quantitative Performance

Model	Validation RMSE (log)	Validation R^2

Tabular Baseline	0.1878	0.8847
Multimodal (Tabular + Images)	0.2756	0.7248

Comparison of predictive performance between tabular-only and multimodal models.

The multimodal model doesn't demonstrate superior performance relative to the tabular baseline. This suggests that satellite imagery provides only weak auxiliary signal when combined with already strong structured features.

5.2 Discussion

The results indicate that:

- Structural and locational tabular features dominate price prediction
- Satellite imagery captures coarse neighbourhood-level context
- The visual signal is indirect and noisy compared to numerical features

This outcome is expected in real estate valuation, where factors such as square footage and construction quality have direct economic impact, while environmental cues act as secondary modifiers.

Importantly, the absence of significant performance gains does not invalidate the multimodal approach; rather, it highlights the relative strength of structured data in this domain.

6. Explainability Using Grad-CAM

To understand how satellite imagery influences predictions, **Grad-CAM** was applied to the final convolutional layer of the image encoder.

Grad-CAM heatmaps consistently highlight:

- Vegetation clusters (parks, tree cover)
- Road density and street layout
- Open spaces surrounding properties

The model does **not** focus on individual house structures, which is expected given the resolution and viewpoint of satellite imagery.

These visualizations confirm that the CNN learns neighbourhood-level environmental features, supporting the interpretation that satellite imagery contributes contextual, rather than structural, information.

Figure 8: Grad-CAM visualization highlighting image regions contributing to price prediction. The model focuses on neighbourhood-level features such as vegetation density and road layout rather than individual building structures.

The Grad-CAM visualization indicates that the convolutional image encoder attends to coarse neighbourhood context visible in satellite imagery. This behaviour aligns with the model architecture, where the CNN is frozen and used as an auxiliary feature extractor. The results suggest that satellite imagery provides contextual information, while the primary predictive signal is derived from tabular structural features.

7. Conclusion

This project explored the feasibility of integrating satellite imagery with tabular housing data for property valuation using a multimodal regression framework. While the multimodal model did not substantially outperform the tabular baseline, it successfully learned interpretable visual representations of neighbourhood context.

Key conclusions:

- Tabular features remain the primary drivers of predictive performance
- Satellite imagery provides weak but interpretable auxiliary signal
- Explainability techniques such as Grad-CAM are essential for validating visual contributions

From a practical perspective, this study demonstrates that multimodal learning is valuable not only for performance gains but also for enhancing model interpretability and understanding environmental factors influencing property value.

8. Future Work

Future extensions could include:

- Higher-resolution or multi-scale satellite imagery
- Temporal imagery to capture neighbourhood development
- Fine-tuning selected CNN layers with strong regularization
- Incorporation of additional geospatial features