Satellite Imagery-Based Property Valuation

1. Overview of Modelling Strategy:

Objective: Predict residential property prices using **tabular + satellite imagery data**.

Pipeline:

- Data Cleaning and Exploratory Data Analysis (EDA)
- Tabular preprocessing & domain-driven feature engineering
- Satellite image preprocessing & CNN-based feature extraction
- Dimensionality reduction of image embeddings (PCA)
- Gradient boosting regression (XGBoost)
- Epoch-wise validation and model comparison

2. Data Preprocessing

Before feature engineering, numerical and categorical variables were cleaned, transformed, and consolidated to improve interpretability and model robustness. This preprocessing step was essential to ensure consistency across the dataset and prepare it for feature derivation.

Numerical Variable:

Sqft_columns *:* It includes Sqft_living, Sqft_lot, Sqft_above, Sqft_living15, Sqft_lot15.

These columns exhibit strong right-skewness and contain extreme values that can disproportionately influence model learning. To mitigate this, a **logarithmic transformation** was applied, which compresses large values while preserving relative differences between properties. This transformation also converts multiplicative relationships into approximately additive ones, improving modelling stability:

$$x' = \log(1 + x)$$

After transformation, features were **standardized** to ensure comparable scales across numerical inputs:

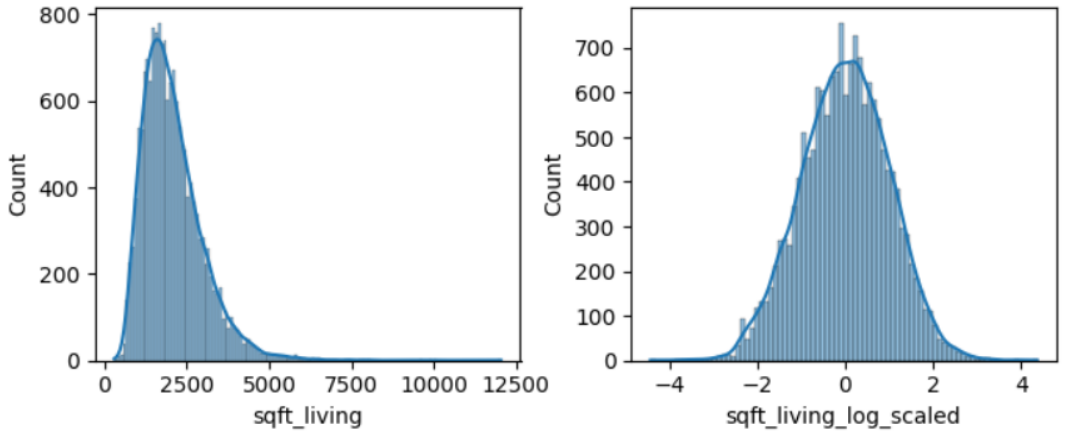
$$x_{scaled} = \frac{x' - \mu}{\sigma}$$

This combination reduces the impact of outliers, stabilizes optimization, and improves robustness when learning interactions between heterogeneous features.

```
log_scale_cols = [
    "sqft_living",
    "sqft_living15",
    "sqft_lot",
    "sqft_lot15",
    "sqft_above"
]

# log transform
log_cols = [f"{c}_log" for c in log_scale_cols]
for col in log_scale_cols:
    df1[f"{col}_log"] = np.log1p(df1[col])

# One scaler for all log columns
scaler_log = StandardScaler()
df1[[f"{c}_log_scaled" for c in log_scale_cols]] =
scaler_log.fit_transform(
    df1[log_cols]
)
```

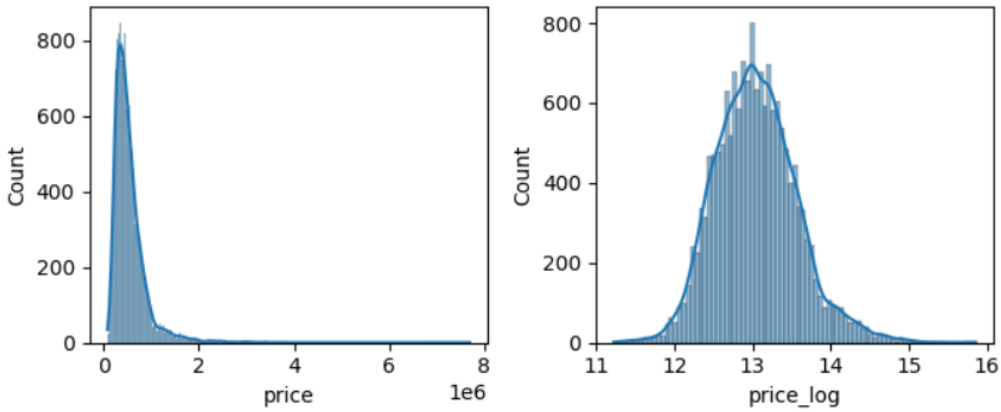



```
Initial skewness: 1.378761
Transformed column skewness: -0.048555
```

Sqft_basement *:* The sqft_basement feature was standardized without log transformation due to the high proportion of zero values representing properties without basements. Applying a log transform would distort the distribution and bias the feature. Standardization preserves the distinction between zero and non-zero values while ensuring comparability with other numerical.

Price *:* The target variable price was transformed to price_log using a logarithmic scale to reduce right-skewness and limit the influence of extreme values. Standardization was not applied to the target, as scaling the output variable does not affect model learning and would complicate interpretation of predictions. Modelling on log(price) preserves relative price differences while maintaining interpretability.

log(y)⇒reduces heteroscedasticity and minimizes outlier leverage




```
Initial skewness: 4.033062
Transformed column skewness: 0.411533
```

Categorical Variable: No scaling or transformation was applied, as these variables represent meaningful counts or binary indicators that are naturally interpretable by tree-based models.

Bedrooms: Converting anomalous entries (e.g., 33 bedrooms corrected to 3) and removing 5 records with 0 bedrooms, as such cases are not realistic for residential properties.

No. of bathrooms, floors which have decimal entries were not rounded off, because it represents realistic things in Real Estate.

3. Exploratory Data Analysis (EDA):

3.1 Price vs. Log-Scaled Living Area with Bathroom Count

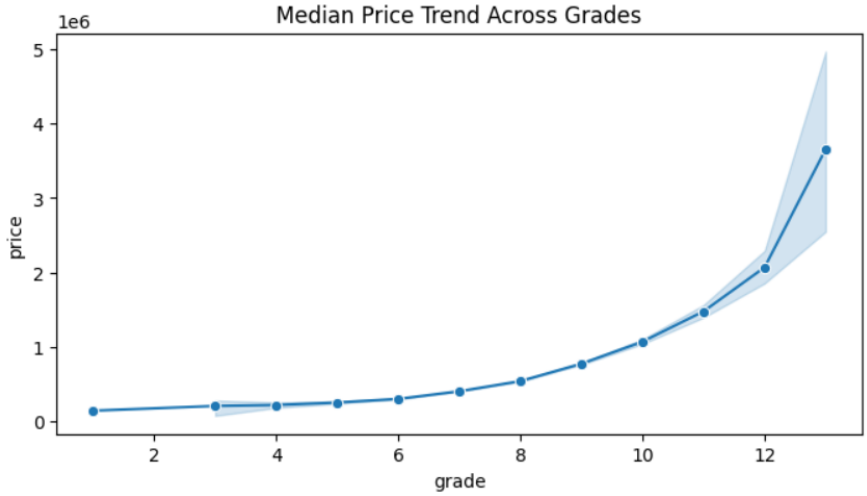
Insights:

- **Living area is the dominant price driver**, with larger homes commanding higher prices; however, **price dispersion widens at higher sizes**, showing greater sensitivity to secondary factors like location, grade, and amenities.
- **Bathrooms amplify the value of space**: for the same living area, homes with more bathrooms consistently fall into higher price bands, indicating a **non-additive interaction** between size and amenities.
- **Log-scaling living area linearizes the size–price relationship**, stabilizing variance and yielding a clear, interpretable trend well-suited for regression and tree-based models.
- **Mathematical Intuition:** Log-scaling the living area linearizes a power-law relationship

$$\text{Price} \propto (\text{Sqft})^{\beta} \Rightarrow \log(\text{Price}) = \alpha + \beta \log(\text{Sqft})$$

revealing a stable and interpretable trend suitable for regression and tree-based models.

3.2 Median Price trend across Grades



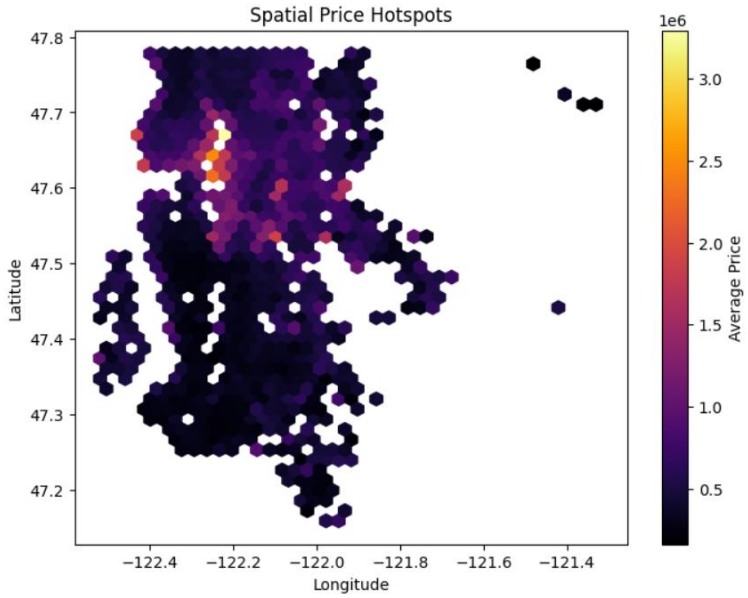
Insights:

- **Property price rises strongly with construction grade**, confirming grade as a key signal of build quality and market value.
- **The relationship is convex**: price gains are modest at lower grades but accelerate sharply beyond grade 9, with **high variability at premium grades** driven by luxury-specific factors.
- **Mid-grade homes (grades 4–8) show stable, predictable pricing**, while higher grades reflect non-linear buyer valuation of qualitative upgrades.
- **Mathematical intuition:**

$$\text{Price} \propto e^{\gamma \cdot \text{Grade}}$$

Even though the plot is in real price space, the convex shape suggests an underlying exponential relationship, which justifies modelling price in log-space during training.

3.3 Spatial Distribution



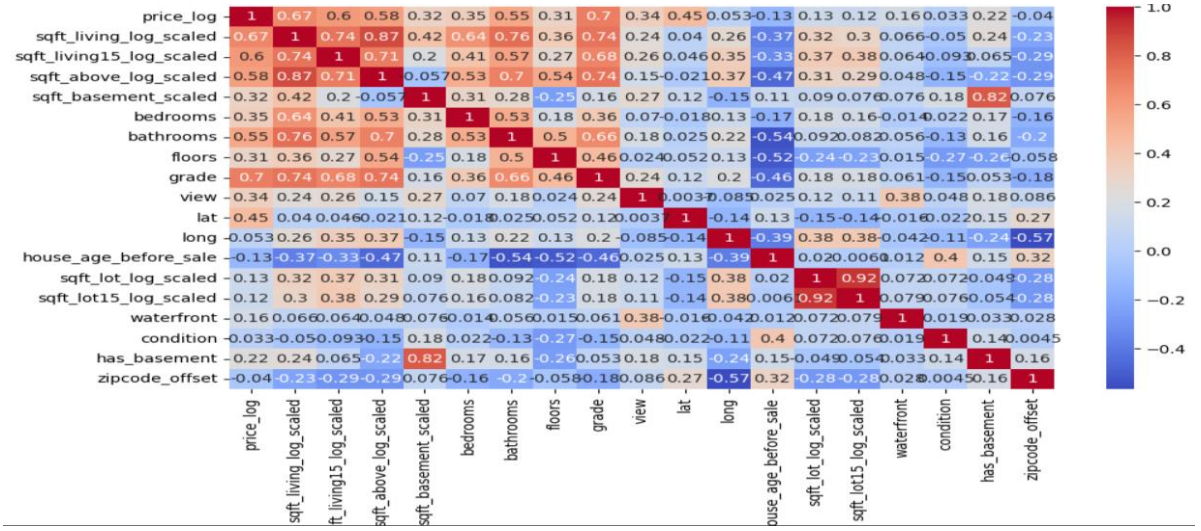
Insights:

- **Property prices exhibit strong geographic clustering**, with distinct high- and low-value zones driven by location.
- **Premium micro-markets show sharp local price gradients**, where nearby areas differ significantly due to amenities and accessibility.
- **Strong spatial heterogeneity means location can dominate property attributes**, justifying spatial and image-based features in pricing models.
- **Mathematical intuition:**

$$\text{Price} = f(\text{Property Attributes}) + g(\text{Latitude,Longitude})$$

where $g(\cdot)$captures non-linear spatial effects that cannot be modelled by tabular features alone.

3.4 Correlation Heatmap



Insights:

- **Strong size–price relationship:** price_log shows high correlation with sqft_living_log_scaled, sqft_above_log_scaled, confirming area as the primary price driver.
- **Quality and layout matter:** grade and bathrooms exhibit strong positive correlation with price, reflecting the impact of construction quality and functional design.
- **Age effect is negative:** house_age_before_sale is negatively correlated with both size and price, indicating depreciation effects over time.
- **Location proxies:** (lat, long, zipcode offset) show moderate but meaningful correlation with price, reinforcing the importance of spatial effects beyond property attributes.

4. Feature Engineering

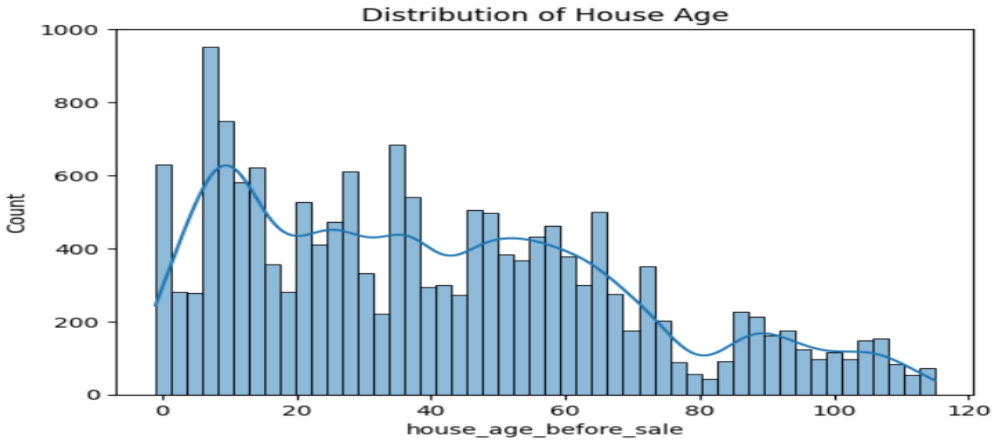
➢ CSV Based Feature Engineering

4.1 House_age_before_sale

Feature Engineering Justification

- The distribution of **house_age_before_sale** is right-skewed, with most properties concentrated in the **0–20 year** range.
- This indicates a market dominated by **newer constructions**, while very old houses are relatively rare.
- Multiple local peaks suggest **distinct construction waves** rather than continuous development over time.
- Older properties form a long tail, representing **niche segments** often driven by location or historical value.
- The clear non-linearity in age effects supports the use of **tree-based models**, which can capture diminishing and heterogeneous impacts on price.

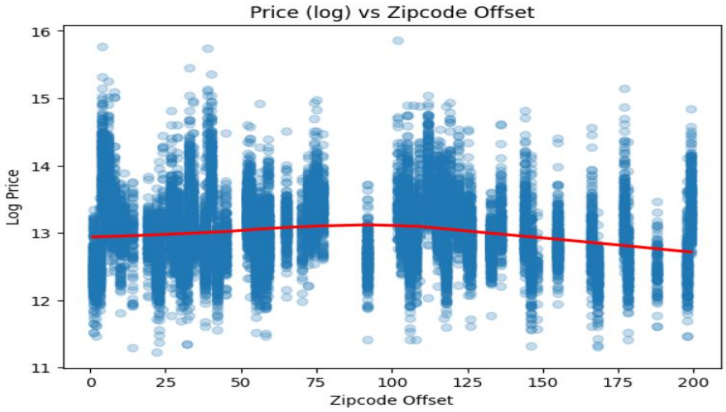
```
df1['house_age_before_sale'] = np.where(
    df1['yr_renovated'] > 0,
    df1['sale_date'].dt.year - df1['yr_renovated'],   # renovated → age from renovation
    df1['sale_date'].dt.year - df1['yr_built']        # not renovated → age from built
)
```



4.2 zipcode_offset (=zipcode – 98000)

Feature Engineering Justification

- **Spatial proxy:** zipcode_offset provides a compact numerical representation of location, capturing broad regional price patterns without high-dimensional encoding.
- **Model stability:** Compared to one-hot zipcodes, it reduces feature sparsity and overfitting, improving generalization in tree-based models.
- **Complementary signal:** It captures administrative and socio-economic zoning effects that latitude–longitude alone cannot fully explain.

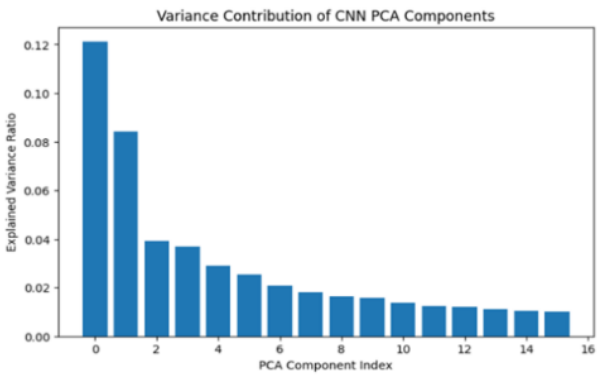


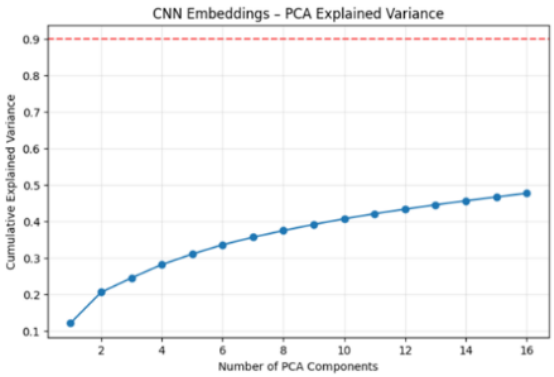
➢ **Image-Based Feature Engineering**

4.3 CNN Embeddings (Pretrained ResNet18)

Feature Engineering Justification

- Extracted high-level visual representations from satellite images using a pretrained ResNet18 with the classification layer removed.
- Embeddings capture complex spatial patterns such as neighbourhood layout, road connectivity, vegetation distribution, and built-up density that are not available in tabular data.
- Used as fixed features (no end-to-end training) to incorporate rich visual context while avoiding overfitting on limited labelled data.





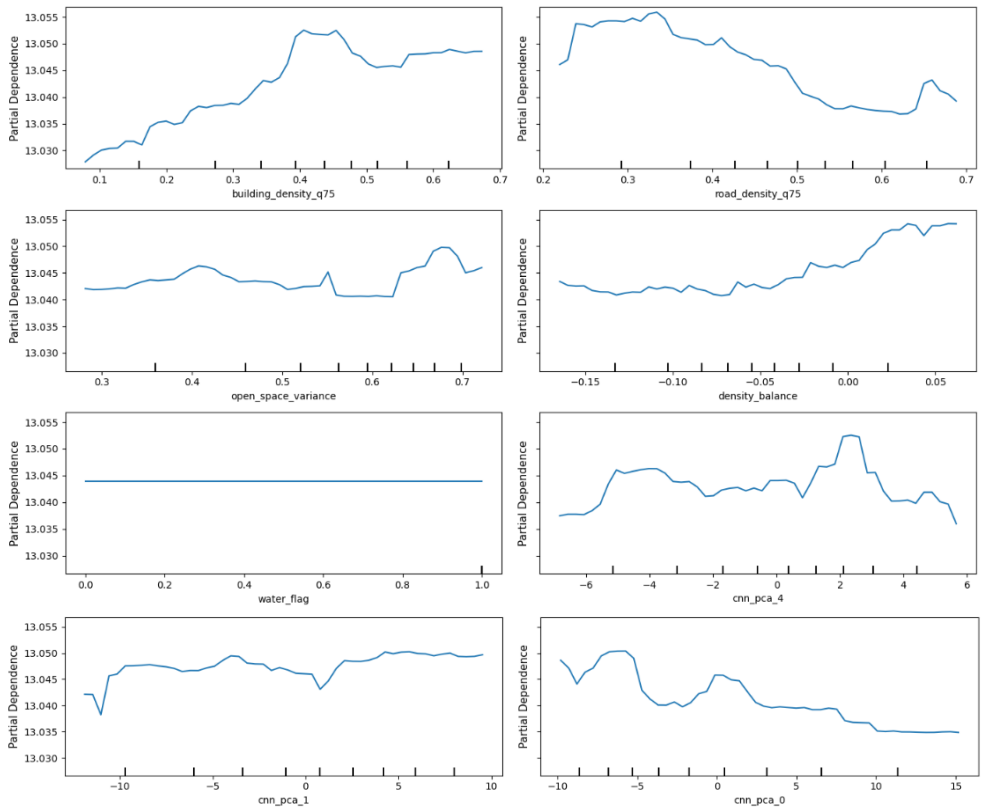
Insights:

- The first few PCA components capture a **disproportionately large share of visual variance**, indicating strong redundancy in raw 512-dim CNN embeddings.
- Variance accumulation shows **diminishing returns beyond ~12–16 components**, justifying dimensionality reduction without significant information loss.
- PCA stabilizes downstream learning by **removing noise and multicollinearity**, enabling effective fusion of image features with tabular data.

4.4 Structural Density Features

Feature Engineering Justification

- Engineered interpretable spatial proxies from satellite imagery, including building density, road density, open space ratio, and water presence.
- These features reflect urban congestion, accessibility, environmental quality, and proximity to amenities—key economic drivers of residential property prices.
- Designed to complement CNN embeddings with human-interpretable signals, improving model transparency and explainability.



Insights:

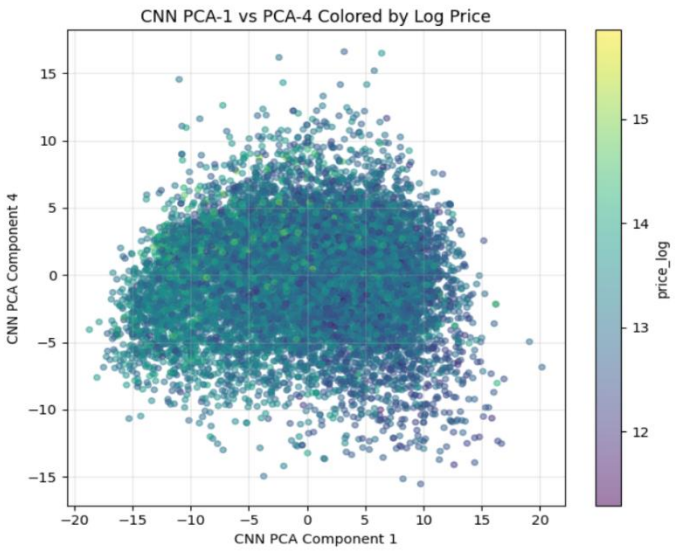
- **Structural density features (building, road, open space)** show clear non-linear effects, indicating that housing prices peak at a balanced trade-off between accessibility and congestion rather than at extreme densities.

- **Water presence exhibits a flat partial dependence**, suggesting its impact is sparse and highly location-specific, affecting only select premium neighbourhood instead of the overall market.
- **CNN PCA components display smooth, non-linear marginal effects**, confirming that deep visual embeddings capture meaningful neighbourhood context beyond simple linear relationships.
- **Image-based features act as refinement signals**, explaining residual price variation after strong tabular drivers (size, grade, location) and justifying their role in a multimodal modelling framework.

4.5 PCA-Reduced CNN Components

Feature Engineering Justification

- Reduced the original 512-dimensional CNN embeddings to 16 principal components using PCA.
- PCA preserves the dominant visual variance while removing redundancy and multicollinearity across embeddings.
- Dimensionality reduction stabilizes tree-based learning, reduces overfitting risk, and enables efficient fusion of image and tabular features.



Insights:

- **Diffuse but structured cloud:** The broad, elliptical spread indicates that PCA-1 and PCA-4 capture *independent, orthogonal visual patterns* (e.g., urban layout vs. neighbourhood texture) rather than redundant information.
- **Weak linear separation, strong local effects:** High-price properties are not isolated along a single direction; instead, price varies *locally* within regions of the PCA space, supporting the use of non-linear models (XGBoost) over linear regression.
- **Visual context as a secondary signal:** The absence of sharp clusters confirms that CNN features do not dominate pricing on their own but act as refinement signals, explaining residual variation after strong tabular drivers like size, grade, and location.

5. Financial / Visual Insights

5.1 Built Environment (Concrete vs. Space)

- **Building density (building_density_q75)** shows a non-linear impact on prices, with mid-density neighbourhoods outperforming highly congested areas.
- **Road density (road_density_q75)** improves value up to an accessibility threshold, after which congestion effects dominate.
- Buyers implicitly price an **optimal balance between connectivity and livability**, rather than raw urban intensity.

5.2 Neighbourhood Livability & Spatial Balance

- **Open space variance** captures spatial heterogeneity, with higher prices in areas combining built structures and open zones.
- **Density balance** highlights planned urban layouts over uniform sprawl or overcrowding.
- These features proxy neighbourhood quality that is not directly measurable through tabular attributes.

5.3 Water Presence & Environmental Premium

- **Water_flag** provides localized price uplift rather than a global effect across the dataset.
- The flat average impact reflects scarcity-driven valuation in specific sub-markets.
- This aligns with water acting as a **luxury amenity**, not a baseline pricing factor.

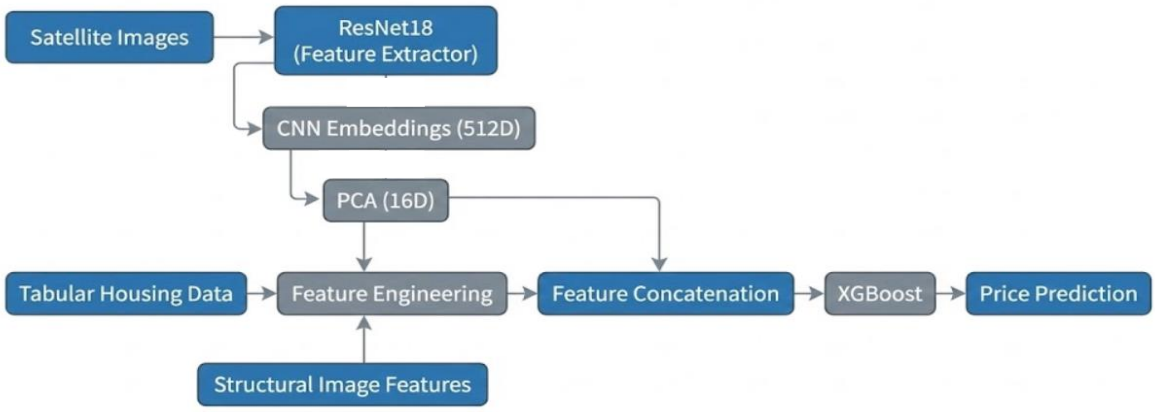
5.4 CNN PCA Visual Embeddings

- **CNN PCA components** encode high-level visual patterns such as urban texture, road geometry, and spatial organization.
- Their smooth, non-linear PDPs indicate meaningful but secondary influence on pricing.
- These features refine predictions among properties with similar size, grade, and location.

5.5 Economic Interpretation

- Tabular features establish **baseline valuation**, while visual features provide **incremental differentiation**.
- Image-derived signals explain residual price variation that standard housing attributes miss.
- This mirrors real-world buyer behaviour: structural filtering first, neighbourhood perception second.

6. Architecture Diagram



ResNet18 was used solely to generate Grad-CAM visualizations for exploratory understanding of satellite images. The Grad-CAM outputs had no role in feature engineering, model training, validation, or prediction, and ResNet18 itself does not contribute to the final modelling pipeline. Grad-CAM was employed purely as a qualitative visualization tool and has no influence on the model architecture, learned parameters, or outputs.

Executive Summary:

- The model combines **tabular housing attributes** with **satellite image–derived features** to improve property price prediction.
- Satellite images are processed using a **pretrained ResNet18** to extract high-level visual embeddings, which are compressed using **PCA** and augmented with **interpretable structural features** (density, open space, water presence).
- All features are fused and passed into an **XGBoost regression model**, where tabular data establishes baseline value and visual features refine predictions by capturing neighbourhood-level context.

A multimodal regression framework integrating structured housing data and satellite imagery to capture both property-level and neighbourhood-level drivers of value.

7. Results: Tabular Data Only vs. Tabular + Satellite Images

7.1 Model Comparison Setup

Two regression models were trained and evaluated using preprocessing steps, and evaluation metrics:

- **Tabular Data Only Model**:
  XGBoost regressor trained on structured housing attributes including size-related features, construction quality, age, and geographic information.

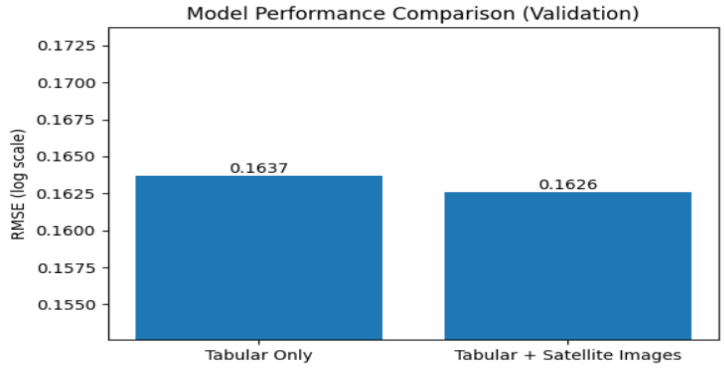
- **Tabular + Satellite Images Model**:
  An extension of the baseline model incorporating additional image-derived information in the form of:
    - PCA-reduced CNN embeddings extracted from satellite imagery
    - Engineered structural density features capturing neighbourhood characteristics

Model performance was evaluated on the validation set using **RMSE and R^2 on log-transformed prices**, ensuring stability against extreme price values.

7.2 Quantitative Performance Comparison

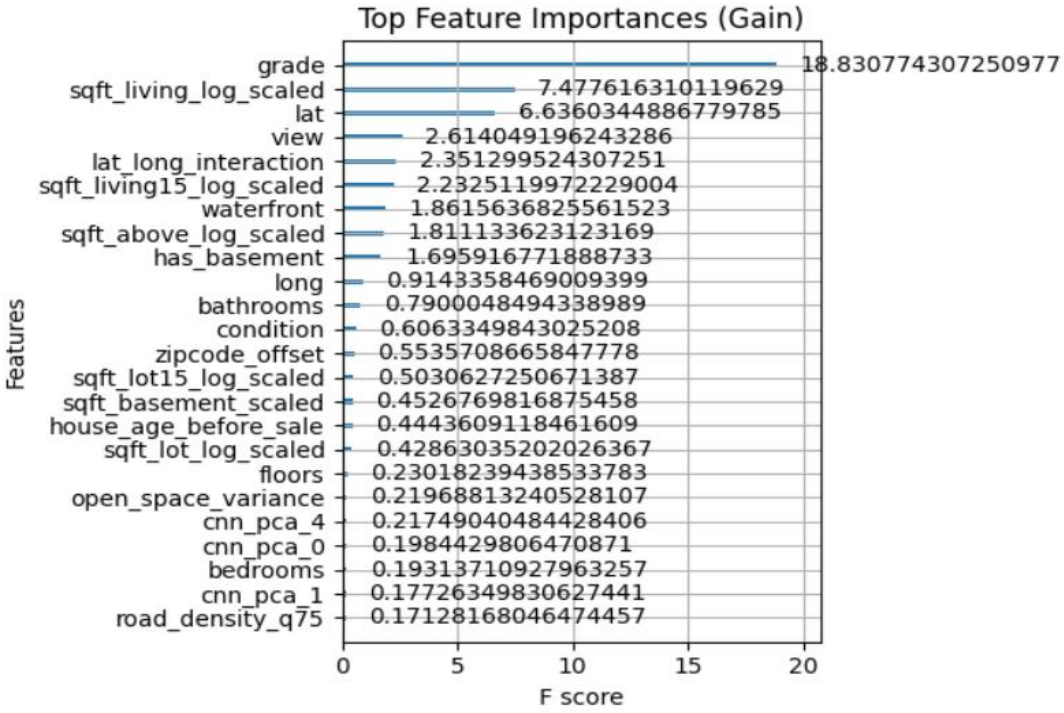
- The **tabular-only model** achieved a validation **RMSE(log) of 0.1637** and **R^2 (log) of 0.9022** of **Cross Validation**, reflecting the strong explanatory power of traditional housing attributes such as square footage, grade, and location.
- The **multimodal model**, which integrates satellite image features, achieved a slightly lower **RMSE(log) of 0.1626** and **R^2 (log) of 0.9042.**
- This represents a small but consistent improvement over the baseline, indicating that **satellite imagery provides complementary information beyond structured data.**

Although the numerical gain is modest, the direction of improvement is stable across training iterations, suggesting that visual features contribute incremental predictive value rather than introducing noise.



7.3 Interpretation of Results

- Tabular features remain the **dominant drivers** of property price prediction, capturing most of the variance through property-level fundamentals.
- Satellite images contribute **neighbourhood-level context**, such as patterns of built density, open space, and surrounding infrastructure, which are not fully captured by tabular variables.
- The multimodal model refines predictions by explaining **residual price variation** among properties with similar structural characteristics.



Tabular Data only Model (Cross Validation) :

```
===== Metrices =====
Mean RMSE (log): 0.16366777254092818
Mean R² (log): 0.9022369423423088
Mean RMSE (real): 117830.35888259669
Mean R² (real): 0.8928475508559404
```

Tabular + Satellite Images Model :

```
--- Best Epoch ---
Epoch: 3493
RMSE (log):  0.1626
R2   (log):  0.9042
RMSE (real): 115847
R2   (real): 0.8931
```

7.4 Key Takeaway

While structured housing attributes account for the majority of predictive power, incorporating satellite image–based features yields a modest but consistent improvement by capturing neighbourhood-level effects absent from tabular data.