Satellite Imagery Based Property Valuation Project

RAJLAXMI_23112082

1. Introduction

Property valuation is a critical task in the real estate industry. Traditionally, property prices are estimated using manual surveys and numerical attributes such as area, number of rooms, and location. These methods are often time-consuming and subjective.

This project presents a machine learning-based approach to predict property prices by combining tabular property data with satellite imagery concepts. Satellite images provide information about surrounding infrastructure, green cover, road connectivity, and urban density, all of which influence property values.

2. Dataset Description

Two datasets were used: a training dataset and a test dataset. The training dataset contains over **16,209** property records with features such as bedrooms, bathrooms, living area, latitude, longitude, construction year, renovation details, and the target variable price. The test dataset contains approximately **5,404** records with identical features, excluding the price column.

3. Satellite Imagery Collection

Satellite imagery was incorporated by using latitude and longitude coordinates to fetch location-based static map images. A representative subset of locations was selected to demonstrate the imagery pipeline while keeping computational costs low. Although the images were not directly used in model training, this step validates the satellite-based workflow and supports future extensions of the project.

4. Data Preprocessing

Data preprocessing included **cleaning** the dataset, removing irrelevant columns, **handling inconsistencies**, and engineering additional features such as house age, renovation status, and total built-up area. Since the price distribution was skewed, a logarithmic transformation was applied to the target variable to improve model stability and learning performance.

5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to analyze feature distributions and relationships with the target variable. The analysis showed that property prices are right-skewed, which is common in real-world datasets. Log transformation resulted in a more balanced distribution, helping the model generalize better.

6. Model Development

A **Random Forest Regressor** was chosen for this task due to its ability to capture non-linear relationships and interactions between features. The dataset was split into training and validation sets, and appropriate feature scaling was applied. The model was trained with tuned hyperparameters to achieve reliable performance.

7. Model Evaluation

Model performance was evaluated using R-squared (**0.86**) and Root Mean Squared Error (**129260**). The model achieved a strong R^2 score, indicating effective learning of data patterns. The RMSE reflects reasonable prediction accuracy considering the wide range of property prices.

8. Final Prediction

The trained model was applied to the test dataset to generate final price predictions. Predictions were converted back from log scale to the original price scale. The final output was saved in the required CSV format with columns id and predicted_price.

9. Conclusion

This project demonstrates an end-to-end machine learning pipeline for property price prediction using tabular data and satellite imagery concepts. The approach successfully integrates data preprocessing, feature engineering, model training, evaluation, and prediction. Future work may include direct use of satellite images with deep learning models and additional spatial features to further enhance prediction accuracy.