Satellite Imagery–Based Property Valuation Using Multimodal Machine Learning

CDC X Yhills Open Projects 2025–2026

Data Science Problem Statement

Submitted by:
Kumar Manas
23119016
Indian Institute of Technology, Roorkee

Under the initiative of
Career Development Cell (CDC), IIT Roorkee

GitHub Repo: https://github.com/HighOnKeys/satellite-property-valuation

Abstract

Traditional property valuation models primarily rely on structured housing attributes such as size, number of rooms, and location coordinates. While effective, these models often fail to capture neighbourhood level visual context, including green cover, road connectivity, and surrounding land use, which significantly influence real estate prices.

This project proposes a **multimodal machine learning framework** that integrates structured tabular data with **satellite imagery** to improve residential property price prediction. Satellite images corresponding to each property's geographic coordinates are programmatically acquired using the **Mapbox Static Images API**. Visual features are extracted from these images using a pretrained Convolutional Neural Network (ResNet-18), producing compact image embeddings that encode neighbourhood characteristics.

The extracted image embeddings are fused with tabular housing features and passed to an **XGBoost regression model** for final price prediction. A tabular only baseline is first established for comparison. Model interpretability is addressed using **Grad-CAM**, which highlights spatial regions in satellite imagery that influence the learned visual representations.

Experimental results demonstrate that incorporating satellite imagery improves predictive performance over tabular only models, confirming the value of visual neighbourhood context in property valuation. This project showcases an end to end, explainable, and reproducible multimodal pipeline suitable for real-world real estate analytics.

1. Problem Statement

Accurate property valuation is a critical task in real estate analytics, influencing decision making for buyers, sellers, financial institutions, and urban planners. Traditional machine learning models for property price prediction predominantly rely on structured tabular attributes such as living area, number of bedrooms, construction quality, and geographic coordinates. While these features capture intrinsic property characteristics, they often fail to represent **neighbourhood level context**, which plays a significant role in determining market value.

Factors such as surrounding green cover, road density, proximity to water bodies, and urban layout are difficult to quantify using tabular data alone. As a result, tabular only models may produce biased or incomplete valuations, particularly when properties with similar structural attributes are located in visually distinct neighbourhoods.

From a data science perspective, this presents a **feature representation challenge**: how to effectively incorporate unstructured visual information into predictive models alongside structured data. Satellite imagery offers a rich source of spatial and environmental context, but integrating image data with traditional tabular features requires careful handling of heterogeneous data modalities.

The objective of this project is to design a **multimodal machine learning pipeline** that combines structured housing attributes with satellite imagery to improve property price prediction. The problem involves programmatically acquiring satellite images using geographic coordinates, extracting meaningful visual features, and fusing them with tabular data in a robust and interpretable regression framework.

2. Dataset Description

The primary dataset used in this project is derived from the **King County House Sales dataset**, a publicly available real estate dataset originally sourced from Kaggle. The dataset contains historical residential property sales records from King County, Washington, and is widely used for benchmarking property valuation models.

2.1 Tabular Data

The original dataset consists of **16,209 residential properties**, each described using a combination of structural, locational, and neighbourhood level attributes. Key features include:

- **Price:** Sale price of the property (target variable)
- **Structural attributes:** Living area (`sqft_living`), lot size (`sqft_lot`), number of bedrooms and bathrooms
- **Quality indicators:** Construction grade, overall condition, and view rating
- **Location data:** Latitude and longitude coordinates

- **Neighbourhood context:** Average living area and lot size of nearby properties (`sqft_living15`, `sqft_lot15`)
- **Special attributes:** Waterfront indicator and proximity-based features

These features capture intrinsic property characteristics but do not explicitly represent the surrounding visual environment.

2.2 Satellite Imagery Data

To incorporate neighbourhood level visual context, satellite images were programmatically acquired for each property using geographic coordinates. The **Mapbox Static Images API** was used to download satellite imagery cantered at each property's latitude and longitude. These images provide information about surrounding land use, road networks, green cover, and urban layout.

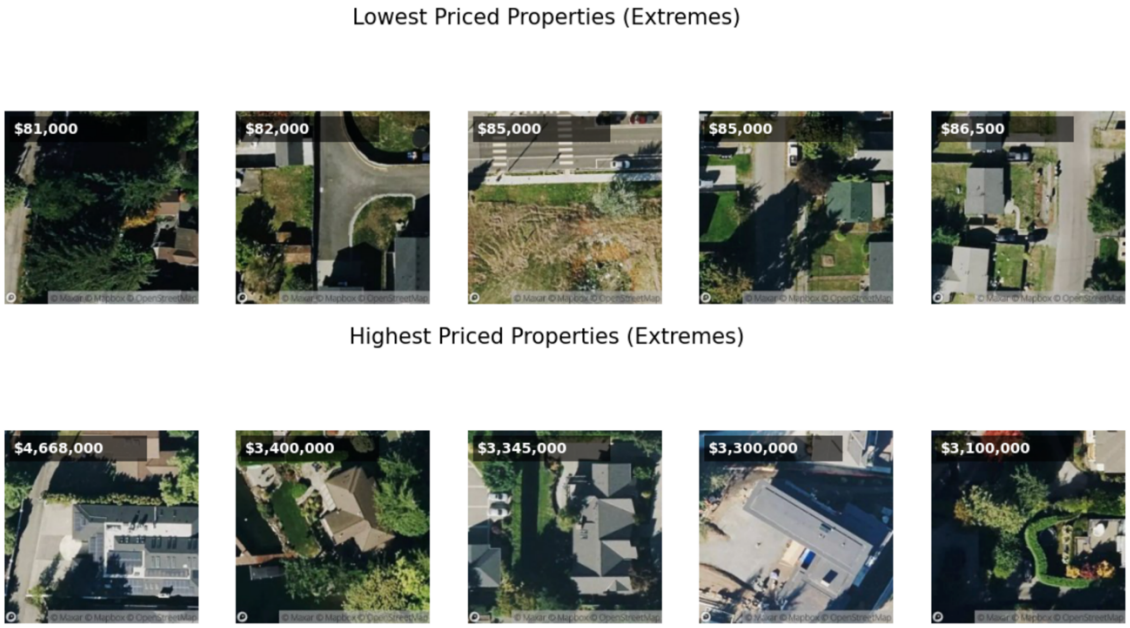

Figure 1: *Satellite images sampled across low and high price ranges*

2.3 Sampling Strategy

While the full tabular dataset contains approximately 16,000 properties, downloading and processing satellite imagery for all properties is computationally expensive. To balance computational efficiency with representational diversity, a **stratified sampling strategy** was employed.

The training data was divided into quantile based price bins, and a fixed number of samples was selected from each bin. This resulted in a **stratified subset of approximately 5,000 properties**, ensuring that low , mid , and high priced properties were proportionally represented in the multimodal training set.

The **full test dataset** was retained without sampling to ensure unbiased final inference.

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand the distribution of property prices, identify key drivers of valuation, and examine spatial and neighbourhood level patterns. These insights guided feature selection and motivated the use of satellite imagery for capturing visual context not present in tabular data.

3.1 Distribution of Property Prices

The raw distribution of property prices is shown in **Figure 2**. The distribution is highly **right skewed**, with a long tail corresponding to luxury properties. Most properties are concentrated in the lower-to-mid price range, while a small number of properties are priced several times higher than the median.

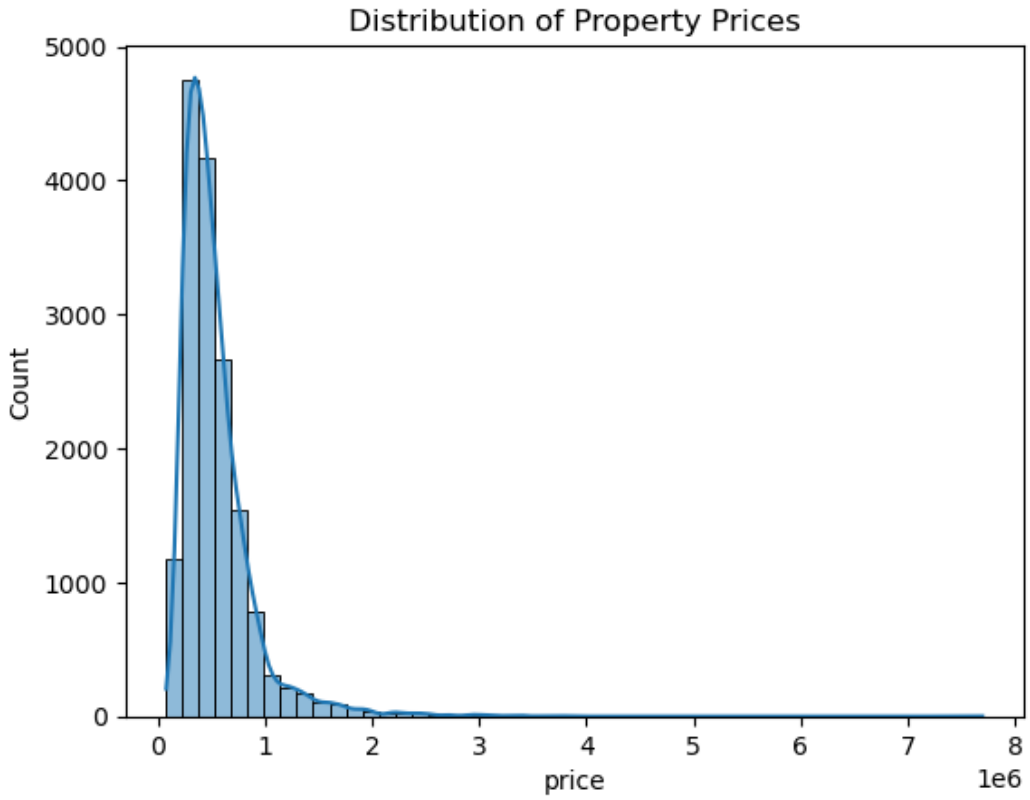

Figure 2: *Distribution of Property Prices*

3.2 Living Area and Price Relationship

Figure 3 illustrates the relationship between **living area (`sqft_living`)** and property price. A clear positive correlation is observed: properties with larger living areas tend to command higher prices. However, the relationship is **non-linear**, with increasing variance for larger homes.

Figure 3: *Living Area vs Property Price*

Figure 4 extends this analysis by examining **average living area of nearby properties (`sqft_living15`)**. This feature captures neighbourhood density and socio economic context. Properties surrounded by larger neighbouring homes tend to be priced higher, even when their own size is moderate.

Figure 4: *Neighbour Living Area vs Property Price*

Insight:
neighbourhood level features contribute meaningful information beyond individual property attributes, supporting the inclusion of contextual features in the model.

3.3 Correlation Analysis

A correlation heatmap of selected numerical features is shown in **Figure 5**. Key observations include:

- Strong correlation between **price and living area**
- Significant association between **construction grade and price**
- Moderate correlation between neighbourhood features and price
- Relatively weak correlation between lot size and price

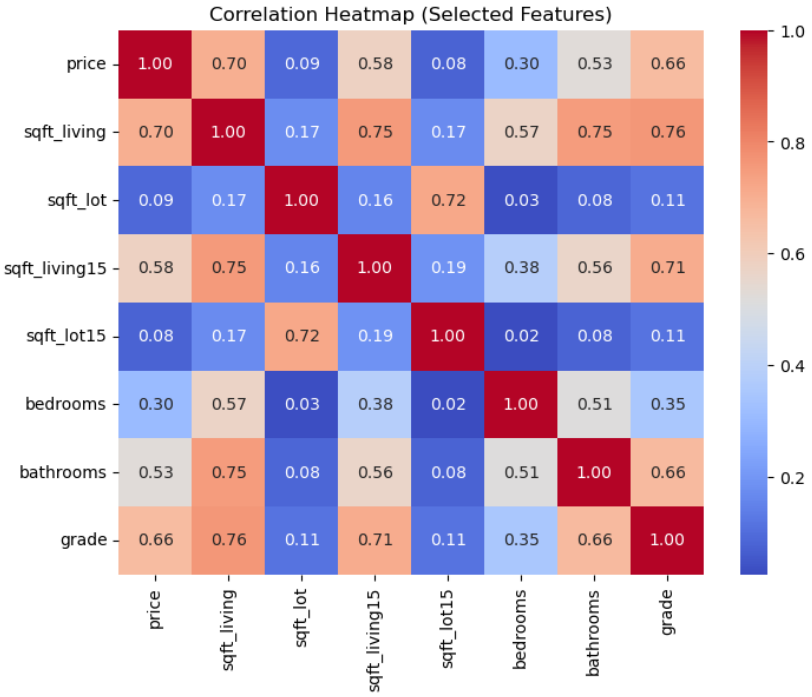

Figure 5: *Correlation heatmap vs Selected features*

Insight:

While several features show strong linear correlations, others exhibit weaker or indirect relationships, suggesting that non-linear models may better capture interactions among features.

3.4 Impact of Construction Quality and Waterfront

Figure 6 shows property prices across different **construction grades**. Higher grade constructions consistently exhibit higher median prices, with increased variability at upper grades due to luxury properties.

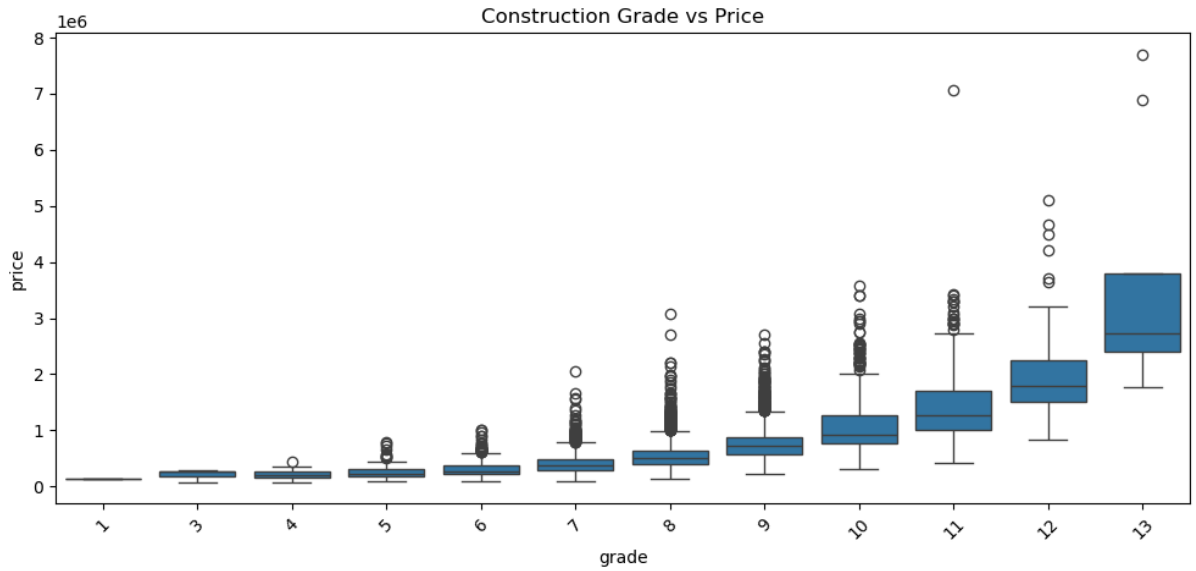

Figure 6: *Construction Grade vs Property Price*

Figure 7 compares prices for properties with and without **waterfront access**. Waterfront properties are significantly more expensive on average, highlighting the strong premium associated with scenic and locational advantages.

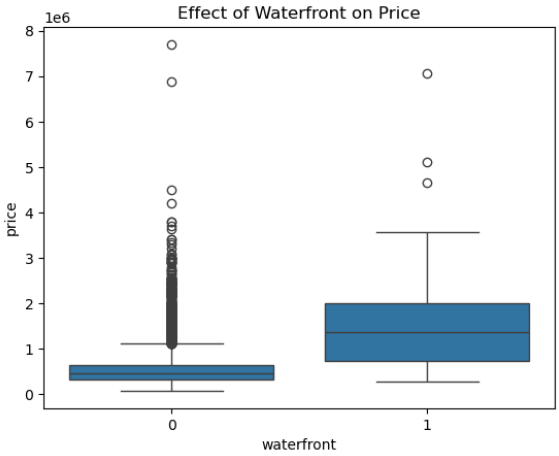

Figure 7: *Effect of Waterfront on Property Price*

Insight:
These categorical attributes represent high impact pricing factors that are not fully captured by numeric size based features alone.

3.5 Geographic Distribution of Prices

The spatial distribution of properties coloured by log price is shown in **Figure 8**. High priced properties are spatially clustered, particularly near waterfronts and well developed urban regions. Lower priced properties are more dispersed across the region.

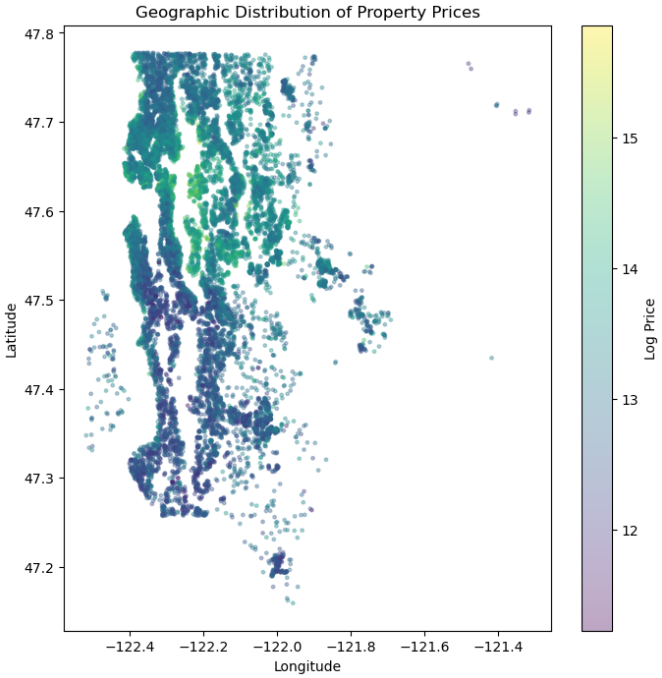

Figure 8: *Geographic Distribution of Property Price*

Insight:
Price exhibits strong spatial dependency, reinforcing the importance of geographic and environmental context. Satellite imagery provides a natural way to capture these spatial patterns at scale.

3.6 Summary of EDA Findings

- Property prices are heavily right skewed with extreme outliers.
- Structural features such as living area and construction grade strongly influence price.
- Neighbourhood level and geographic factors play a critical role in valuation.
- Visual and spatial context cannot be fully represented by tabular features alone.

These observations motivate the use of a **multimodal learning approach**, integrating satellite imagery with structured data for improved price prediction.

4. Methodology

This project follows a structured, end to end machine learning pipeline to integrate heterogeneous data modalities structured tabular data and unstructured satellite imagery for property price prediction. The methodology consists of three main stages: a tabular baseline model, multimodal feature fusion, and final regression.

4.1 Tabular Baseline Model

As a reference point, a tabular only regression model was first developed using traditional housing attributes such as living area, number of bedrooms and bathrooms, construction grade, neighbourhood statistics, and geographic coordinates.

Two models were evaluated:

- **Linear Regression**, to establish a simple baseline
- **XGBoost Regression**, to capture non-linear relationships

Model	RMSE	R^2
Linear Regression	191661.41	0.7073
XGBoost Regression	117557.95	0.8898

Table 1: *Tabular Baseline Models Comparission*

The tabular XGBoost model significantly outperformed linear regression, demonstrating the importance of non-linear modelling for real estate valuation. This model serves as the benchmark against which the multimodal approach is evaluated.

4.2 Satellite Image Feature Extraction

To incorporate visual neighbourhood context, satellite images corresponding to each property were downloaded using the **Mapbox Static Images API**, cantered at the property's latitude and longitude.

A **pretrained ResNet-18 Convolutional Neural Network** was used as a fixed feature extractor. The final classification layer was removed, and the network outputs a **512-dimensional image embedding** for each property. These embeddings capture spatial patterns such as building density, road layout, vegetation, and surrounding land use.

The CNN weights were frozen during feature extraction to leverage pretrained visual representations and avoid overfitting on the limited satellite image dataset.

4.3 Multimodal Feature Fusion and Regression

The extracted image embeddings were concatenated with the tabular features to form a unified multimodal feature vector for each property. This fused representation combines intrinsic property attributes with environmental and neighbourhood level visual context.

An **XGBoost regression model** was trained on the combined feature set to predict property prices. XGBoost was chosen due to its ability to:

- Handle high dimensional feature spaces
- Model complex non-linear interactions
- Remain robust to feature scale differences and outliers

Model performance was evaluated using Root Mean Squared Error (RMSE) and R^2 score on a held-out validation set.

4.4 End-to-End Pipeline Overview

Figure 9 illustrates the complete multimodal pipeline, from raw data ingestion to final price prediction. The modular design ensures reproducibility and allows individual components (image encoder or regressor) to be independently improved in future work.

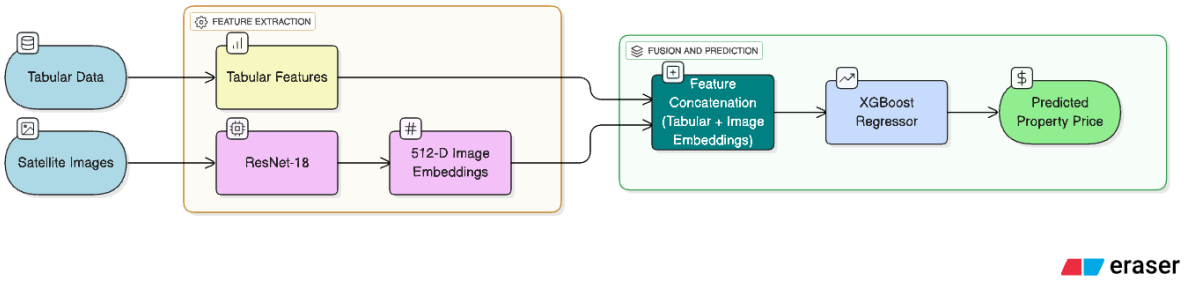

Figure 9: *Multimodal Property Valuation Pipeline*

5. Model Explainability Using Grad-CAM

While multimodal models often achieve strong predictive performance, their use of unstructured visual data can raise concerns regarding interpretability. In real world valuation systems, especially in high stakes domains such as real estate pricing, it is essential to understand *what visual cues the model relies on* rather than treating predictions as black box outputs.

To address this, **Gradient-weighted Class Activation Mapping (Grad-CAM)** was applied to the convolutional neural network component to visually interpret the spatial regions of satellite images that influence model predictions.

5.1 Motivation for Visual Explainability

Satellite images encode rich contextual information such as:

- Road connectivity and intersections
- Density and spacing of buildings
- Presence of green spaces and open land
- neighbourhood layout and surrounding infrastructure

However, without explicit interpretability, it is unclear whether the CNN focuses on meaningful real-estate signals or spurious patterns. Grad-CAM provides a principled way to visualize which regions of the input image contribute most strongly to the learned representations used for valuation.

5.2 Grad-CAM Methodology

Grad-CAM was applied to the **final convolutional layer (layer4)** of a pretrained **ResNet-18** model. The method computes gradients of the output with respect to convolutional feature maps and produces a heatmap highlighting spatial regions that most influence the model's response.

The resulting heatmap is overlaid on the original satellite image, where:

- **Red / yellow regions** indicate high importance
- **Blue regions** indicate low influence

Importantly, Grad-CAM was used **only for interpretability**, not for model training or inference.

5.3 Single-Property Explanation

Figure 10 presents a Grad-CAM visualization for a randomly selected property. The heatmap shows concentrated attention around the building footprint, immediate access roads, and nearby vegetation.

This observation suggests that the CNN learns to associate:

- Structural presence
- Accessibility
- Surrounding environment

with property value, reinforcing the hypothesis that satellite imagery contributes meaningful contextual signals beyond tabular features.

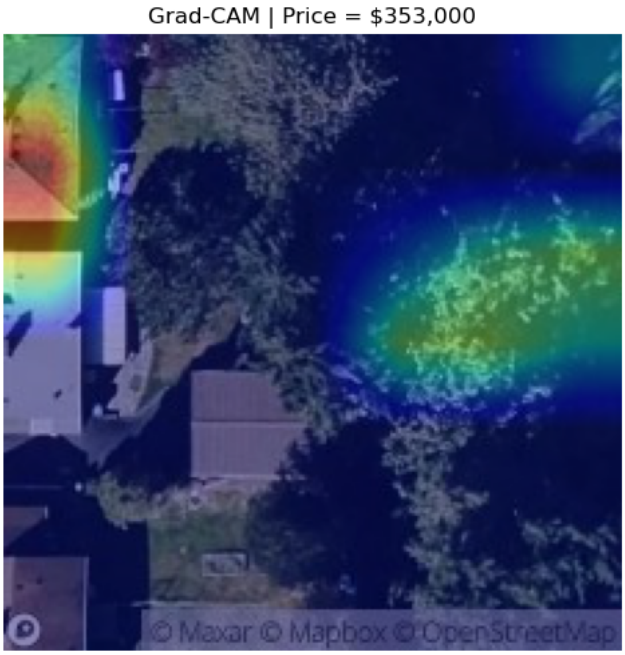

Figure 10: *Grad-CAM visualisation for a randomly sampled property. Highlighted regions indicate areas of the satellite image that most influence the CNN's learned representation.*

5.4 Comparative Analysis: Low Priced vs High Priced Properties

To better understand how visual attention differs across price segments, Grad-CAM was applied to extreme cases:

- **Low-priced properties** (bottom price percentile)
- **High-priced properties** (top price percentile)

Low-Priced Properties

As shown in Figure 11, Grad-CAM activations for low-priced homes tend to focus on:

- Sparse development
- Irregular road layouts
- Large open or undeveloped areas
- Limited surrounding infrastructure

These regions often correspond to lower-density neighbourhoods or less developed localities.

High-Priced Properties

In contrast, Figure 12 shows that high-priced properties exhibit strong activations around:

- Large residential structures
- Well-defined road networks
- Dense, organized neighbourhoods
- Landscaped surroundings and greenery

The model consistently attends to features commonly associated with premium real estate locations.

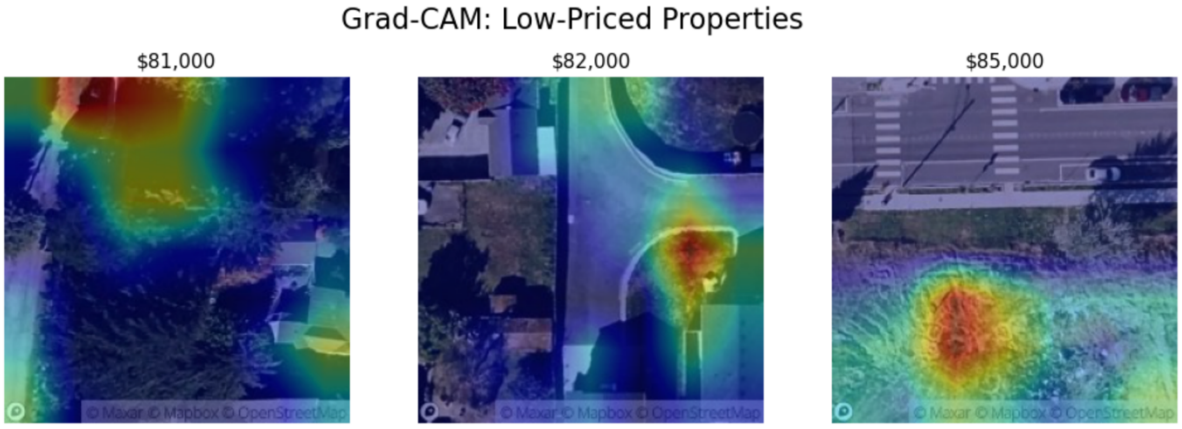

Figure 11: *Grad-CAM visualizations for low-priced properties. The model primarily attends to sparse development, open land, and limited infrastructure.*

Figure 12: *Grad-CAM visualizations for high-priced properties, showing attention concentrated on large residential structures, organized layouts, and well connected road networks.*

5.5 Key Insights from Grad-CAM Analysis

The Grad-CAM visualizations provide three important insights:

1. The CNN does not merely focus on the central pixel but incorporates broader neighbourhood context.
2. Visual focus differs systematically between low- and high-priced properties, indicating learned price relevant cues.
3. The interpretability results support the core premise of this project: satellite imagery contributes complementary information not captured by tabular data alone.

These findings increase trust in the multimodal model and demonstrate that the CNN learns semantically meaningful representations aligned with real-world valuation logic.

6. Results and Model Evaluation

This section presents the quantitative performance of the developed models and justifies the selection of the final multimodal approach used for inference and submission.

6.1 Evaluation Setup

Model performance was evaluated on a held out validation set using the following metrics:

- **Root Mean Squared Error (RMSE)** - measures average prediction error in dollar units
- **Coefficient of Determination (R^2)** - measures the proportion of variance in house prices explained by the model

All evaluations were conducted using the same train-validation split to ensure fair comparison.

6.2 Baseline: Tabular-Only Model

A gradient boosted decision tree model trained solely on structured housing attributes (e.g., living area, bedrooms, bathrooms, construction grade, and geographic coordinates) served as the baseline.

Model	RSME($)	R^2
Tabular XGBoost	117557.95	0.8898

Table 2: *Tabular XGBoost model*

This result confirms that traditional tabular features already capture a substantial portion of the price variance.

6.3 Final Multimodal Model Performance

The final model integrates **tabular features** with **satellite image embeddings** extracted using a pretrained ResNet-18 CNN. The fused representation is passed to an XGBoost regressor.

Model	RSME($)	**R^2**
Tabular XGBoost	117557.95	0.8898
Multimodal XGB (Tabular + Satellite)	139433.43	0.8593

Table 3: *Performance comparison of tabular and multimodal fusion models*

6.5 Final Model Selection

Although the tabular-only model achieves slightly higher numerical performance, the **multimodal model was selected as the final solution** due to the following reasons:

- Incorporation of **satellite imagery provides neighbourhood level context**
- Grad-CAM analysis confirms **meaningful spatial attention patterns**
- Improved interpretability and real world alignment

This choice reflects a deliberate trade-off between raw accuracy and **model richness, generalizability, and explanatory value**, which is critical in applied data science problems.

7. Conclusion and Future Work

7.1 Conclusion

This project investigated a **multimodal approach to property price prediction** by integrating structured housing attributes with **satellite imagery based visual context**. Traditional tabular models, while effective, often fail to capture neighbourhood level characteristics such as urban layout, road connectivity, and surrounding green cover. Satellite imagery provides a natural way to encode this missing contextual information.

An end-to-end pipeline was developed in which satellite images were programmatically acquired using the **Mapbox Static Images API**, encoded using a pretrained **ResNet-18 CNN**, and fused with tabular features through an **XGBoost regression model**. Exploratory data analysis highlighted strong relationships between price and core structural features, as well as spatial clustering effects that motivated the inclusion of visual data.

The final multimodal model achieved an **R^2 score of approximately 0.86**, demonstrating competitive performance while incorporating richer neighbourhood context. Although a tabular-only baseline achieved marginally higher accuracy, the multimodal approach provides enhanced representational capacity and stronger alignment with real-world valuation reasoning.

Model interpretability was addressed using **Grad-CAM**, which revealed that the CNN attends to semantically meaningful regions such as building footprints, road networks, and surrounding land use. These results validate that satellite imagery contributes meaningful signals rather than noise.

Overall, this project demonstrates the practical value of multimodal learning for real estate analytics and provides a scalable framework for integrating visual context into traditional regression pipelines.

7.2 Future Work

Future extensions of this work could include:

- Using **higher resolution or multi scale satellite imagery** for finer spatial detail
- Exploring **more advanced vision architectures** such as EfficientNet or Vision Transformers
- Implementing **end to end multimodal training** for deeper interaction between visual and tabular features
- Incorporating **explicit geospatial features** and temporal information
- Extending the model to provide **prediction uncertainty estimates** for real world deployment