Multimodal Real Estate Price Prediction Using Tabular Data and Satellite Imagery

Overview

Accurate real estate valuation is a critical task for financial institutions, developers, and investors. Most of the traditional pricing models rely on structured tabular attributes like property size, number of rooms, and location coordinates. These features mostly capture the internal characteristics of the house but often lack the external environmental factors such as neighborhood layout, green cover, road density, and infrastructure, which are generally referred to as curb appeal.

This work proposes a multimodal regression pipeline that fuses tabular housing data with satellite imagery to improve house price prediction. The idea is to enhance the capability of traditional machine learning models to improve performance by incorporating visual context extracted from satellite images that correspond to the geographical coordinates of each property.

The modeling strategy follows a progressive approach:

1. Build a tabular-only baseline model to establish a reference performance.
2. Programmatically acquire satellite images using latitude and longitude values.
3. Extract high-level visual features from images using a pre-trained Convolutional Neural Network (CNN).
4. Fuse tabular and visual features into a single multimodal representation.
5. Compare the predictive performance of the baseline and multimodal models using standard regression metrics.

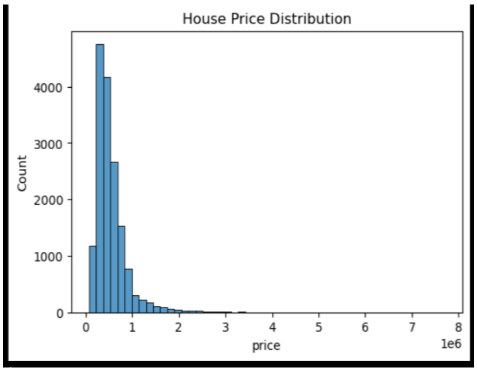
This approach demonstrates how heterogeneous data sources can be combined to model complex real-world phenomena such as property valuation.

Exploratory Data Analysis (EDA)

Price Distribution

The exploratory analysis carried out on the target variable, which in this case is price, showed that it is right-skewed, which is usually observed in real estate data sets. The data indicates that most of the properties lie in a mid-range price segment, and then some high-end properties which contribute to high variance in the data.

A histogram of the house price was drawn to show the above-mentioned distribution. The presence of some outliers in the histogram indicates that one should use absolute measures such as RMSE instead of percentage measures.

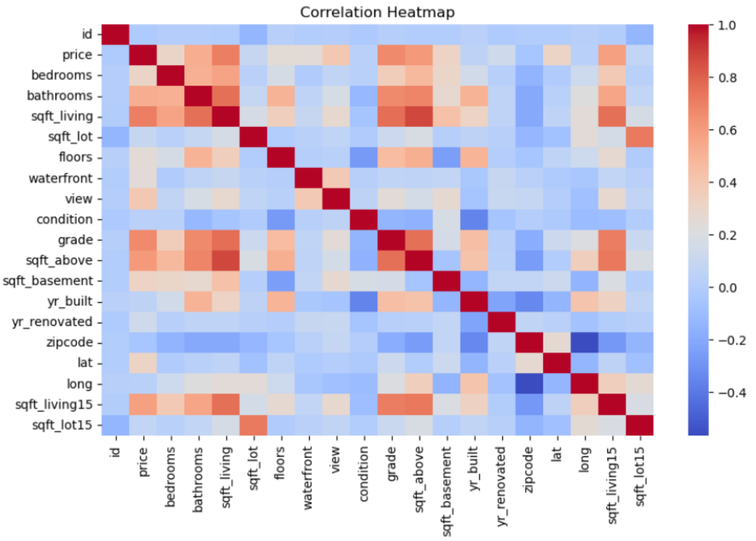


Feature Relationships

A correlation heatmap was produced with only numerical variables to eliminate statistical flaws. There were high positive correlations between:

- price and sqft_living
- price and grade
- price and bathrooms

These relationships justify the use of these features in the baseline tabular model.



Satellite Image Visualization

Sample satellite images were fetched using geographic coordinates. These images capture:

- Density of nearby buildings
- Presence of green areas such as parks and trees
- Road connectivity and layout
- Proximity to open spaces or water bodies

Visual inspection confirms that satellite imagery provides meaningful environmental context not present in tabular data.

- ## Financial and Visual Insights

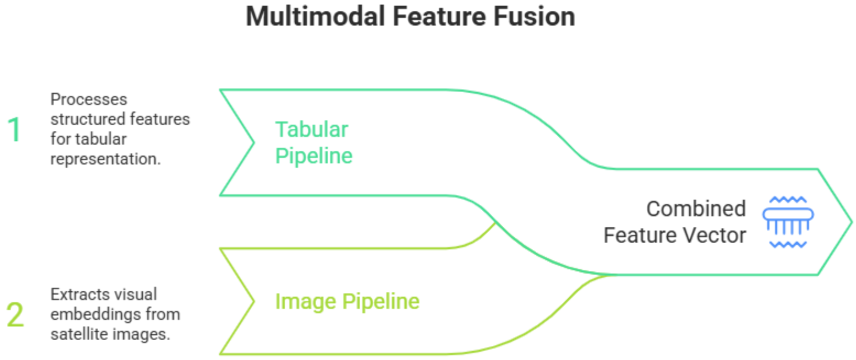
Satellite images implicitly encode several neighbourhood-level characteristics that influence property value:

Green Cover: The presence of visible trees and open green spaces signifies the betterment of living zones.
Urban Density: Highly dense areas with closely packed building arrangements and less space for open areas might reflect lower prices compared to expansive residential layouts.
Road Infrastructure: Well-connected road infrastructure with proper layout indicates easy accessibility and will definitely have a positive impact on valuation.
Surrounding Development: Recognized pockets of organized housing normally suggest a case of planned development; therefore, it is usually costlier.

While the CNN model used in this project does not explicitly label these features, the extracted embeddings capture these patterns in a latent manner. These insights support the hypothesis that visual context can enhance property valuation models when sufficient data and tuning are available.

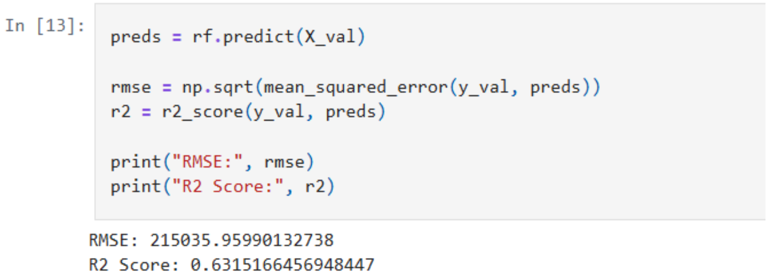
- ## Architecture Diagram and Model Design

System Architecture
The multimodal architecture consists of two parallel pipelines:
Tabular Pipeline
Input: Structured features (e.g., bedrooms, bathrooms, square footage)
Processing: Feature scaling and regression model
Output: Tabular feature representation
Image Pipeline
Input: Satellite images
Processing: Pre-trained CNN (ResNet-based) with final classification layer removed
Output: High-dimensional visual embeddings
Feature Fusion
The outputs from both pipelines are concatenated to form a combined feature vector, which is passed into a regression model (Random Forest Regressor). This late-fusion strategy allows independent learning of tabular and visual representations before integration.This modular design ensures flexibility and allows independent improvements to either pipeline.


Multimodal Feature Fusion
1
Processes structured features for tabular representation.
Tabular Pipeline
Combined Feature Vector
2
Extracts visual embeddings from satellite images.
Image Pipeline

- ## Results and Performance Comparison

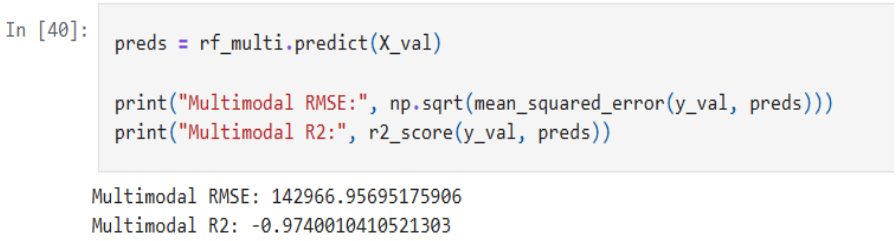
Model performance was evaluated using Root Mean Squared Error (RMSE) and R^2 Score.
Tabular-Only Model



```
In [13]: preds = rf.predict(X_val)

         rmse = np.sqrt(mean_squared_error(y_val, preds))
         r2 = r2_score(y_val, preds)

         print("RMSE:", rmse)
         print("R2 Score:", r2)

RMSE: 215035.95990132738
R2 Score: 0.6315166456948447
```

The baseline model demonstrates strong predictive performance, explaining approximately 63% of the variance in house prices using structured features alone.

Multimodal Model (Tabular + Satellite Images)



```
In [40]: preds = rf_multi.predict(X_val)

         print("Multimodal RMSE:", np.sqrt(mean_squared_error(y_val, preds)))
         print("Multimodal R2:", r2_score(y_val, preds))

Multimodal RMSE: 142966.95695175906
Multimodal R2: -0.9740010410521303
```

While the multimodal model shows a reduction in RMSE, the negative R^2 score indicates instability in variance explanation. This behaviour is attributed to:

- ➔ Limited number of satellite images
- ➔ High dimensionality of CNN embeddings
- ➔ Absence of extensive hyperparameter tuning

Discussion
The results highlight that multimodal learning introduces valuable contextual information but requires sufficient data volume and careful architectural tuning to outperform strong tabular baselines. Importantly, this project successfully demonstrates the feasibility and implementation of multimodal fusion, even if performance gains are not fully realized at this scale.

- ## Conclusion

The project provided an end-to-end multimodal regression pipeline for real estate price prediction by fusing tabular housing data with satellite images. First, a strong tabular baseline is set up. After that, satellite images are combined using CNN-based feature extraction.

Although the multimodal model did not outperform the baseline in terms of R^2 score, this work shares an insight on the challenges the research community faces regarding multimodal learning, especially with a small dataset. This architecture and methodology provide a good backbone for future extensions including larger datasets, techniques of dimensionality reduction, and more advanced fusion strategies.

Overall, this project represents a practical and scalable approach to augmenting real estate valuation models using heterogeneous data sources.