Overview :

Problem Statement

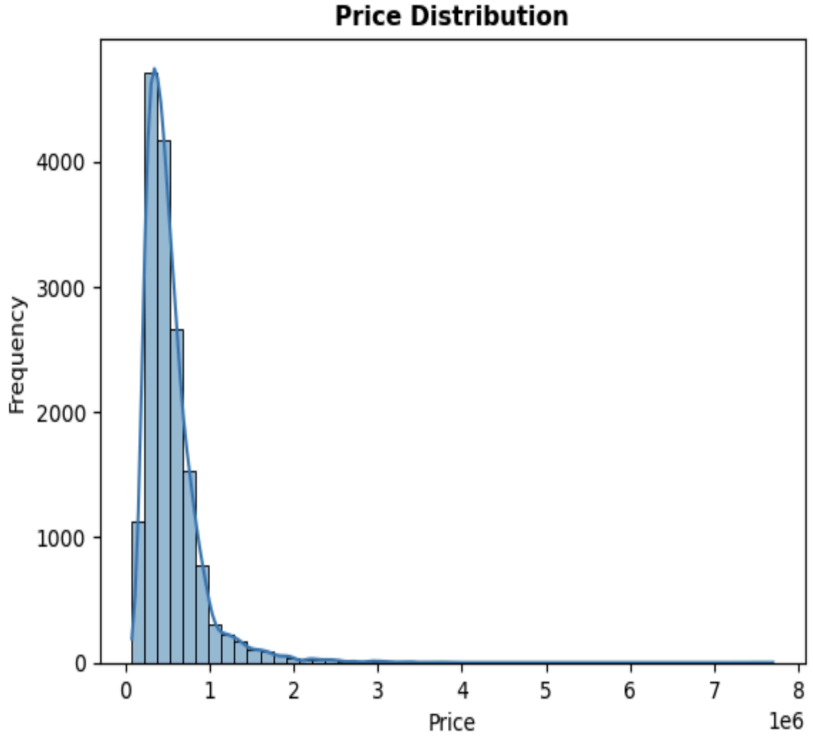
1. Predict house prices using structured housing attributes.
2. Hypothesis: satellite imagery contains neighbourhood-level signals not present in tabular data.

Approach

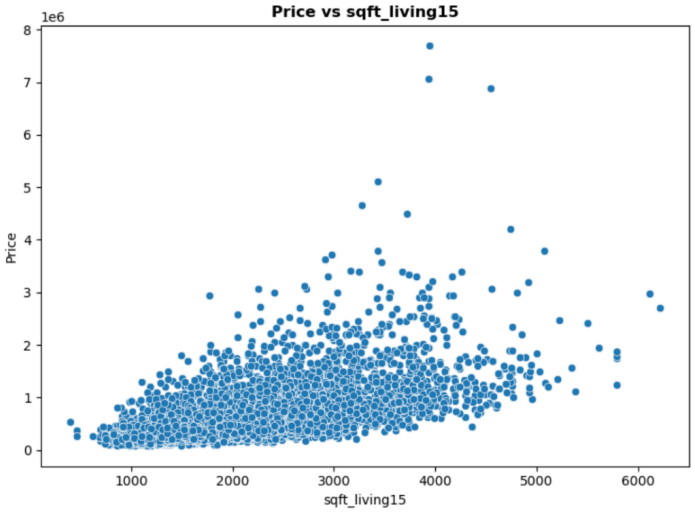
1. XGBoost on tabular features
2. CNN (ResNet-based) on satellite images
3. Fusion model combining tabular and image features

EDA :

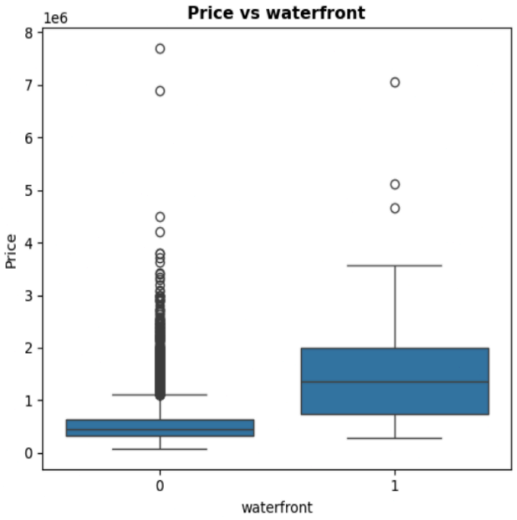
1. The price distribution being right skewed motivates us to use a tree based model , hence we use XGBoost.

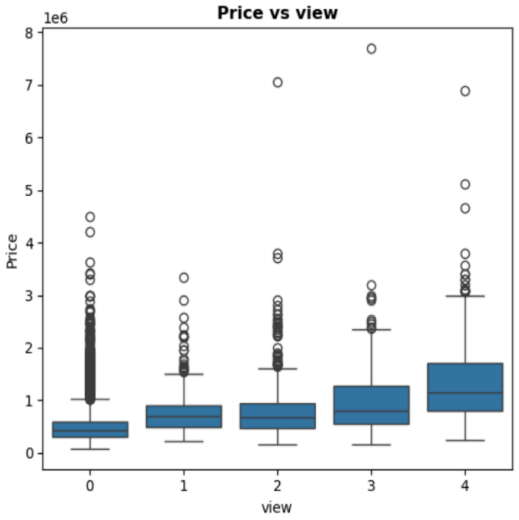


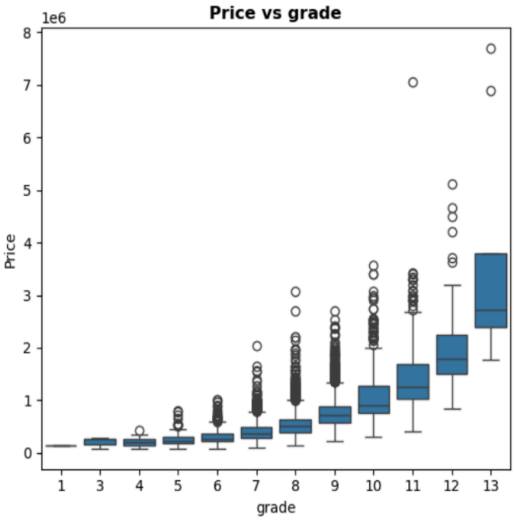
2. There is a strong positive relationship between sqft_living , sqft_living15 and price captured via the following scatterplot.



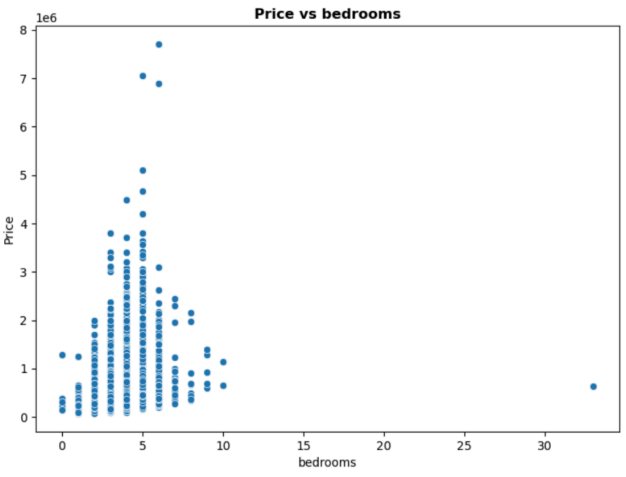
3. Grade largely affects the price of the house. Along with it , presence of waterfront and view quality also affects the price.







4. Bedrooms and Bathrooms being key features affect the price as well.



Visual Insights:

These are sample satellite images fetched via google maps static API.

Observations :-

1) Higher priced houses are surrounded by dense tree cover while lower priced houses show more concrete-dominated surroundings.
2) Expensive properties show more spacing between neighbouring structures while cheaper properties appear smaller in visible roof area.
3) Some lower priced houses are directly adjacent to main roads while higher priced ones are located on quieter internal streets.
4) Presence of waterfront and view quality positively correlates with the house price.

Architecture Diagram:

Satellite Images → CNN → Image Embeddings

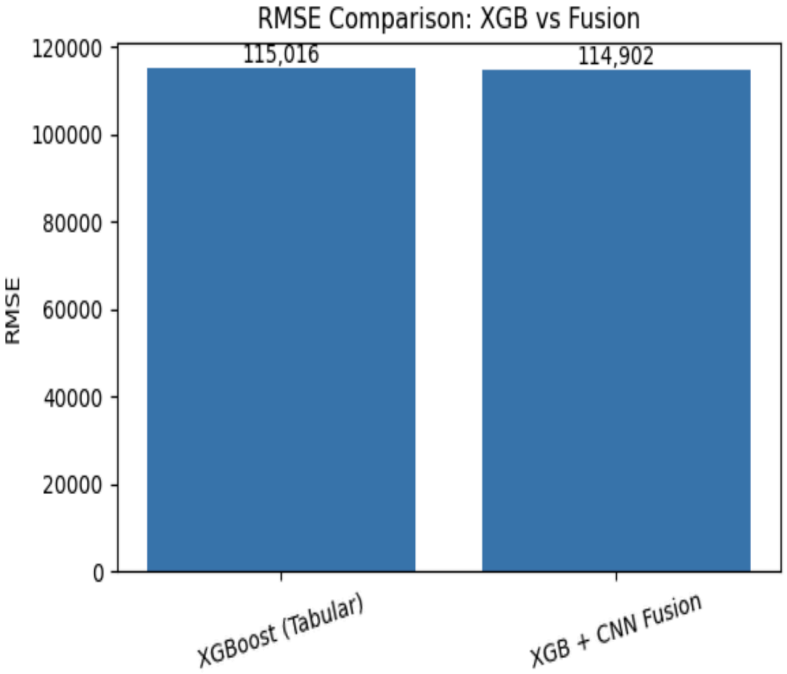
Tabular Features → MLP → Concatenation → MLP → Price

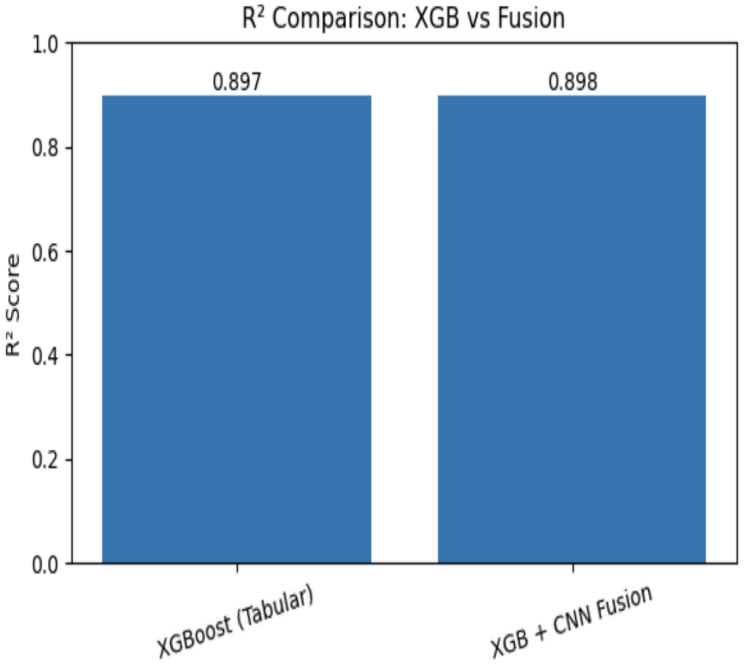
Satellite images are first processed using a Convolutional Neural Network (CNN), which automatically extracts high-level visual features such as neighborhood layout, green cover, road structure, and building density. The output of the CNN is a fixed-length image embedding that represents visual context around each property.

In parallel, structured tabular features (like bedrooms, bathrooms, area, year built, location-based attributes) are passed through a Multilayer Perceptron (MLP) to learn nonlinear interactions among numeric variables and produce a tabular embedding.

The image embedding and tabular embedding are then concatenated to form a joint representation that captures both visual and structured information. This combined feature vector is passed through a final MLP regression head, which learns cross-modal interactions and outputs the predicted house price.

Results:





The multimodal fusion model (Tabular + Satellite Images) shows a marginal improvement over the tabular-only XGBoost model in terms of RMSE and R^2. While the numerical gains are small, the performance remains consistent and stable, indicating that the image features provide complementary but not dominant information beyond structured attributes.

In real estate applications, even small RMSE improvements translate to significant monetary impact at scale, making the fusion approach practically relevant.