📄 PROJECT REPORT

Satellite Imagery-Based Property Valuation

Author: *Yogeshwar Singh*
Institution: *IIT Roorkee*
Year*: 3rd year*
Branch*: Mechanical Engineering*
Enrollment*: 23117149*
Date: *07 January, 2026*

---

1. Overview

Property price prediction usually relies on structured data such as location and basic property attributes. However, satellite images can provide extra context about the surrounding area, such as how urban or developed a region is.

In this project, I combine **tabular property data** with **satellite image features** to predict property prices. Satellite images are passed through a pretrained CNN to extract fixed image embeddings. These embeddings are then combined with tabular features and fed into an **XGBoost regression model**.

To improve performance and avoid manual tuning, **Bayesian optimization with cross-validation** is used to select XGBoost hyperparameters. Model performance is compared for **tabular-only** and **tabular + image** setups.

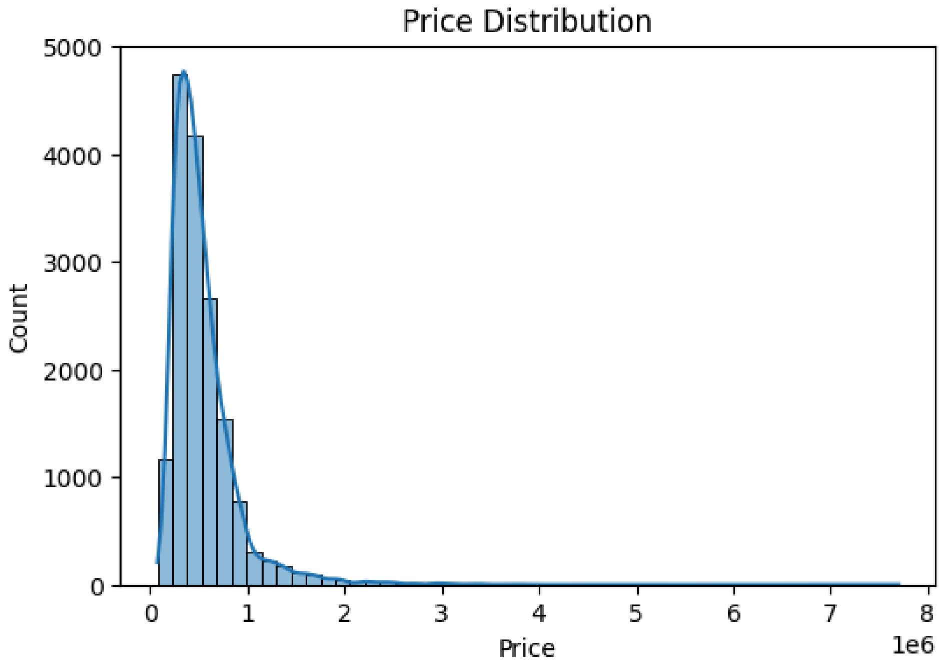
---

2. Exploratory Data Analysis (EDA)

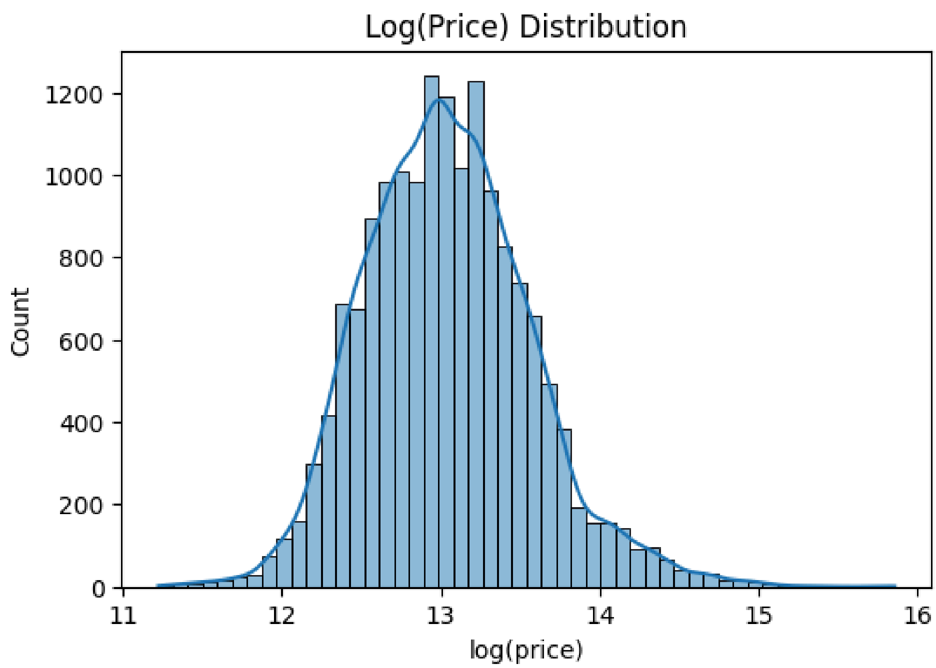
2.1 Price Distribution

The property price data is highly right-skewed, with a small number of very expensive properties. Training directly on raw prices leads to unstable predictions.

To handle this, the target variable is log-transformed before training, which results in a more balanced distribution and better model behavior, as prices are highly right-skewed



[Histogram of Raw Property Prices]



[Histogram of Log-Transformed Prices]

2.2 Sample Satellite Images

A few sample satellite images are shown below to understand the visual data used in the project.

[Sample Satellite Images with High v/s Low Price Labels]

High-priced properties often appear in dense urban regions, while lower-priced properties are more common in sparsely developed areas. This motivates the use of satellite imagery as an additional data source.

---

3. Financial and Visual Insights

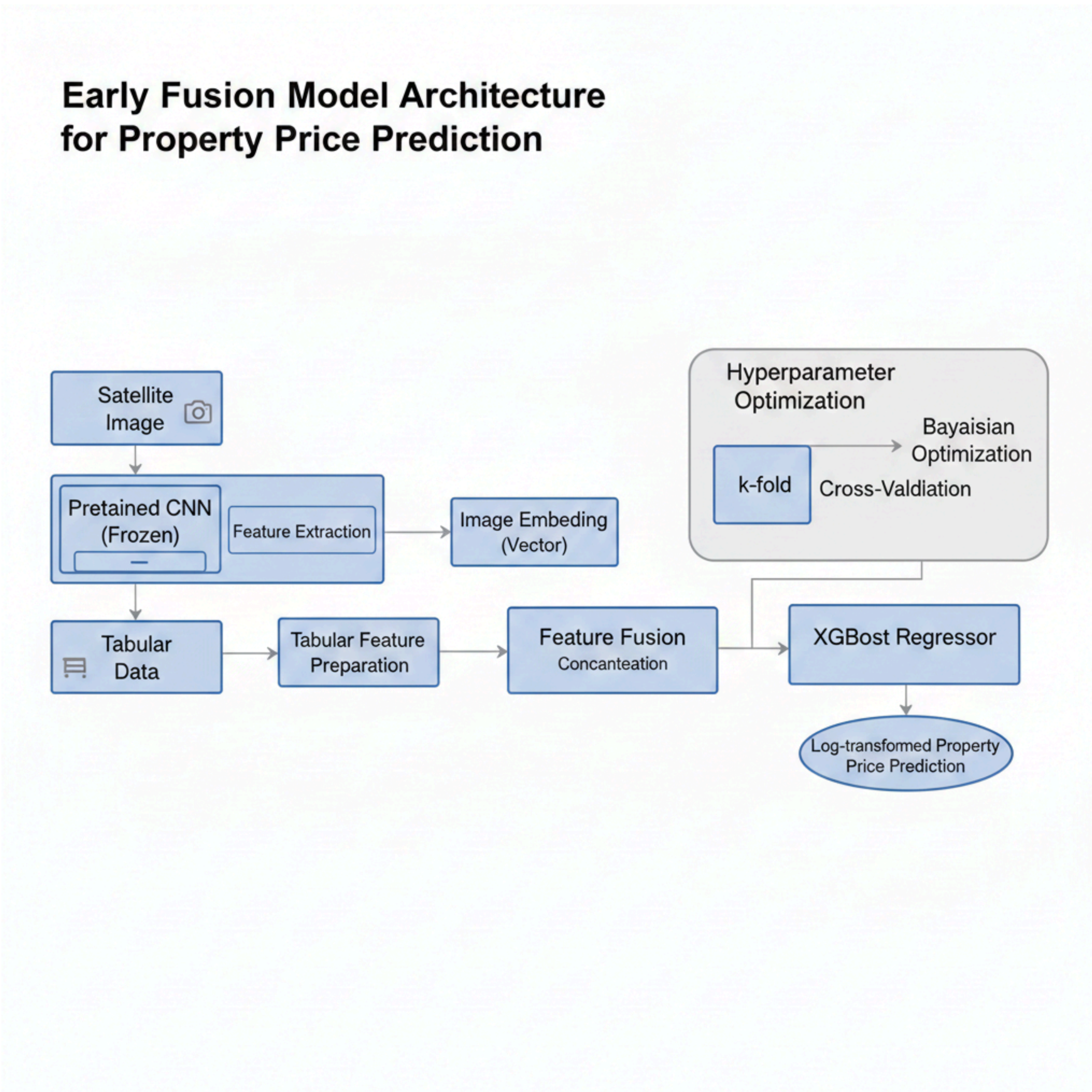
The image model produces numerical embeddings, which are not directly interpretable at the pixel level. As a result, the model does not explicitly identify features such as trees, buildings, or roads.

However, feature importance analysis shows that image embeddings contribute useful information when combined with tabular data. This indicates that satellite images capture high-level visual patterns related to property value, such as development density and surrounding infrastructure.

Visual inspection also suggests that developed areas are generally associated with higher prices, while open or less developed regions tend to have lower prices.

4. Model Architecture

The model uses an **early fusion** approach, where image features and tabular features are combined before prediction.



The modeling pipeline consists of the following stages:

1. **Satellite Image Processing:**
   Each satellite image is passed through a pretrained CNN (kept frozen during training) to

extract a fixed-dimensional embedding.

2. **Tabular Feature Processing:**
   Tabular features are cleaned and prepared without extensive normalization, as tree-based models are scale-invariant.

3. **Feature Fusion:**
   Image embeddings and tabular features are concatenated into a single feature vector.

4. **Regression Model:**
   An XGBoost regressor predicts the log-transformed property price.

5. **Hyperparameter Optimization:**
   Bayesian Optimization with k-fold cross-validation is used to tune key model hyperparameters.

---

5. Experimental Results

5.1 Model Configurations

Two primary model variants are evaluated:

- **Tabular Only:**
- **Tabular + Satellite Images**

Performance is evaluated on a held-out test set using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

5.2 Quantitative Results

[INSERT TABLE 1: Model Performance Comparison]

Model Variant	RMSE	R^2
Tabular Data Only	0.179	0.884
Tabular + Satellite Images	0.168	0.897

Including satellite image embeddings leads to **slight** improvement over the tabular-only model. Most of the predictive power comes from structured data, with images providing additional but limited gains.

---

6. Discussion

The results show that tabular features remain the main driver of property price prediction. Satellite images help capture broader spatial context but do not dominate performance when used as fixed embeddings.

Bayesian optimization helps automate hyperparameter tuning and produces performance comparable to carefully hand-tuned models. The small difference between optimized and manual configurations suggests the model is already close to its best achievable performance.

---

7. Conclusion

This project demonstrates a simple and effective way to combine tabular data and satellite imagery for property price prediction. Using pretrained image embeddings with an XGBoost model provides a practical balance between performance and complexity.

While image features offer modest improvements, the approach is scalable and can be extended in future work by training image models end-to-end or using more advanced fusion strategies.

---

8. References

1. **Yang, L., Chen, Z., & Li, G.** (2025).
   Satellite Image Price Prediction Based on Machine Learning.
   Remote Sensing, 17(12), 1960.
   DOI : https://doi.org/10.3390/rs17121960

2. **Chen, T., & Guestrin, C.** (2016).
   XGBoost: A Scalable Tree Boosting System.
   Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).