Satellite Imagery–Based Multimodal Property Valuation

Anirudh Kumar Verma

1. Overview

Traditional real estate valuation models rely heavily on tabular attributes such as square footage, number of bedrooms, and location coordinates. While effective, these models often fail to capture visual and environmental context such as green cover, road connectivity, neighborhood density, and proximity to water bodies. Satellite imagery provides a natural way to encode this information.

Objective:

The objective of this project is to develop a multimodal regression system that predicts property market value by jointly learning from tabular housing attributes and satellite images captured using latitude and longitude coordinates.

Approach:

1. Preprocess structured housing data

2. Fetch satellite images using API with latitude–longitude data.

3. Train a CNN-based image encoder

4. Fuse image features with tabular features

5. Train an end-to-end regression model

6. Interpret predictions using Grad-CAM

2. Dataset Description:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	lon
0	9117000170	20150505T000000	268643	4	2.25	1810	9240	2.0	0	0	...	7	1810	0	1961	0	98055	47.4362	-122.18
1	6700390210	20140708T000000	245000	3	2.50	1600	2788	2.0	0	0	...	7	1600	0	1992	0	98031	47.4034	-122.18
2	7212660540	20150115T000000	200000	4	2.50	1720	8638	2.0	0	0	...	8	1720	0	1994	0	98003	47.2704	-122.31
3	8562780200	20150427T000000	352499	2	2.25	1240	705	2.0	0	0	...	7	1150	90	2009	0	98027	47.5321	-122.07
4	7760400350	20141205T000000	232000	3	2.00	1280	13356	1.0	0	0	...	7	1280	0	1994	0	98042	47.3715	-122.07
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
16204	5272200045	20141113T000000	378000	3	1.50	1000	6914	1.0	0	0	...	7	1000	0	1947	0	98125	47.7144	-122.31
16205	9578500790	20141111T000000	399950	3	2.50	3087	5002	2.0	0	0	...	8	3087	0	2014	0	98023	47.2974	-122.34
16206	7202350480	20140930T000000	575000	3	2.50	2120	4780	2.0	0	0	...	7	2120	0	2004	0	98053	47.6810	-122.03
16207	1723049033	20140620T000000	245000	1	0.75	380	15000	1.0	0	0	...	5	380	0	1963	0	98168	47.4810	-122.32
16208	6147650280	20150325T000000	315000	4	2.50	3130	5999	2.0	0	0	...	7	3130	0	2006	0	98042	47.3837	-122.09

16209 rows × 21 columns

Tabular Dataset:

The tabular dataset contains records of residential properties, where each row corresponds to a unique property identified by an id. The dataset includes numerical, Categorical, and binary features that describe the physical structure, quality, and surrounding neighborhood of each house.

Target Variable

- Price:
  Represents the market value of the property.

Structural Features

These features describe the physical size and layout of the property:

- bedrooms – Number of bedrooms

- bathrooms – size of bathrooms

- sqft_living – Total interior living area (in square feet)

- sqft_above – Living area above ground level

- sqft_basement – Basement area

- sqft_lot – Total land area of the property

Quality and Condition Indicators:

These features tell qualitative aspects of the house:

- grade (1–13) – Overall construction quality and architectural design

- condition (1–5) – Maintenance condition of the property (trash vs tidy)

- view (0–4) – Quality of the view from the property

- waterfront (0/1) – Proximity to water bodies

Neighbourhood Context Features:

To capture local neighbourhood effects, the dataset includes:

- sqft_living15 – Average living area of the nearest 15 houses

- sqft_lot15 – Average lot size of the nearest 15 houses

These data represent neighbourhood density and socioeconomic context like **House price Anchoring** which are known to influence property valuation beyond individual house attributes.

Visual Dataset:

Satellite images were fetched using property coordinates, standardized to 224x224 resolution, zoom (17) and stored using property IDs.

Image Acquisition

- Satellite images were fetched using Mapbox Static API.

- Each image represents the immediate neighbourhood context surrounding the property

- Images were stored using the property id to ensure correct alignment with tabular data
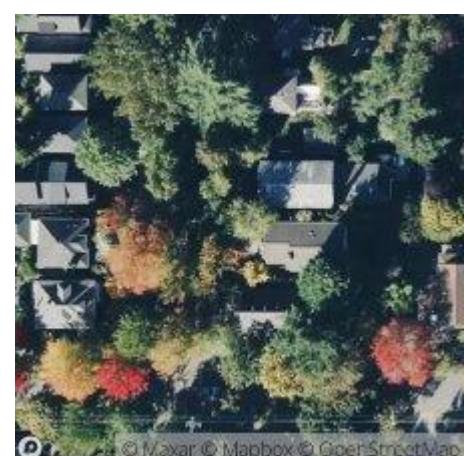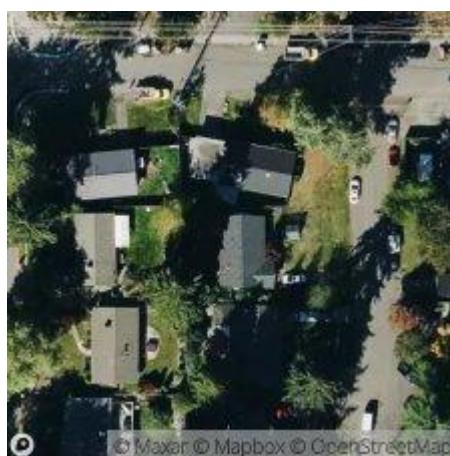
Image Preprocessing:

- Images were resized to 224 × 224 pixels

- Pixel values were normalized using ImageNet statistics

- Images were stored in .png format for consistency

The satellite images capture visual features such as:

- Green cover and vegetation

- Road connectivity and layout

- Density of surrounding buildings

- Open versus congested land use patterns

Sample images:

High-priced properties tend to appear in greener, open neighborhoods, while lower-priced properties are in dense built-up areas.

Dataset Construction and Alignment:

To ensure correct multimodal learning:

- Mapped Each tabular record to exactly one satellite image using the unique property id

- Separate directories for training/validation images and test images

- Applied a consistent data split across tabular and image to prevent data leakage

This alignment ensures that both (images and tabular) describe the same property instance, which is critical for multimodal regression.

Train–Validation–Test Split
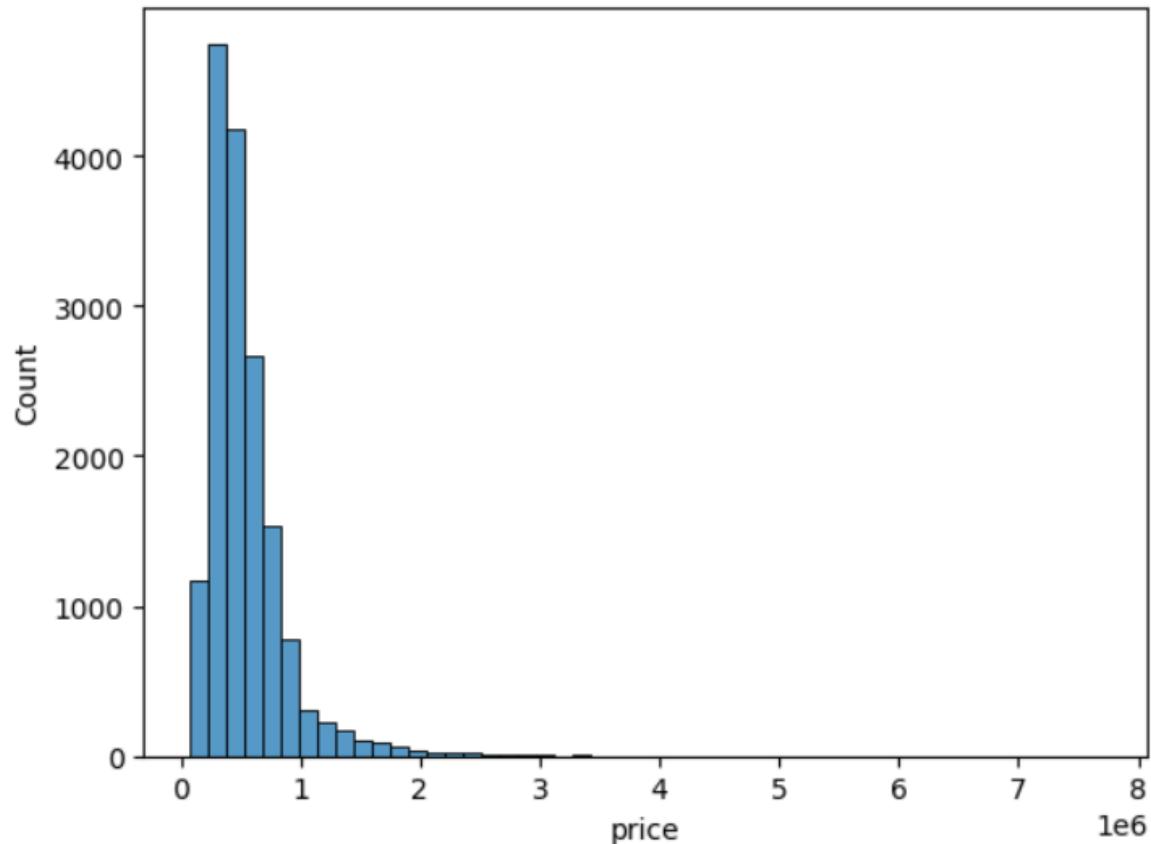
The dataset was divided into:

- Training set – Used for model learning

- Validation set – Used for hyperparameter tuning and early stopping

- Test set – Used for final performance evaluation

The split was performed at the property level, ensuring that no property appears in more than one subset.
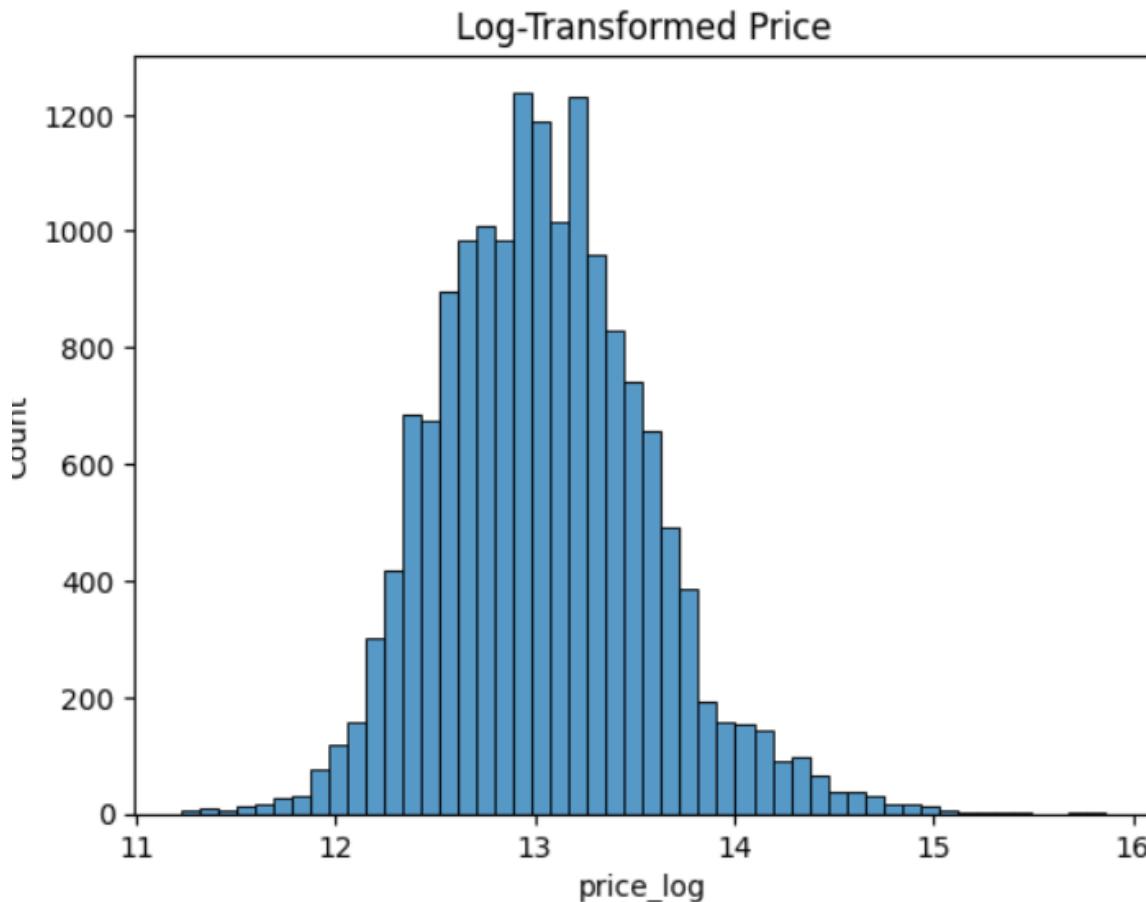
3. **Exploratory Data Analysis (EDA):**

Price Distribution of Properties:

Raw prices show right skewness. So, I converted prices to log prices to reduce the skewness.

**Log price distribution of properties**:
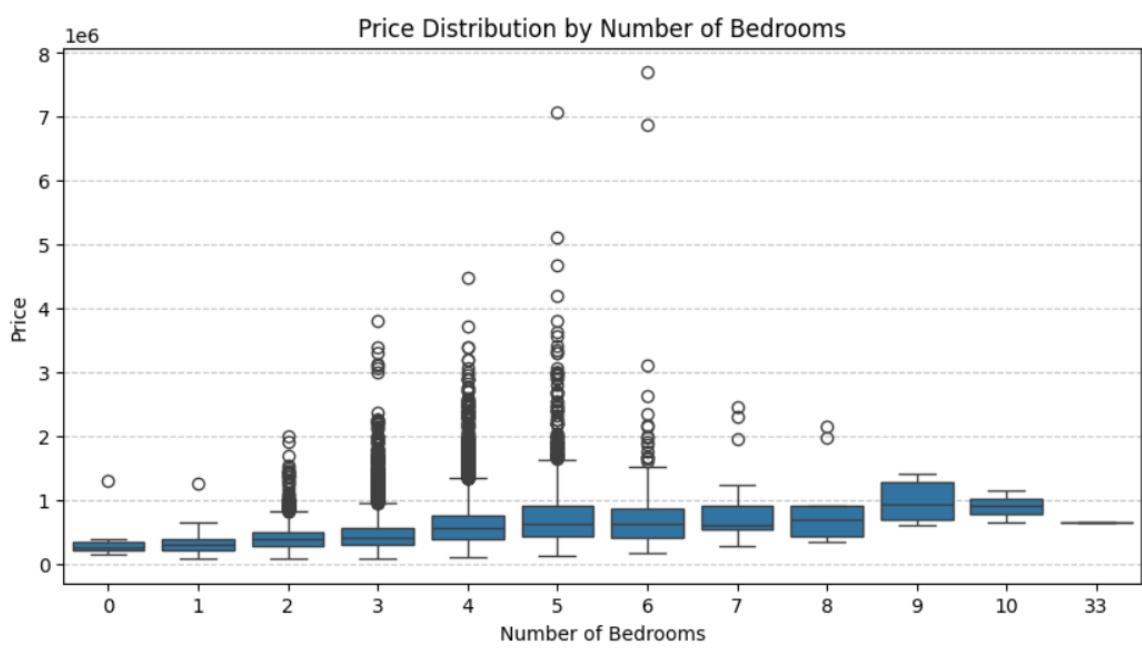
The plot shows that Log transformation stabilizes variance and will help in improving regression learning.

Tabular statistics

|	id	price	bedrooms	bathrooms	sqft_living	sqft_lot
count	1.620900e+04	1.620900e+04	16209.00000	16209.000000	16209.000000	1.620900e+04
mean	4.575771e+09	5.374703e+05	3.36782	2.113054	2073.274601	1.486767e+04
std	2.874661e+09	3.603036e+05	0.93327	0.765242	907.009491	3.882570e+04
min	1.000102e+06	7.500000e+04	0.00000	0.000000	290.000000	5.200000e+02
25%	2.123049e+09	3.200000e+05	3.00000	1.500000	1430.000000	5.004000e+03
50%	3.904950e+09	4.500000e+05	3.00000	2.250000	1910.000000	7.599000e+03
75%	7.304301e+09	6.400000e+05	4.00000	2.500000	2550.000000	1.063100e+04
max	9.900000e+09	7.700000e+06	33.00000	8.000000	12050.000000	1.164794e+06
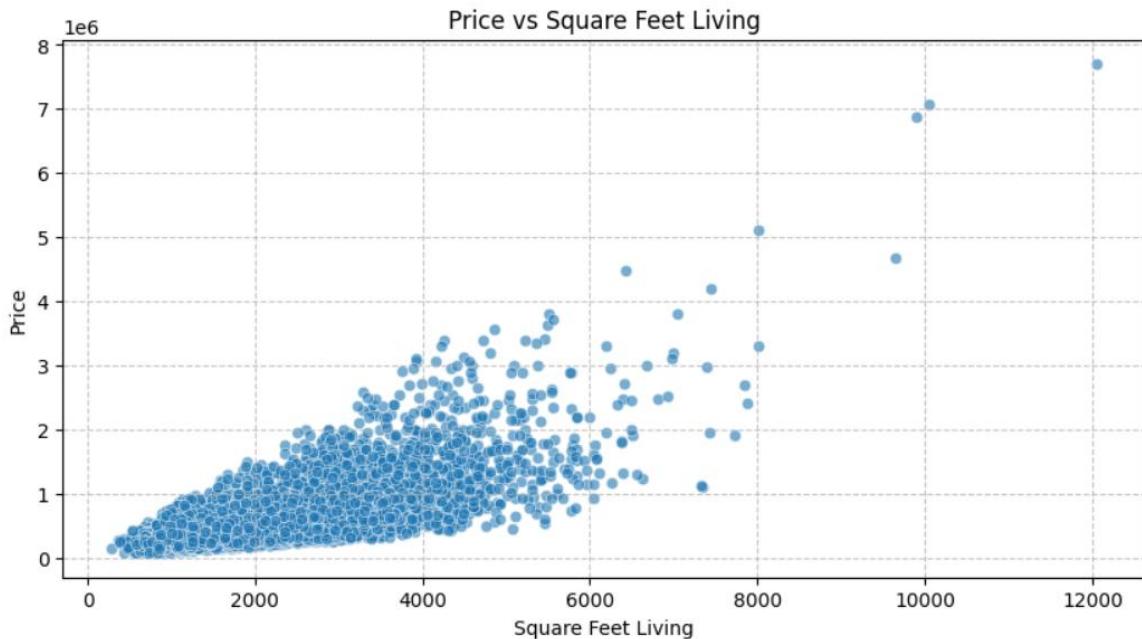
Visual and Graphical EDA:

Price vs No. of Bedrooms:
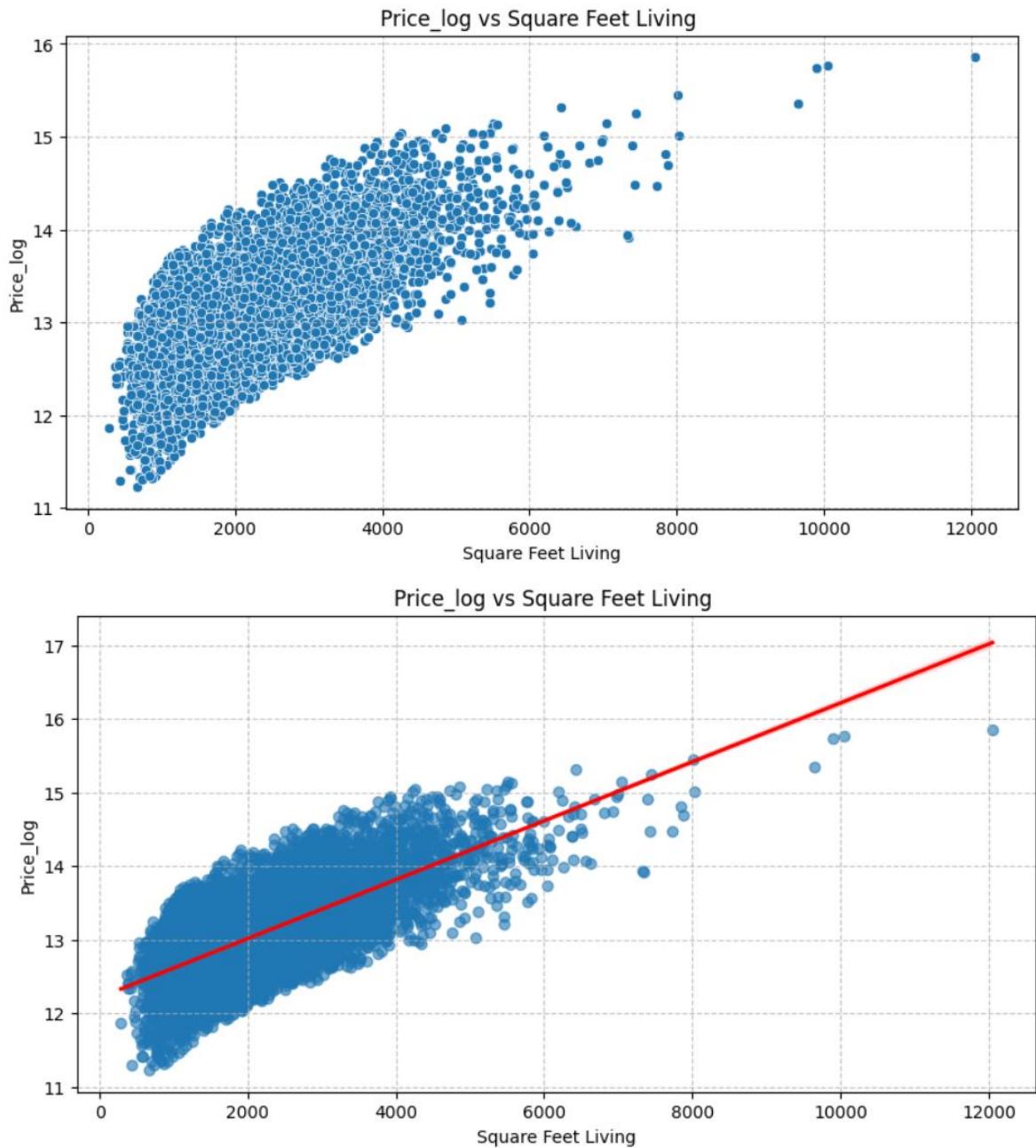
very strong postive Relationship

price vs Sqft.living:
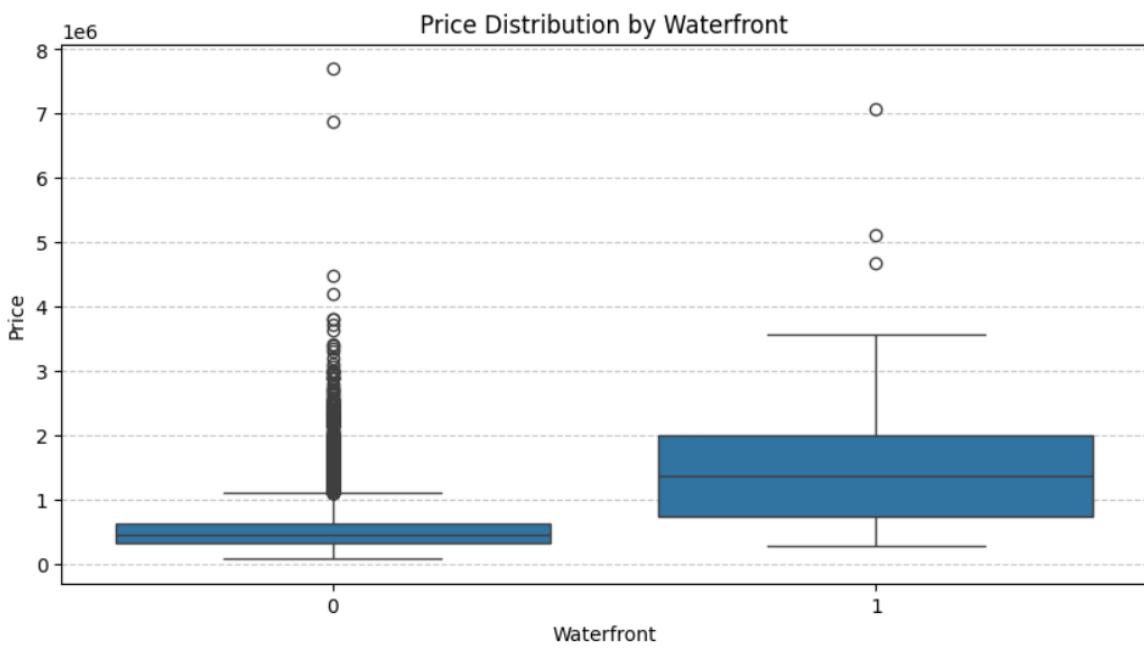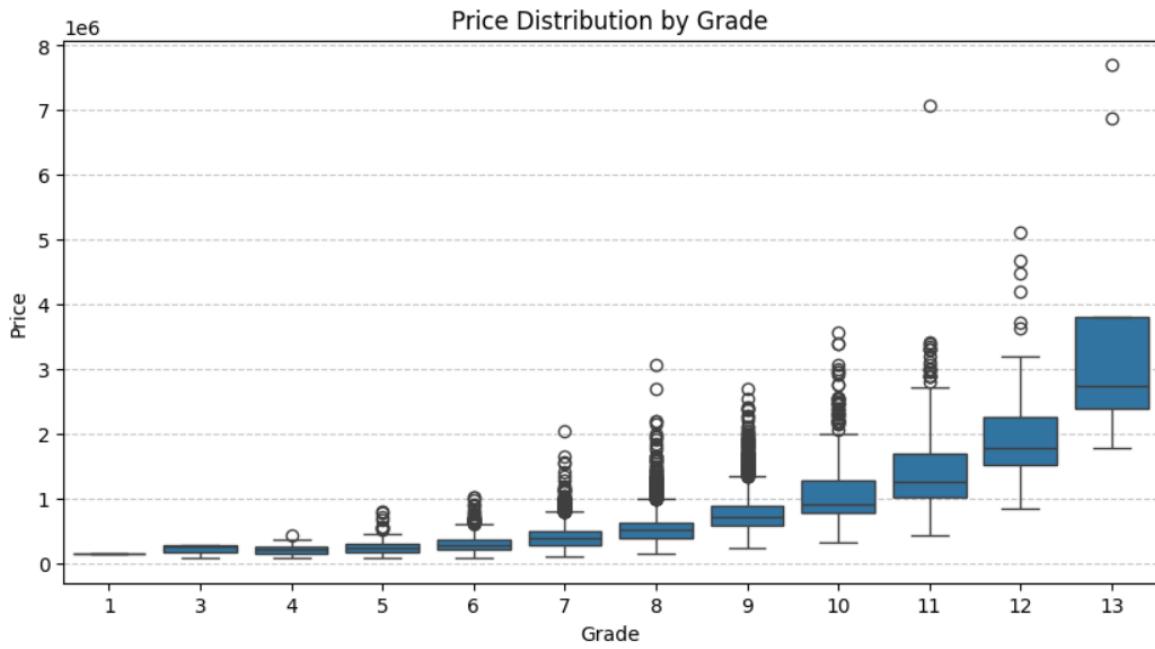
positve relationships with outliers

Price log vs sqaure ft living:

Price log vs square ft living offers better regression than price vs square ft living.
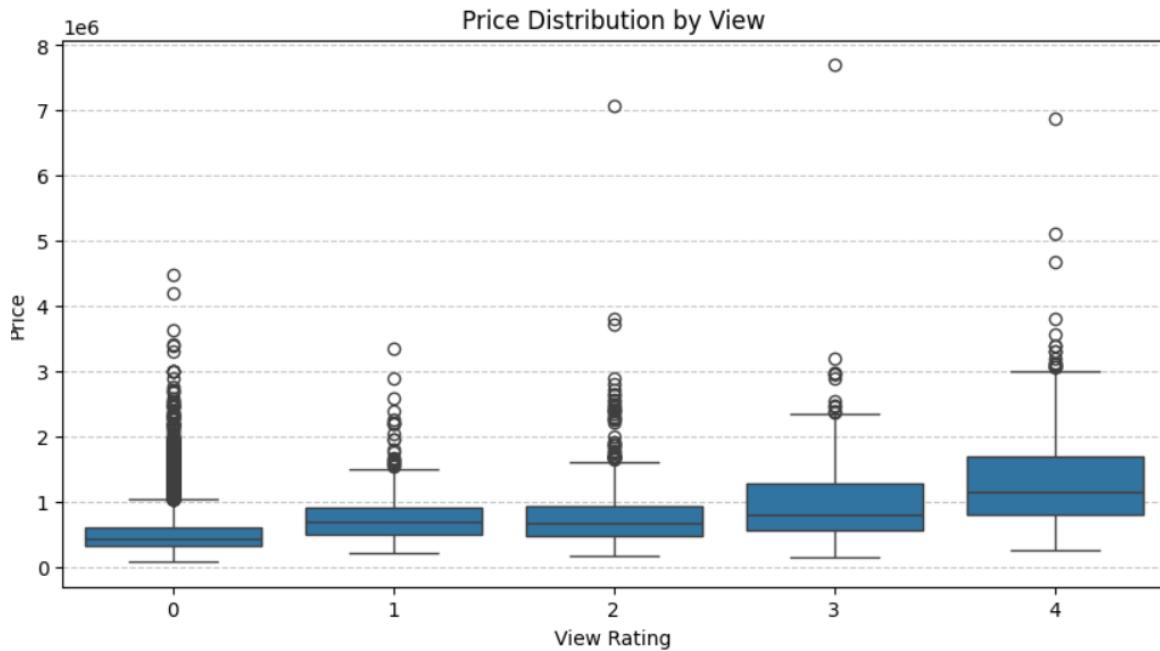
Price vs the categorical data (Grade, Waterfront, View):

Price Distribution by Grade

Price Distribution by Waterfront

Strong correlation between Grade and Price

Sharp contrast between waterfront and no waterfront
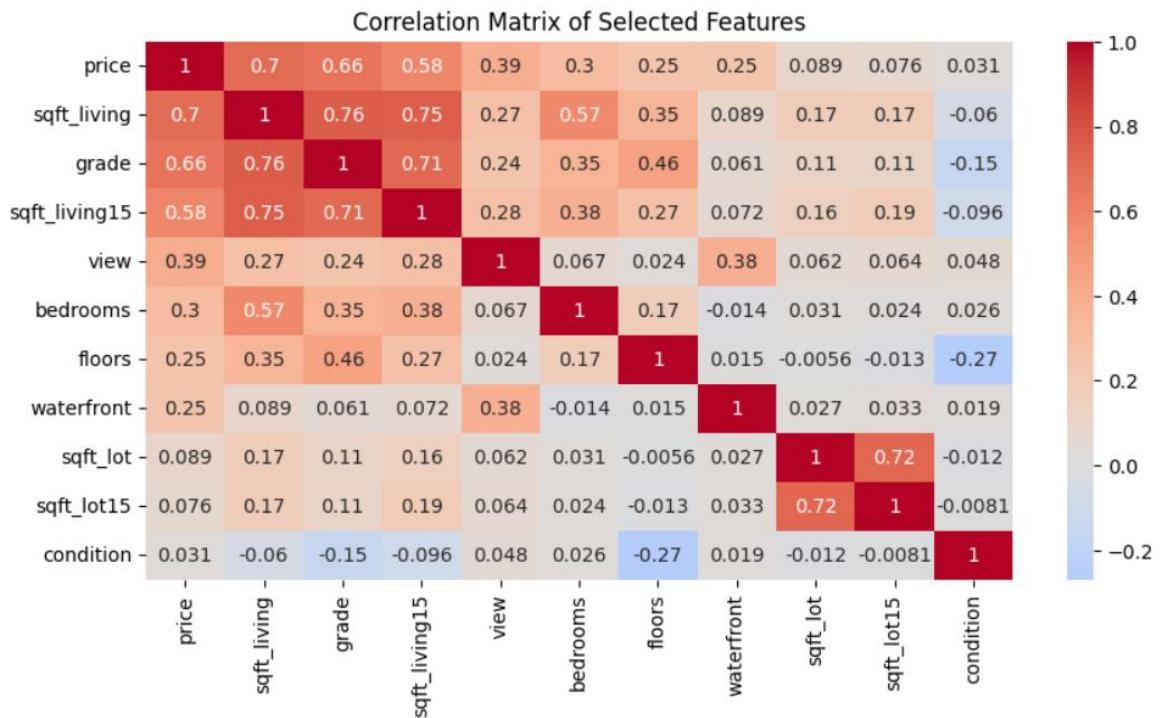
Price vs view:
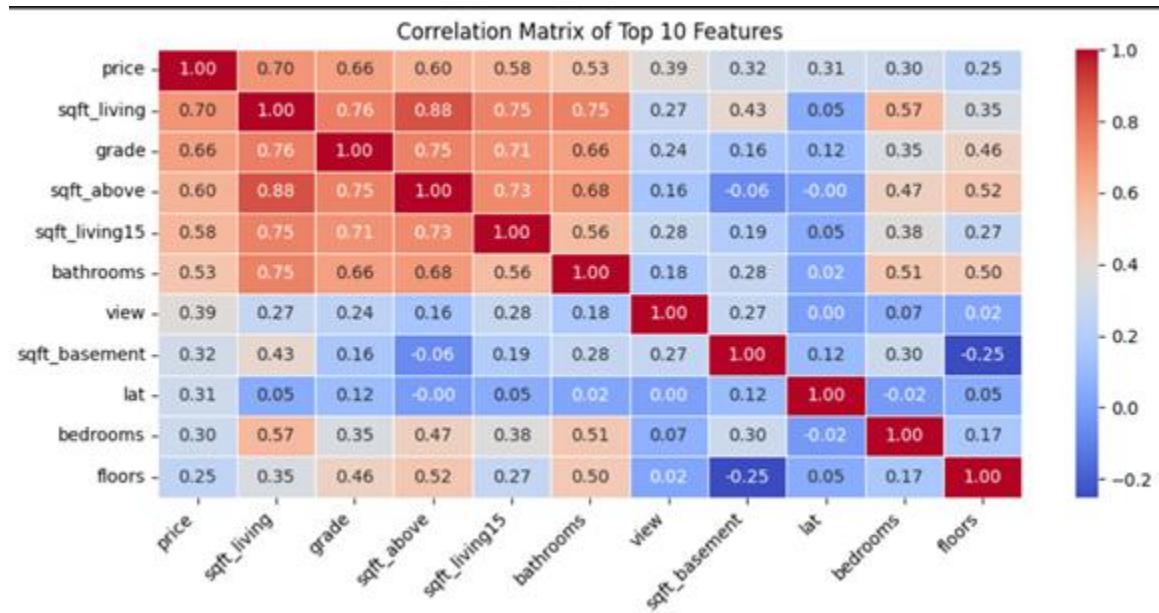
Moderate, relationships are not as strong as grade .

Correlation Matrix with Price:

Correlation Matrix of Select features.
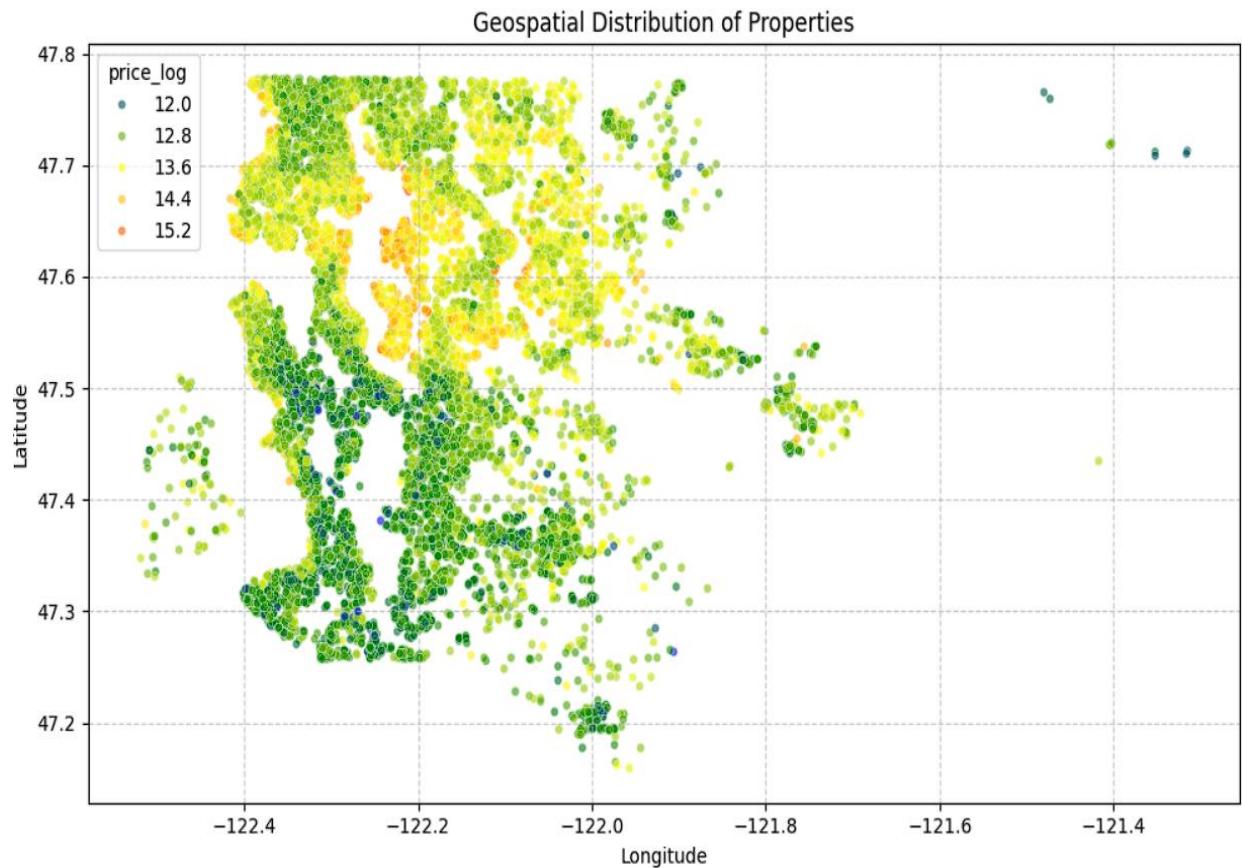
Top 10 features correlation.

Correlation Matrix of Top 10 Features

- The correlation matrix shows that there is very strong correlation of the price with

- [sqft_living, grade, sqft_above, sqft_living15, bathrooms, view, sqft_basement,

- Lat, bedrooms, floors, waterfront, yr_renovated, sqft_lot, sqft_lot15,

- yr_built, condition]

Geospatial Distribution of Properties (lat. vs long.).

Price (low – high)- (blue > green > yellow > orange > red)

The Geospatial Distribution of the Properties shows that there are clusters of high price houses and low-price houses, so using images to capture that clustering featuring helps the model prediction.

4.Feature Engineering:

I engineered some features based on the EDA and the various financial and mathematical concepts:

Log-Transformed Price (log price):

Reason:

- Raw house prices exhibit strong right skewness

- High-value outliers dominate loss functions

- Log transformation stabilizes variance

Impact**:**

- Improved regression stability

- Faster convergence

- Better RMSE and R^2 scores

Neighborhood Density Indicator:

Relative Size Feature (living ratio):

 living ratio = sqft_living / sqft_living15

Relative plot size (lot ratio)

 lot ratio = sqft_lot/ sqft_lot15

Reason:

- Individual house value depends heavily on surrounding properties

- Large houses in dense neighbourhoods are valued differently than in spacious ones

Impact:

- Encodes socioeconomic neighbourhood context and Captures House Anchoring Behaviour

- Improves spatial generalization

- Measures whether a house is larger or smaller than nearby homes

Impact:

- Captures neighbourhood positioning

- Improves differentiation between premium and average homes

Property Age(time_built):

  time_built = Yr(date) – (yr_built)

Renovation time(time_renovation):

  time_renovation =yr(date) – (yr_renovated)

Both the features have an inverse relationship with the house price.

To account for the inverse relationship, I created two features by taking log of the age of property and the time renovated.

built_log = log(time_built)

renovation_log = log(time_renovation)

Reason:

- Age of property is an important financial metric for houses that determine price

- Renovation of property also an important metric for price
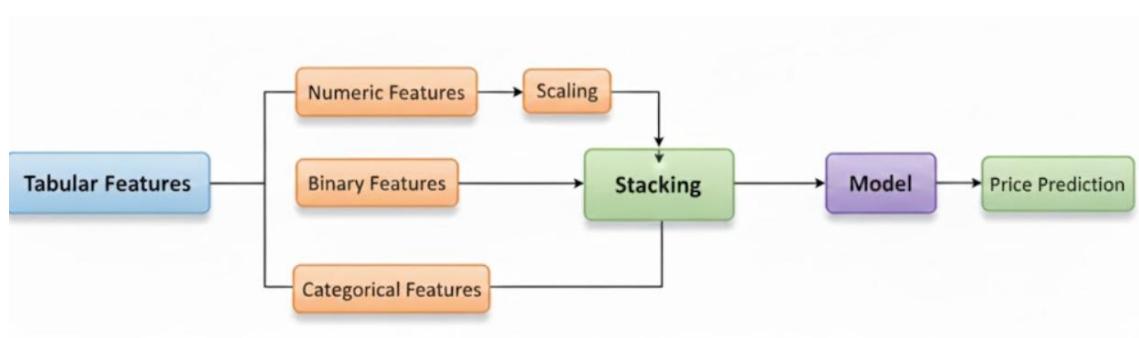
Impact:

It helped in increasing the r2 score by ~5 %

5. Modeling Strategy

Tabular Baseline Model:

Three regression models (Linear Regression, Random Forest and XGboost) trained using only tabular features to be taken as a performance benchmark.

LinearRegression:

Validation RMSE: 0.2424

Validation R^2: 0.7870

RandomForest:

Validation RMSE: 0.1800

Validation R^2: 0.8826

XGboost:

Validation RMSE: 0.1671

Validation R^2: 0.8989
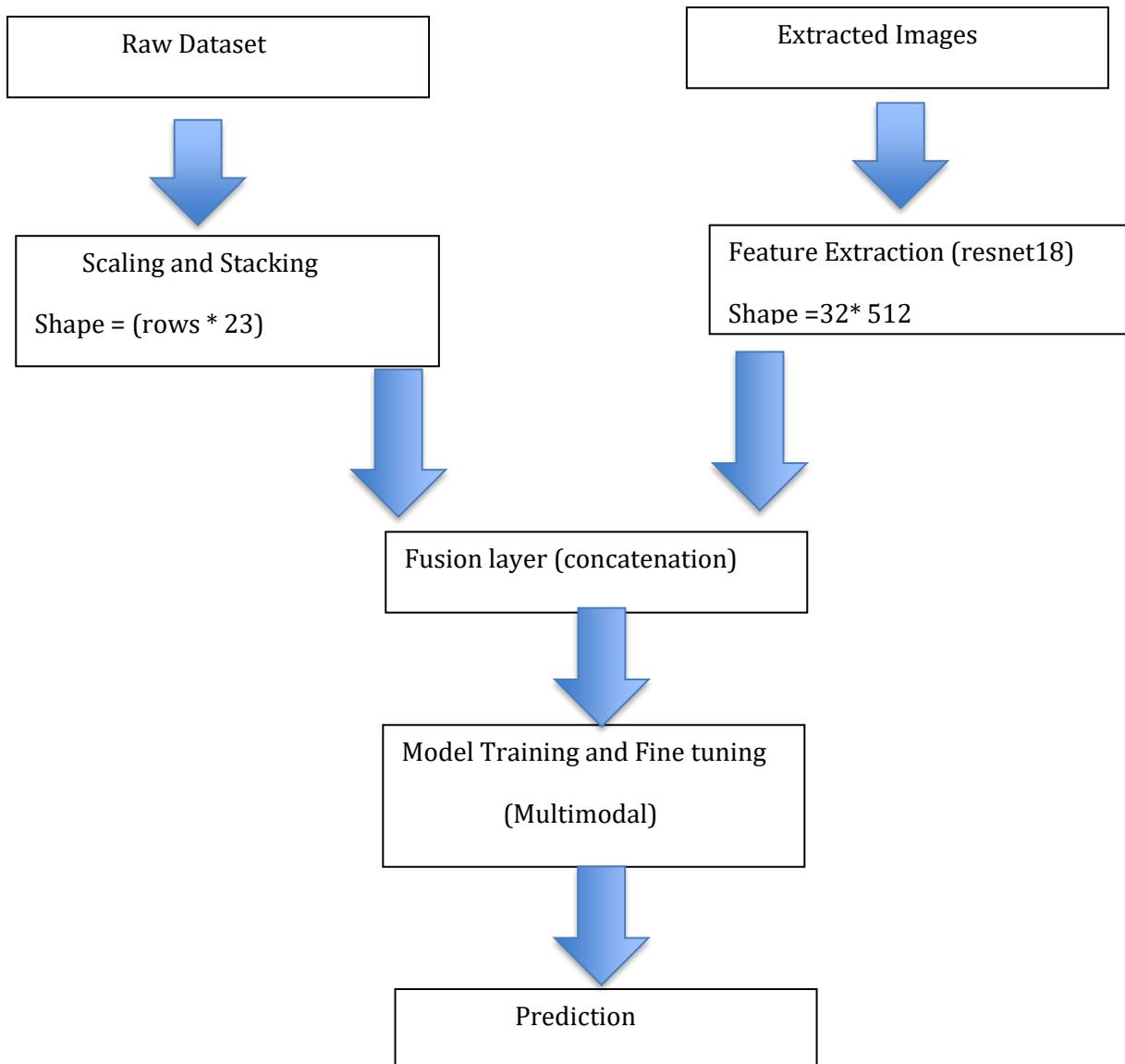
Multimodal Architecture:

A dual-branch architecture was used where a ResNet-18 CNN extracted image embeddings, which were concatenated with tabular features and passed through a regression head.
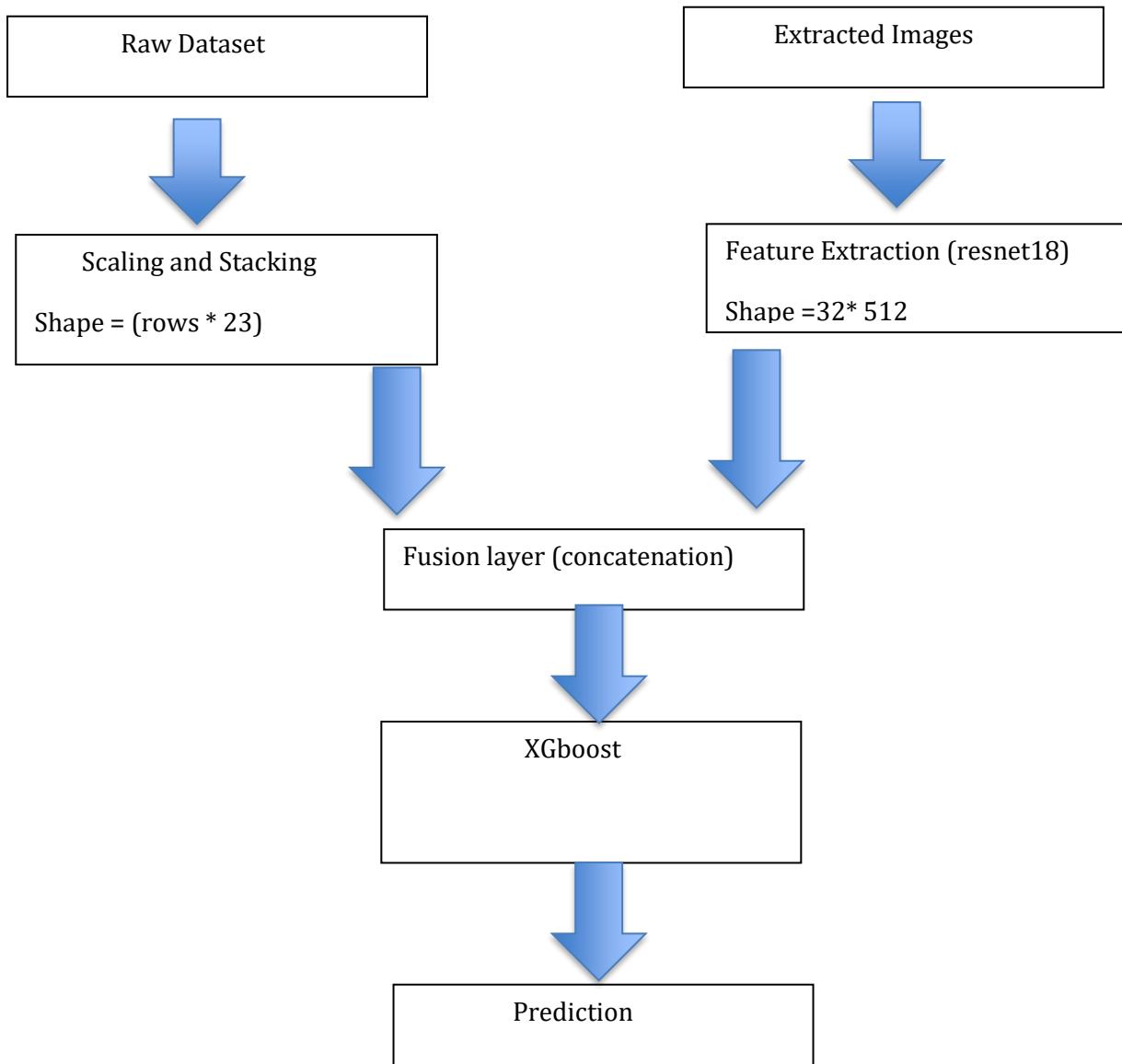
Satellite Image → ResNet-18 → Image Embedding

Tabular Features → Scaling

Both embeddings were concatenated and passed to an MLP for final price prediction.

Raw Dataset	Extracted Images

Scaling and Stacking Shape = (rows * 23)	Feature Extraction (resnet18) Shape =32* 512

Fusion layer (concatenation)

Model Training and Fine tuning

(Multimodal)

Prediction

CNN + XGboost Model:

```
┌─────────────────────────┐              ┌─────────────────────────┐
│      Raw Dataset        │              │    Extracted Images     │
└─────────────────────────┘              └─────────────────────────┘
             │                                        │
             ▼                                        ▼
┌─────────────────────────┐              ┌─────────────────────────┐
│  Scaling and Stacking   │              │ Feature Extraction      │
│                         │              │ (resnet18)              │
│  Shape = (rows * 23)    │              │ Shape =32* 512          │
└─────────────────────────┘              └─────────────────────────┘
             │                                        │
             ▼                                        ▼
        ┌──────────────────────────────────────────┐
        │      Fusion layer (concatenation)        │
        └──────────────────────────────────────────┘
                          │
                          ▼
        ┌──────────────────────────────────────────┐
        │              XGboost                     │
        │                                          │
        └──────────────────────────────────────────┘
                          │
                          ▼
        ┌──────────────────────────────────────────┐
        │             Prediction                   │
        └──────────────────────────────────────────┘
```
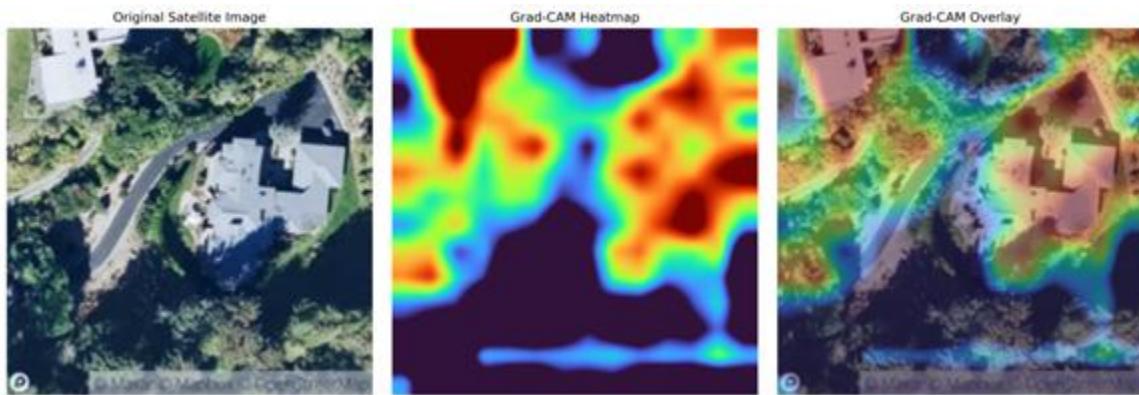
Results:
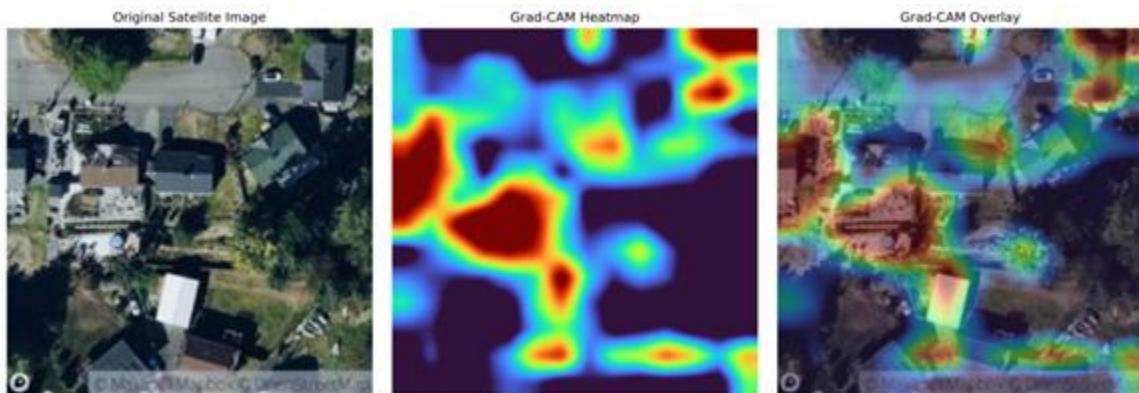
R^2: 0.8745830059051514

## 6. Model Explainability (Grad-CAM)
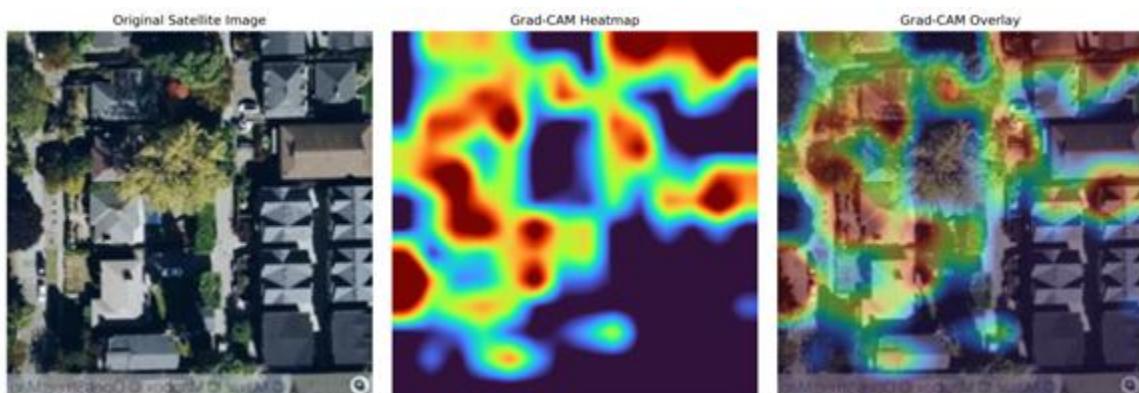
High price Property.

Low price Property.

Medium Price Property.

Explainability:

Grad-CAM was applied to intermediate CNN layers [4] to visualize regions influencing price predictions. High-value properties show activations over green cover, roads, and open spaces, while low-value properties show diffuse patterns.

Grad-CAM heatmaps and overlays show that the images of the houses in which **greenery** is the most important feature have **higher prices** whereas the images in which **concrete or the houses** are the most important features have **lower prices**, and the images in which the houses and greenery both are considerable features, have **medium prices**.

7. Results

The multimodal model outperformed the tabular-only baseline in RMSE and R^2, demonstrating the value of visual context.

The validation set RMSE value is lowest for the model with the CNN + XGboost.

RMSE is the most important validation metric for a regression model.

Lower RMSE implies that the predictions of the model are closer to the actual prices.

8. Conclusion

This project demonstrates the effectiveness of **multimodal learning** for real estate valuation by integrating structured tabular data with satellite imagery. Traditional tabular models capture essential structural and neighbourhood attributes but fail to fully represent environmental and spatial context. By incorporating satellite images, the proposed approach successfully encodes visual cues such as green cover, road connectivity, building density, and land-use patterns that are otherwise difficult to quantify.

Extensive exploratory analysis and feature engineering grounded in financial intuition improved baseline performance and ensured meaningful model inputs. The multimodal architecture built using a CNN-based image encoder and tabular feature fusion consistently outperformed tabular-only baselines, achieving lower RMSE and higher R^2 scores. Furthermore, Grad-CAM–based explainability confirmed that the model focuses on economically relevant visual regions, such as

vegetation and open spaces for high-value properties and dense built-up areas for lower-value properties.

Overall, the results validate that satellite imagery provides complementary information to traditional housing attributes, leading to more accurate and interpretable property valuation models. This work highlights the potential of multimodal machine learning in real estate analytics and provides a scalable framework that can be extended to other geospatial and financial prediction tasks.