SATELLITE IMAGERY BASED PROPERTY VALUATION

Overview

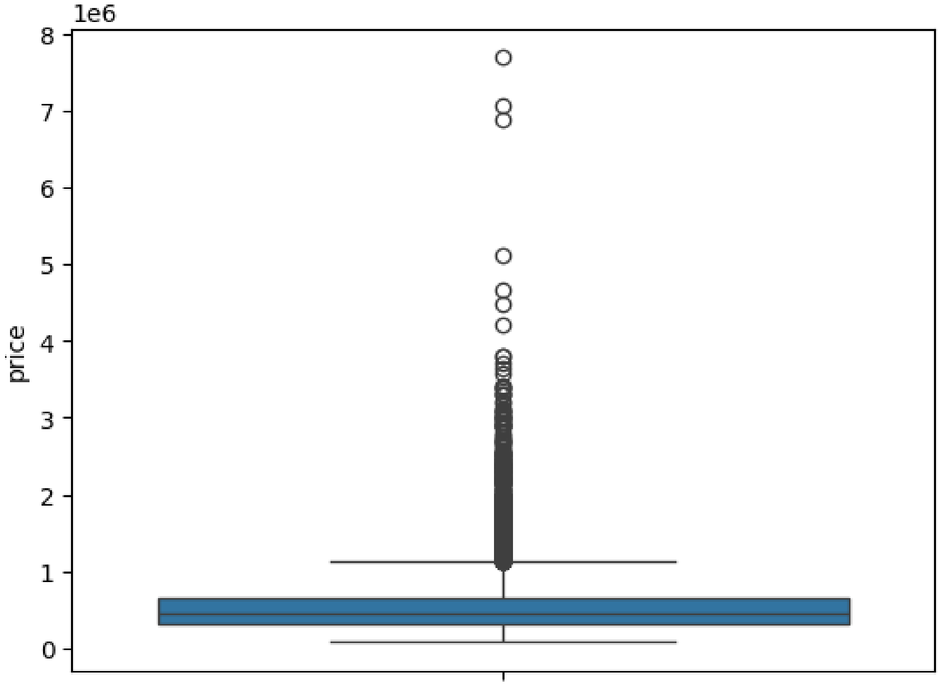
The model uses a combination of CNN architecture ResNet18 and XGBoost for tabular data. Satellite images from Sentinel-2 via Google Earth Engine are collected for each property location based on latitude and longitude given, and preprocessed (e.g. normalized RGB and vegetation indices like NDVI, NDBI, NDWI). Numeric features (square footage, year built, location, etc.) are cleaned, encoded (categorical → numeric) and scaled. For tabular data, XGBoost gives the best performance amongst Random Forest and Multi Layer Perceptron. For imagery, a pretrained ResNet CNN extracts visual features (shape of building, greenery, roads). These two feature vectors are concatenated and fed through fully-connected (ReLU + dropout) layers to predict the price.

In essence, the CNN captures “neighborhood context” (green cover, concrete, street patterns) while XGBoost captures intrinsic house characteristics.

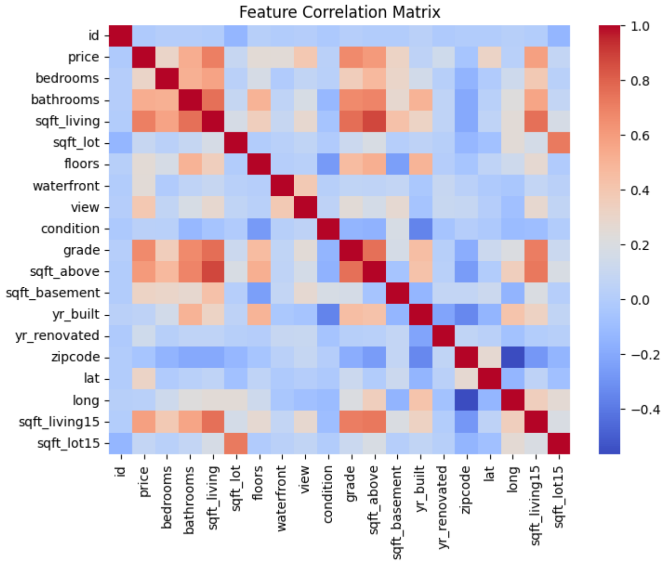
Exploratory Data Analysis

Tabular Data Analysis

It can be observed that houses lie mostly within a moderate price range with few high value outliers, representing the expensive properties. To reduce the effect of these values and improve numerical stability during training, the target variable was log-transformed. Thus, it is common in housing markets to use log(price) in modeling to normalize residuals.

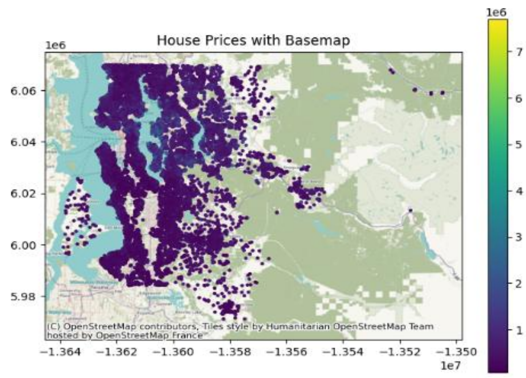


Feature correlation matrix shows price was dependent more on features like sqft living area and construction grade. Thus, due to such non-linear dependencies, XG boost was used to learn these relationships.

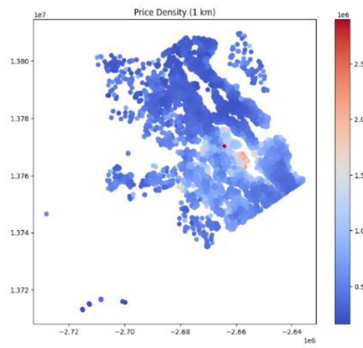


GEOPANDAS

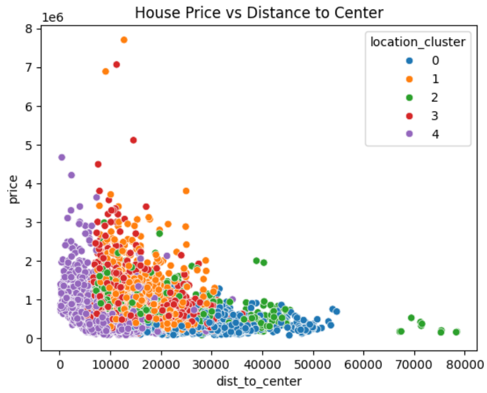
Geopandas was used to explore spatial data, plotting the point locations showed that the data was taken from house properties in seattle, usa.



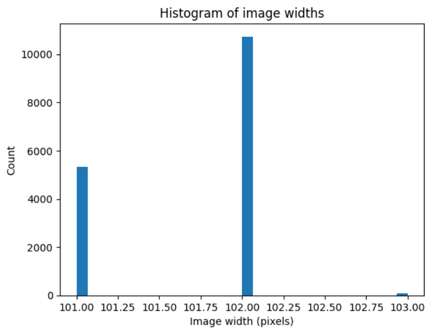
Spatial price density visualization showed mostly uniform pricing across the city, with a few exceptions of significantly higher prices. These high-value points reflect neighbourhood-level effects and spatial correlation in housing prices. This again justifies using log(prices) for model stability.

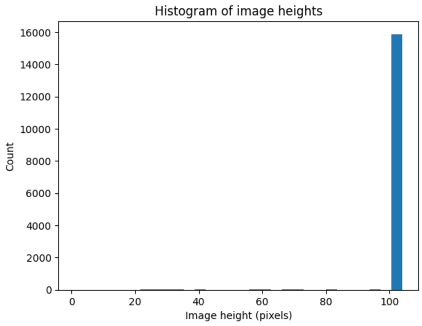


Location clusters vs Price
Location clusters were generated using K-means clustering which shows a strong correlation with house prices, with properties within same cluster having similar prices.



IMAGES
Satellite images were downloaded from sentinel hub sentinel-2 API with spatial resolution of 10m. It was observed that around 0.37% of images were missing due to required quality filters(cloud coverage) so dummy images were used instead of that and a binary 'has_image' attribute was added in the dataframe to turn on/off cnn-based feature projections.
Also, most images had height and width around 100 pixels, so to maintain consistency with ResNet architecture, these images were resized to 96x96 pixels.





Thus, images used were tensors of dimensions 96,3,3 and these were normalised using ImageNet mean and std deviation statistics.

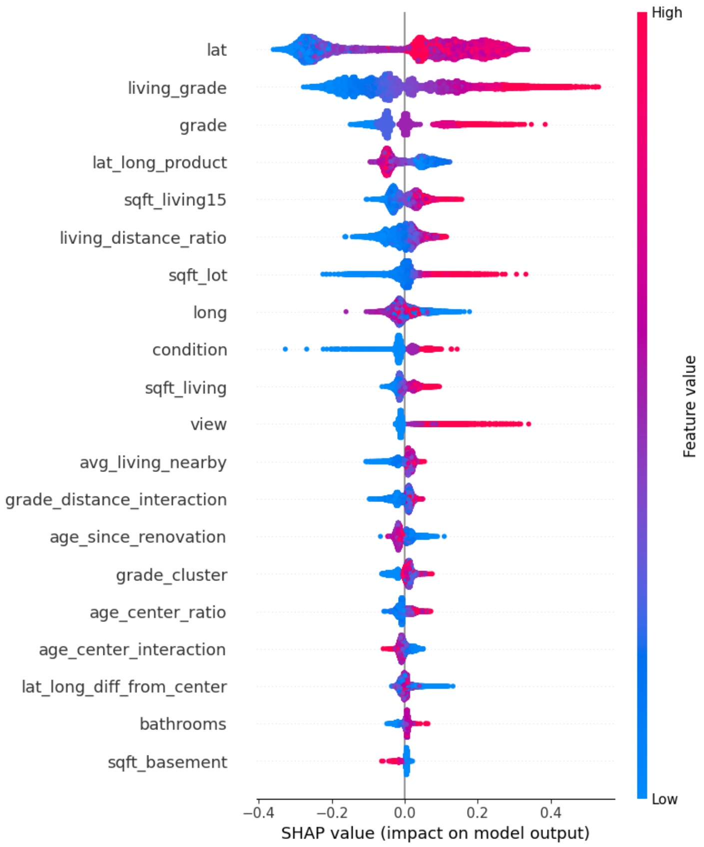
A typical img fetched looked like this:

Financial and Visual Insights

FOR TABULAR MODEL (XGBoost):

The SHAP summary plot provides a global interpretation of the XGBoost model by showing the importance of individual features.

Location-driven effects dominate the model.
Latitude (lat) is the most influential feature. It has a strong spatial dependency in house prices. Higher latitude values affect prices positively, reflecting the expensive properties associated with specific geographic zones within Seattle. This shows the presence of neighbourhood level price dependence and spatial correlation.



Quality and size are key financial drivers.
Features related to construction grade and living space like living_grade, grade, sqft_living, and sqft_living15, show a strong positive influence on price. Higher values of these features consistently show higher prices, highlighting that larger homes with better construction quality have significantly higher market value. This relation is quiet intuitive, however, the SHAP analysis empirically validates its importance within the model.

Engineered features and diminishing returns of absolute features.
Combination of features like lat_long_product, lat_long_diff_from_center, and living_distance_ratio show meaningful contributions, suggesting that relative positioning within the city (e.g., distance from central or high-value zones) plays an important role. These features help the model to understand nonlinear spatial pricing gradients that raw latitude and longitude cannot represent.

At the same time, certain absolute size-related features,eg. sqft_lot and sqft_basement, exhibit more heterogeneous SHAP distributions, suggesting diminishing marginal returns. While increases in these attributes can add value, their impact is not uniform and tends to weaken beyond a certain threshold. This can be explained using these engineered features which shows relation between these absolute size features and the grade and the location of the property.

Neighborhood context matters beyond the individual property.
Variables such as avg_living_nearby, grade_cluster, and grade_distance_interaction indicate that surrounding properties influence a home's valuation. As seen before, houses in location clusters having higher grade have more market value, reflecting the idea that neighborhood quality exerts a spillover effect on individual property prices.

Views and condition provide secondary but consistent uplift.
The view and condition features are less dominant than size or location, but they show a clear positive effect when present at higher values. This suggests that amenities such as scenic views and good maintenance condition contribute incremental but reliable gains in property value.

FOR CNN CONTRIBUTION IN THE FUSION MODEL

Grad-CAM Analysis

To interpret the role of satellite imagery in the proposed CNN + XGBoost fusion model, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the CNN branch. The Grad-CAM visualizations highlight spatial regions in the input Sentinel-2 images that contribute most strongly to the learned CNN embeddings used by the multimodal model.

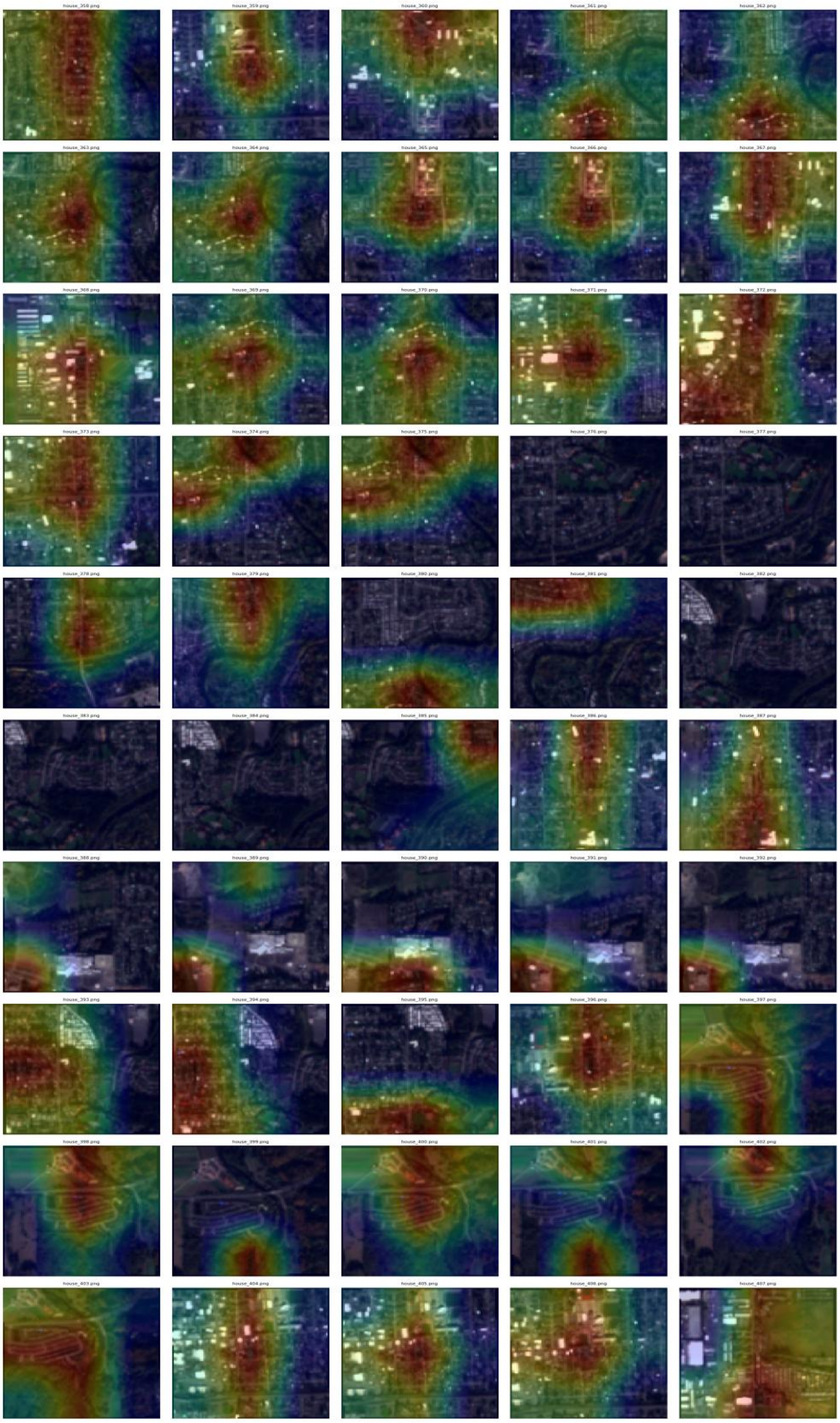
Visual Interpretation

Across a wide range of samples, the Grad-CAM maps consistently shows that large-scale urban structures have high-activation regions rather than the individual buildings. High-activation regions (shown in red and yellow having high price predictions) are predominantly aligned with:

- Major road networks and transportation connections

- Dense residential or commercial clusters
- High-intensity urban cores

In contrast, regions like open land, vegetation, or sparsely developed areas had low activation. This indicates that the CNN learned to focus on neighbourhood-level spatial data, such as accessibility and urban density, instead of fine-grained rooftop textures.



Financial and Economic Insights

From a real-estate valuation perspective, the highlighted regions correspond to intuitive of property prices, which are proximity to transportation, dense areas, and urban connectivity ,all associated with higher land and housing values. By capturing these spatial signals, the CNN provides complementary information that is not directly available in tabular property attributes.

Within the fusion framework, the CNN embeddings effectively encode location-driven price premiums, while XGBoost models intrinsic property characteristics such as size, age, and structural features.

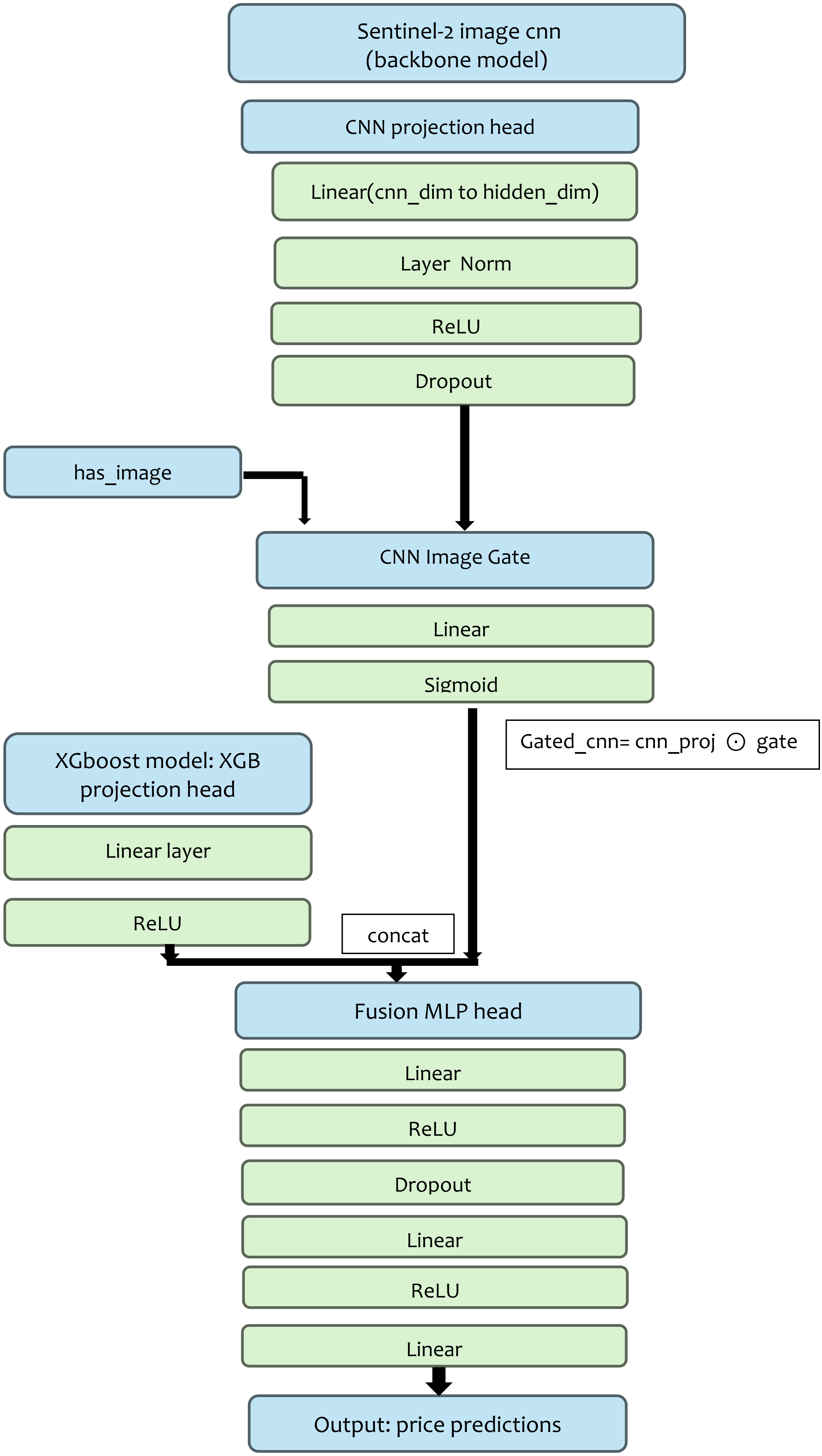
This enables the model to explain price variations between properties with similar tabular features but different surrounding urban features.

For example, two houses with same structural features may receive different price predictions depending on their proximity to major roads or dense urban zones, as reflected in the Grad-CAM activations.

Model Architecture

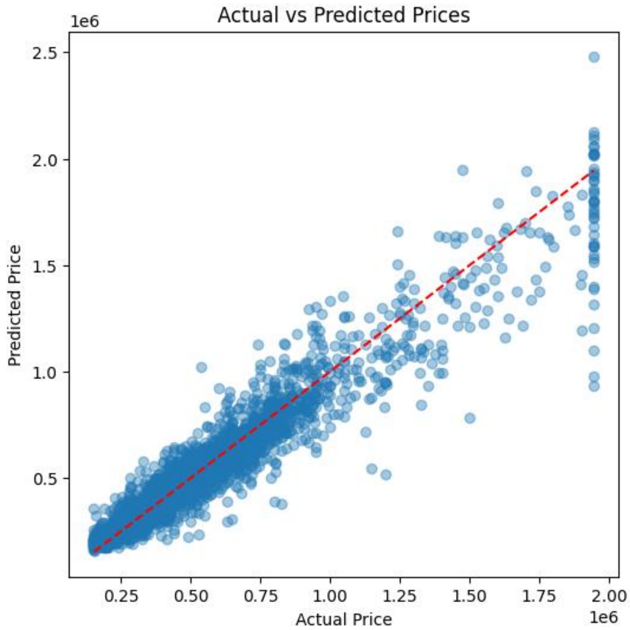
The network is a multimodal model, which takes the data in two forms, i.e. tabular features and images for each location. The left branch takes a satellite image into a CNN (ResNet), producing an image embedding. The right branch processes structured features (through XGBoost). These two embeddings are **concatenated** and passed through further fully connected layers with ReLU activation and dropout, yielding the final price output.

This architecture merges visual and tabular pipelines: the pretrained CNN's last layers (512-D) become a feature vector, which is combined with the tree-based features and fed to a regression head.

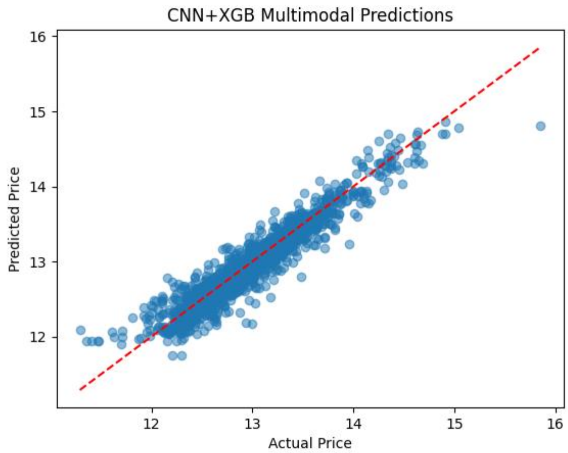

Sentinel-2 image cnn
(backbone model)
CNN projection head
Linear(cnn_dim to hidden_dim)
Layer Norm
ReLU
Dropout
has_image
CNN Image Gate
Linear
Sigmoid
Gated_cnn= cnn_proj ⊙ gate
XGboost model: XGB projection head
Linear layer
ReLU
concat
Fusion MLP head
Linear
ReLU
Dropout
Linear
ReLU
Linear
Output: price predictions

Results:

XGboost only model achieved an R2 score of around 0.89-0.91 on test data.



XGboost +ResNet model achieved an R2 score of around 0.9 for test data.



Although overall performance gains are not large, the fusion model provides more informative feature importance and consistently yields lower prediction errors in spatially heterogeneous regions.

Thus, the multimodal model is the preferred choice because of its robustness and improved feature interpretability.