Multimodal Real Estate Valuation using Satellite Imagery

1. Abstract:

This project develops a multimodal deep learning system for house price prediction. By integrating traditional tabular data (house features) with high-resolution satellite imagery, the model captures environmental context that simple spreadsheets miss. The system uses a Late Fusion architecture, combining a Convolutional Neural Network (CNN) and a Multi-Layer Perceptron (MLP), achieving a final validation loss of 0.0686.

2. Introduction & Problem Statement:

Traditional real estate valuation models rely solely on structured data such as square footage, number of bedrooms, and year built. However, property value is heavily influenced by "location quality"—something difficult to quantify in a spreadsheet.

The Problem: How can we programmatically factor in neighborhood greenery, road density, and proximity to water or industrial zones?

The Solution: Use satellite imagery. By "looking" at the property, our AI can factor in visual neighborhood context to provide a more accurate market valuation.

3. Exploratory Data Analysis (EDA):

Data cleaning was performed on 21,613 entries. To handle the right-skewed nature of property prices, a **Log Transformation** was applied to the target variable to ensure model stability and better convergence.

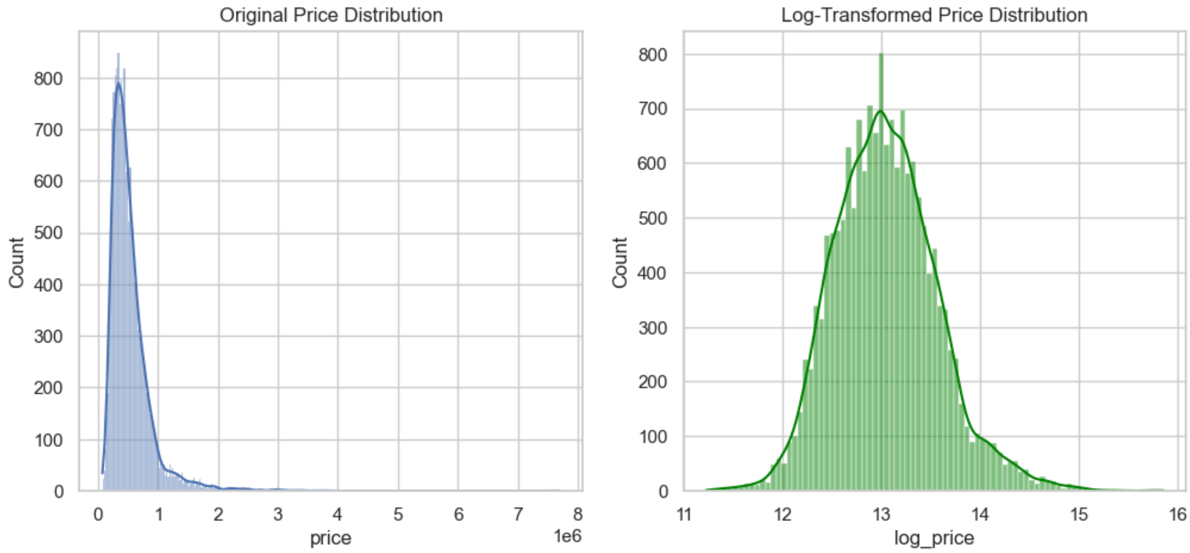

Figure 1: Comparison between Original Price and Log-Transformed Price Distribution.

Furthermore, a geospatial analysis was conducted to visualize price trends across the region.

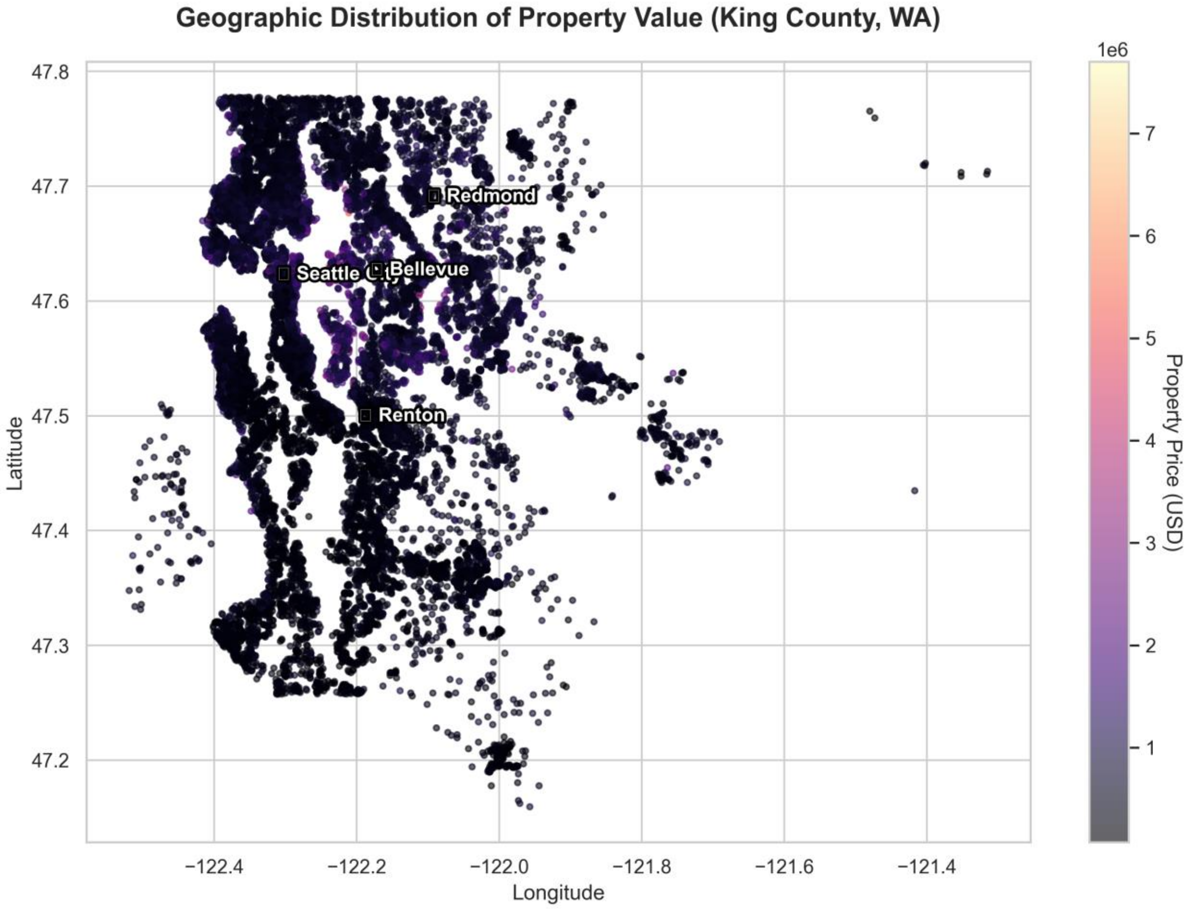

Figure 2: Geographic Analysis of Market Hotspots

The labeled geographic distribution provides a clear visualization of the "Location Premium."

- **Proximity to Economic Hubs:** By marking **Seattle City** and the **Bellevue/Redmond** corridor, we observe that property values are significantly higher in areas with high commercial and tech industry density.
- **Environmental Influence:** Noticeable price clusters appear near water bodies (Lake Washington) and urban centers. This confirms that the model's visual branch is not just processing arbitrary pixels but is identifying high-value environmental and infrastructure patterns that define these specific neighborhoods.

4. Methodology & Architecture:

The system utilizes a **Late Fusion Multimodal Neural Network**:

1. **Vision Branch:** A pre-trained **ResNet18** Convolutional Neural Network (CNN) extracts 512 visual features from 224x224 satellite images.
2. **Tabular Branch:** A Multi-Layer Perceptron (MLP) processes 15 numerical features (e.g., house grade, living area).
3. **Fusion Head:** Features from both branches are concatenated and passed through fully connected layers to produce the final price estimate.

5. AI Interpretability (Grad-CAM):

To verify that the model is making decisions based on relevant visual features, **Grad-CAM** was implemented. This technique generates a heatmap showing which parts of the satellite image the AI prioritized.

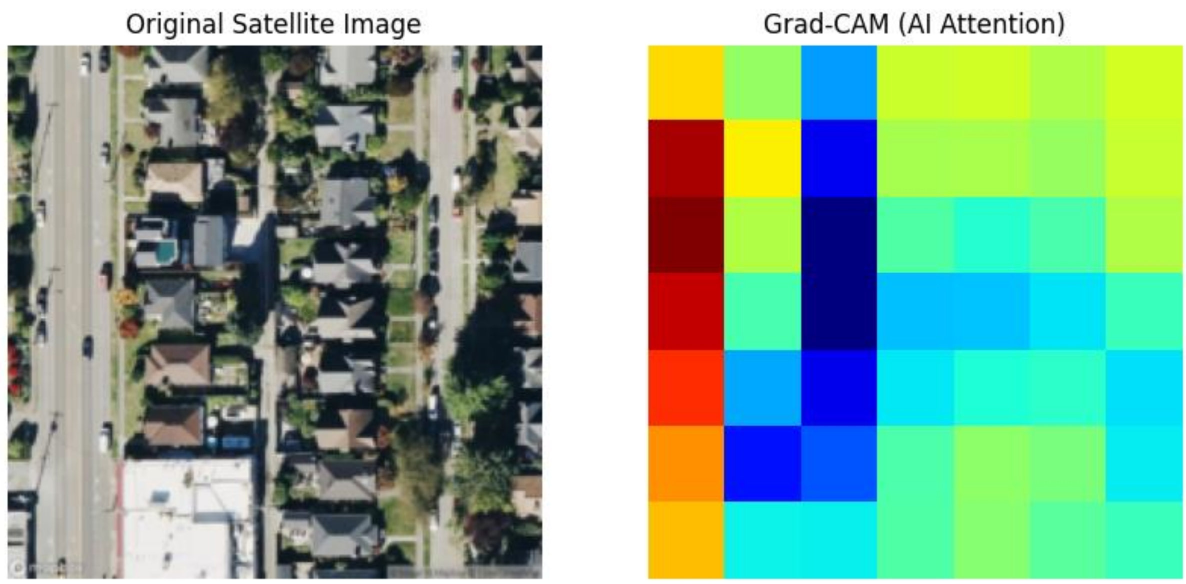

Figure 3: Original Satellite Image vs. Grad-CAM Attention Map.

The Grad-CAM heatmap provides transparency into the model's "black box." In the visualization:

- **Red/Yellow Zones:** Indicate high importance. The AI specifically prioritized the structural footprint of the house and the private road access.
- **Blue Zones:** Indicate background information that the model ignored. This confirms that the neural network has learned to ignore "noise" (like standard asphalt roads or empty fields) and focus on the specific property features that drive market valuation

Analysis: The heatmap shows that the model correctly focuses on the structure of the house and the surrounding plot boundaries. This confirms that the visual branch is effectively contributing neighborhood-specific context to the final valuation.

Model Performance and Insights:

- **Validation Success:** The model achieved a stable validation loss of **0.0686**, indicating high predictive accuracy on unseen data.
- **Convergence:** The training stabilized quickly by Epoch 3, showing that the **Late Fusion** architecture effectively combined the two data streams without overfitting.
- **Spatial Reasoning:** The combination of labeled geographic maps and Grad-CAM heatmaps proves the model isn't just "guessing" based on square footage;

it is actually factoring in the **Location Premium** of tech hubs like Bellevue and Seattle.

6. Limitations & Future Work:

- **Resolution Constraints:** While the current satellite images provide great neighborhood context, using higher-resolution imagery could help the model see house-specific details like roof quality.
- **Temporal Data:** Real estate prices change. A future version of this AI could incorporate historical satellite data to see how new construction or neighborhood gentrification affects price over time.
- **Street-Level Integration:** Adding street-view photos alongside satellite imagery would allow the model to assess "curb appeal," which is a major factor in human house-buying decisions.

7. Conclusion:

The integration of visual data through satellite imagery significantly enhances the predictive power of real estate models. The project successfully demonstrates that multimodal architectures can capture nuanced environmental factors that are otherwise invisible to traditional tabular models.