Property Price Prediction Using Tabular Data and Satellite Image Residual Learning

- Suhani Jain (23411039)

1. Overview: Approach and Modeling Strategy

The goal of this project is to improve property price prediction by combining **tabular property data** with **satellite imagery**, using a **residual learning framework**.

Instead of training a single end-to-end multimodal model, the approach is designed in **three clear stages**:

Stage 1: Tabular Baseline Model

- A LightGBM model is trained using only tabular features.
- This model captures structured factors such as location coordinates and property attributes.
- It provides an initial price prediction (`lgb_pred`).

Stage 2: Hard-Sample Selection Using Residuals

- Residuals are computed as:

```
residual=  true price- LightGBM prediction
```

- Only **hard samples** (large prediction errors) are selected for image-based learning.
- The threshold is set at the **69th percentile** of absolute residuals.

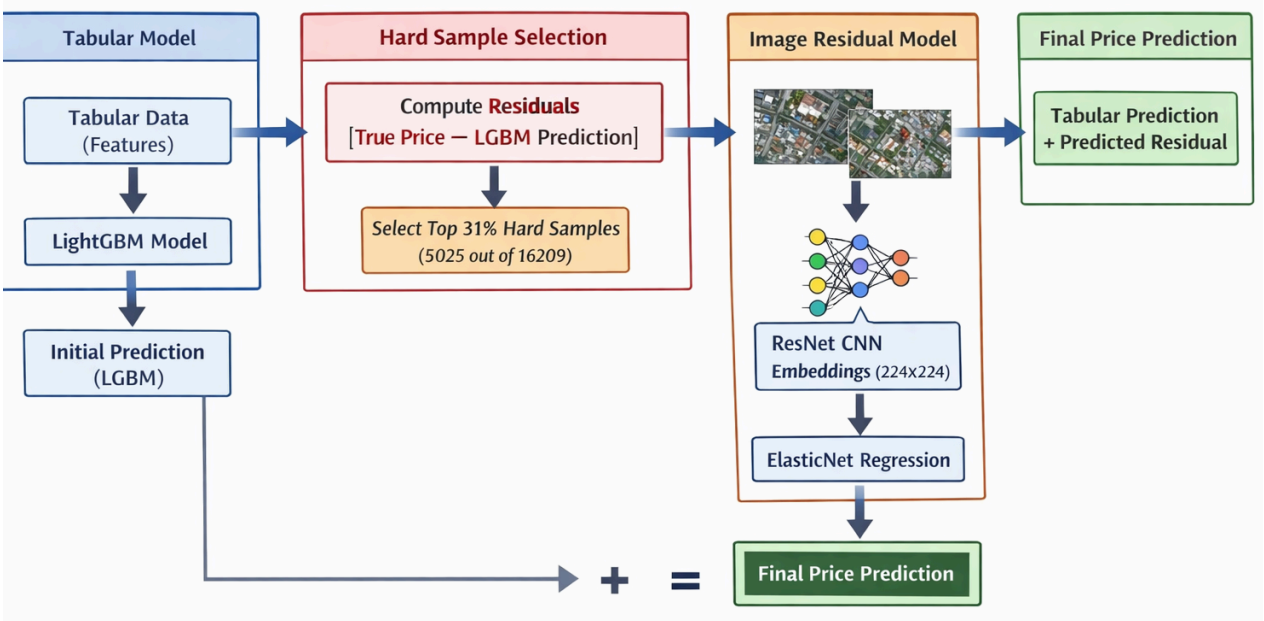
```
Using 5025 hard samples out of 16209
```

This step ensures that satellite images are used **only where tabular data fails**, making the model efficient and focused.

Stage 3: Image-Based Residual Prediction

- Satellite images are processed using a **pretrained ResNet CNN**.
- The CNN is used only as a **feature extractor** (no fine-tuning).
- Extracted image embeddings are fed into an **ElasticNet regression model** to predict residual corrections.
- Final price is computed as:

```
Final Prediction = Tabular Prediction + Predicted Residual
```



2. Exploratory Data Analysis (EDA)

This section explores the key characteristics of the dataset used for property price prediction. The analysis focuses on the **distribution of property prices** and a **visual inspection of satellite images** to understand the additional information they provide beyond tabular features.

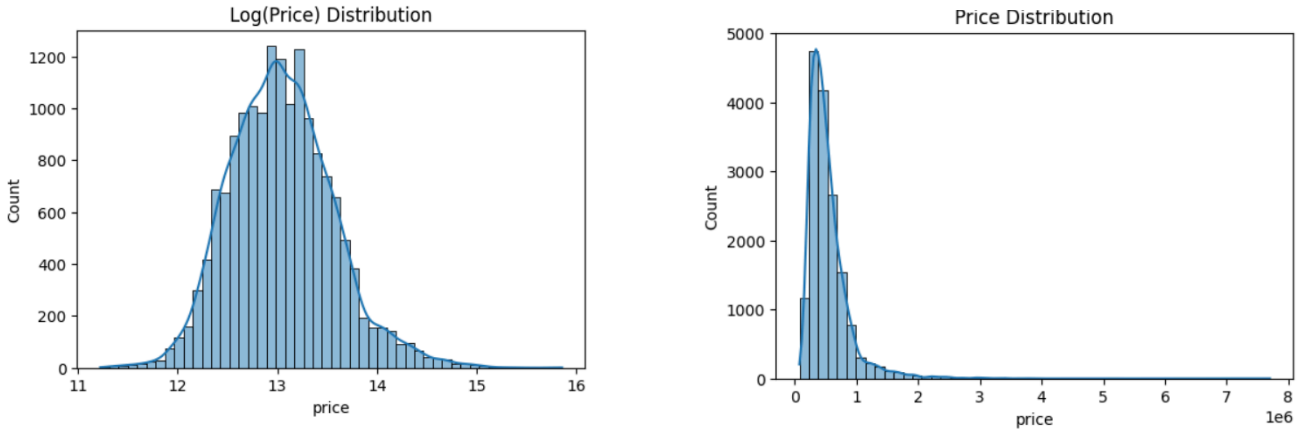
2.1 Property Price Distribution

The target variable in this project is the **property price**.
A visualization of the price distribution shows that the data is **right-skewed**:

- Most properties fall within the **lower to mid price range**
- A smaller number of properties have **very high prices**
- Some extreme values (outliers) are present, which is common in real estate data

This distribution suggests that:

- Predicting property prices is not trivial due to large variance
- Models must handle both common and high-value properties effectively
- Error metrics such as **RMSE and R^2** are appropriate for evaluation



Histogram plot of property prices

2.2 Engineered Tabular Features

In addition to raw tabular attributes, several **derived features** were engineered to better represent property characteristics and location effects. These features aim to capture non-linear relationships that are difficult for models to learn from raw inputs alone.

Examples of engineered features include:

- Location-based aggregations
- Distance-related or density-based measures
- Transformed versions of original features

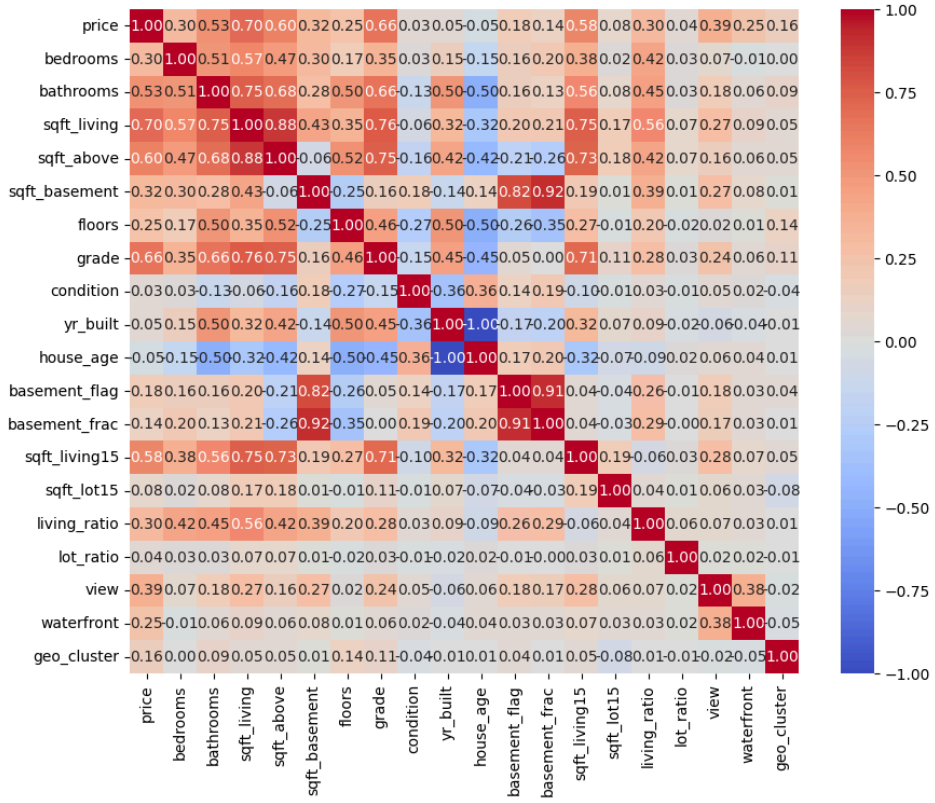
These engineered features help the tabular model capture **economic and spatial patterns** more effectively.

2.3 Correlation Analysis of Tabular Features

To understand relationships between engineered features and property price, a **correlation matrix** was computed.

The correlation analysis reveals:

- Certain engineered features show **strong positive correlation** with price
- High correlations between some features suggest redundancy, which is handled well by tree-based models



2.4 Tabular Feature Limitations

The tabular dataset contains structured information such as location-based and property-specific attributes. While this data captures important economic signals, it **does not describe the physical surroundings** of a property.

For example, tabular data cannot directly represent:

- Amount of greenery near a property
- Road layout and connectivity
- Urban density or congestion
- Neighborhood planning quality

This motivates the inclusion of **satellite imagery** as an additional data source.

2.5 Satellite Image Data Overview

Each property is associated with a satellite image representing its surrounding area.
These images provide a **top-down view** of the neighborhood and capture spatial patterns that are difficult to encode numerically.

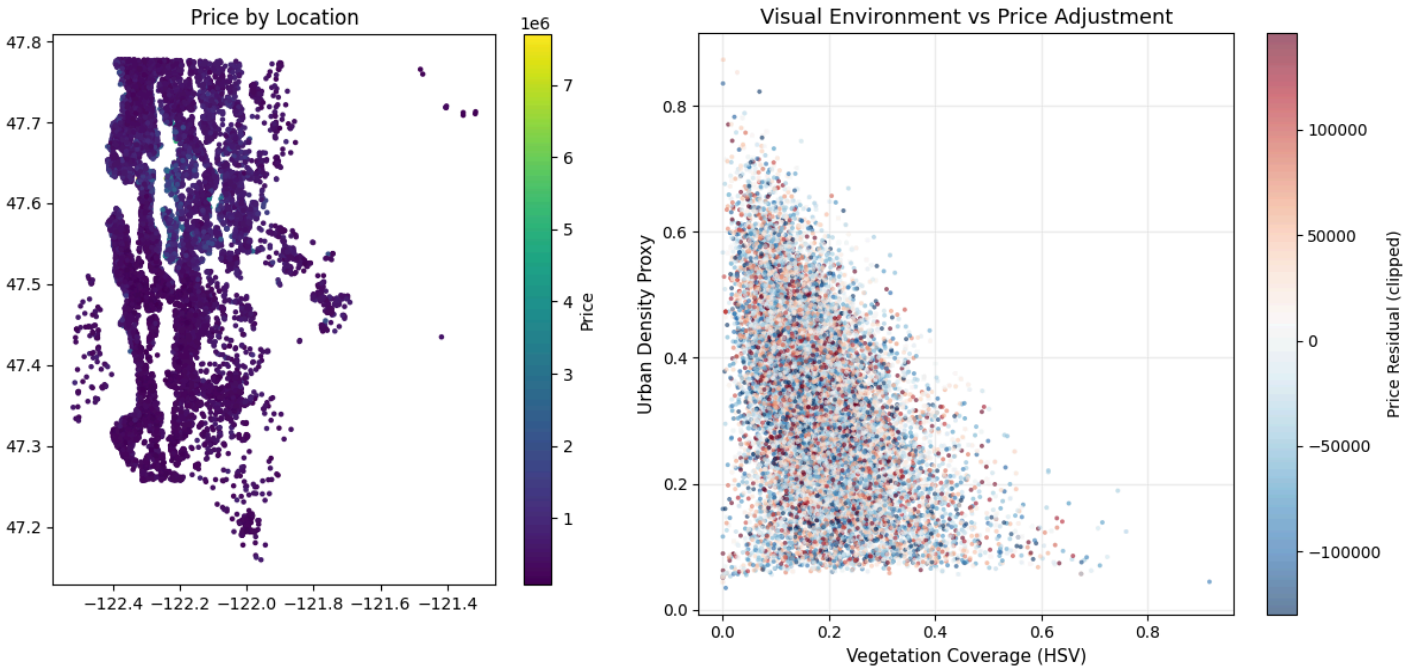
2.6 Visual Environment Features vs Price Adjustment

To better understand how visual surroundings influence property prices, a plot was created relating **visual environment indicators** (such as green cover and urban density) to **price adjustments made by the image-based model**.

The analysis shows that:

- Properties with **higher green cover** tend to receive positive price adjustments
- Areas with **high urban density and concrete dominance** often receive negative adjustments
- Visual features provide complementary information to tabular data

This confirms that satellite imagery captures economically meaningful signals.



2.7 Summary of EDA Findings

The analysis shows that engineered tabular features explain most of the variance in property prices. However, satellite images capture additional environmental factors such as greenery and urban density that are not present in tabular data, helping improve predictions in visually distinct locations.

3. Financial and Visual Insights

This section analyzes how **visual characteristics extracted from satellite images** influence property value and explain price adjustments beyond tabular data. These insights are derived from qualitative inspection of satellite images and the residual corrections produced by the image-based model.

3.1 Role of Visual Environment in Property Valuation

While tabular data explains most of the variation in property prices, it does not capture the **physical and environmental context** of a location. Satellite images provide a direct view of the surrounding environment, allowing the model to account for factors that influence buyer perception and long-term value.

The image-based model primarily acts as a **price adjustment mechanism**, refining predictions made by the tabular model.

3.2 Visual Features Associated with Higher Property Value

Properties that receive **positive price adjustments** from the image-based model typically exhibit the following visual characteristics:

- **High green cover**, including trees, parks, and open spaces
- **Low building congestion**, with visible spacing between structures
- **Organized road layouts**, indicating planned residential areas
- **Clean neighborhood structure**, with minimal industrial activity

From a financial perspective, these features are associated with:

- Better living conditions
- Higher demand and desirability
- Long-term appreciation potential

3.3 Visual Features Associated with Lower Property Value

Negative price adjustments are commonly observed in properties located in areas with:

- **High concrete density** and limited greenery
- **Dense and irregular building layouts**
- **Congested road networks**
- **Industrial or mixed land-use regions**

These visual signals often correspond to:

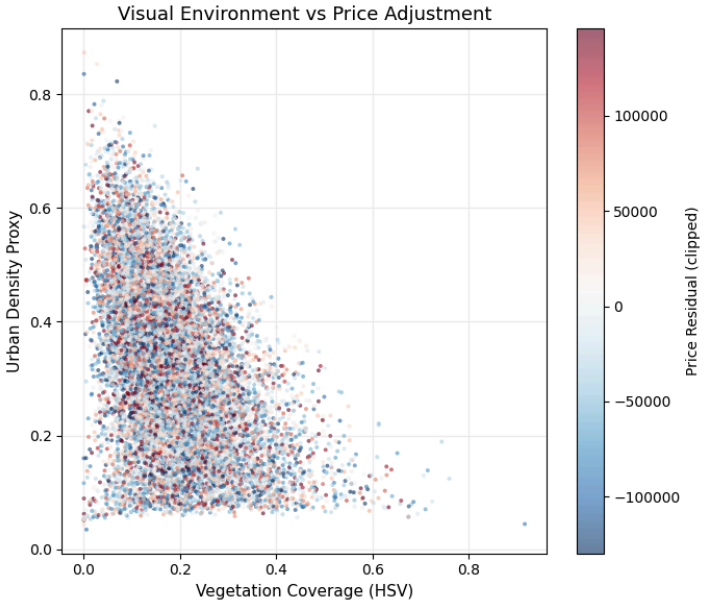
- Reduced environmental quality
- Higher noise and congestion
- Lower perceived livability

3.4 Green Cover and Urban Density vs Price Adjustment

A direct comparison between **visual environment indicators** and **price adjustments** highlights a clear trend:

- Properties with **higher green cover** tend to receive positive residual corrections
- Properties in **highly dense urban regions** often receive negative corrections

This confirms that satellite imagery captures economically meaningful information that complements tabular features.



(Both positive and negative price adjustments are observed across visual environments. Blue points indicate cases where satellite imagery suggests a lower valuation than what tabular features alone would predict. This commonly occurs in dense urban areas, regions with limited accessibility, or visually less desirable neighborhoods that are not fully captured by structured data. The presence of both positive and negative adjustments highlights that the image model acts as a corrective signal rather than a one-sided booster.)

3.5 Key Takeaways

- Visual features significantly influence property valuation
- Greenery and planned layouts or concrete-heavy and congested environments have an effect on the perceived and predicted value
- Satellite images improve model accuracy by capturing these effects

4. Architecture Diagram

The architecture consists of two main components:

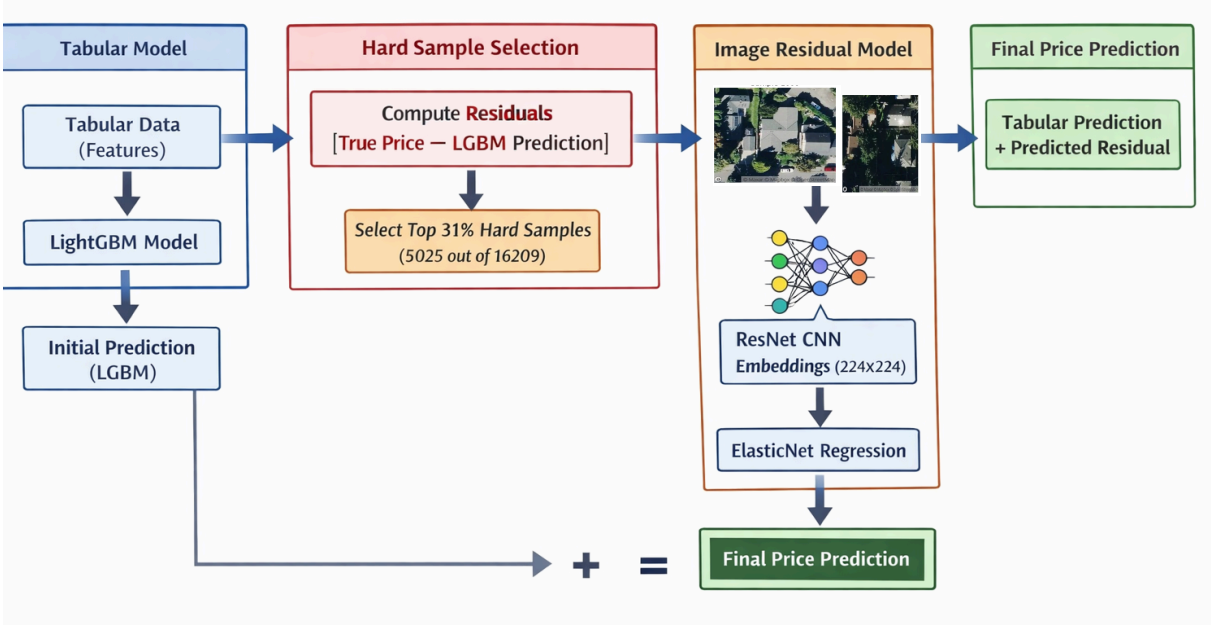
1. **Tabular Data Model**
    - Tabular features are used as input to a LightGBM model.
    - This model produces an initial property price prediction.

2. **Image-Based Residual Model**
    - Satellite images are processed using a pretrained Convolutional Neural Network (CNN).
    - The CNN extracts high-level visual features (embeddings).
    - These embeddings are used to predict residual corrections to the tabular model output.

The final price prediction is obtained by adding the predicted residual to the tabular model's output. The complete workflow is as follows:

1. Tabular features → LightGBM → Initial price prediction
2. Residuals computed from tabular predictions
3. Satellite images → CNN → image embeddings
4. Image embeddings → regression model → predicted residual
5. Final prediction = tabular prediction + predicted residual

This modular design allows each model to focus on what it does best:
tabular data captures structured economic factors, while images capture spatial and environmental context.



5. Results and Model Comparison

This section compares the performance of the **tabular-only model** with the **combined tabular and satellite image model** using standard regression metrics.

Model	RMSE ↓	R^2 Score ↑
Tabular Data Only (LightGBM)	~133,127	~0.855
Tabular + Satellite Images (Residual Learning)	106,083	0.8948

5.1 Interpretation of Results

- The tabular model explains a large portion of price variance, demonstrating strong baseline performance.
- Adding satellite images improves the **R^2 score**, indicating better overall explanatory power.
- The image-based model corrects errors made by the tabular model by incorporating visual and environmental information.
- The combined model is especially effective in visually diverse neighborhoods where tabular features alone are insufficient.

5.2 Key Outcome

The results confirm that:

- **Tabular data explains most of the variance in property prices.**
- **Satellite images provide complementary information** that improves predictions.
- The residual learning approach allows efficient and interpretable multimodal modeling.