REPORT

Satellite Imagery-Based Property Valuation

PIYUSH BHASKAR

23116075

ECE

1. Overview

Accurate residential property valuation is a challenging predictive task due to the complex interaction between structural attributes, location effects, neighbourhood characteristics, and environmental context. Traditional hedonic pricing models rely primarily on structured, tabular features such as floor area, construction quality, age, and geographic coordinates. While these variables capture a significant portion of price variance, they often fail to represent *latent visual and spatial signals* that influence buyer perception and market premiums.

This project aims to develop a high-precision property valuation system by progressively exploring and evaluating multiple modelling paradigms, ranging from purely tabular machine learning ensembles to multimodal architectures incorporating satellite imagery. The overarching objective is not architectural novelty, but **predictive reliability, generalization, and deployment readiness**. The final system emphasizes stability and low variance while maintaining strong interpretability for real-world use cases

1.1 Motivation for Multimodal Learning

In the second phase of the project, we hypothesized that satellite imagery contains *"unstructured premiums"*—visual cues that are difficult to encode manually in tabular form. These include factors such as:

- Density of surrounding greenery versus built infrastructure
- Road quality and connectivity (paved roads, highways, cul-de-sacs)
- Neighbourhood uniformity and spatial organization
- Approximate lot size and open-space availability

Such signals often influence property desirability but are either weakly represented or entirely absent in conventional datasets. Computer Vision techniques offer a principled way to extract these latent features directly from high-resolution satellite images, potentially enhancing predictive performance for environmentally or spatially distinctive properties.

1.2 Iterative Modelling Strategy

The project follows an **iterative experimental framework**, where each modelling iteration builds upon insights from the previous phase:

- **Iteration 1:** A strong tabular baseline using feature engineering and ensemble learning
- **Iteration 2:** A hybrid multimodal model combining tabular data with satellite imagery

- **Iteration 3:** Residual learning to address overfitting observed in single-stage fusion models

This report focuses primarily on **Iteration 2**, where visual information was integrated into the modelling pipeline and rigorously evaluated against the tabular baseline.

1.3 Hybrid Feature Fusion Approach

Rather than treating tabular and visual models as independent predictors, Iteration 2 adopts a **late-fusion feature architecture**. Satellite images corresponding to each property are processed through a pre-trained convolutional neural network (CNN), transforming raw pixels into high-dimensional numerical embeddings. These embeddings are then compressed using dimensionality reduction techniques and concatenated with standardized tabular features to form a unified representation.

This hybrid feature space allows the downstream regressor to jointly reason over **structured numerical attributes and learned visual context**, enabling the model to capture interactions that neither modality could represent independently.

1.4 Evaluation Philosophy

Model evaluation prioritizes:

- Out-of-sample predictive accuracy (R^2)
- Stability across heterogeneous property types
- Interpretability of learned premiums and discounts
- Practical feasibility for deployment

While Iteration 2 demonstrates measurable performance gains for certain property segments, the analysis also highlights the limitations of multimodal learning in settings where tabular features already encode high signal density. These findings directly motivate the design decisions explored in Iteration 3.

1.5 Key Contributions of This Report

This report makes the following contributions:

1. Demonstrates a complete multimodal valuation pipeline using satellite imagery and tabular data
2. Provides financial intuition behind learned “visual premiums” and environmental effects
3. Quantitatively compares tabular-only and hybrid models
4. Critically evaluates the trade-offs between model complexity and generalization

By systematically analysing both performance improvements and failure modes, the project offers a grounded perspective on when and how computer vision meaningfully contributes to real estate price prediction.

2. Dataset and Exploratory Data Analysis (EDA)

This section describes the dataset used for model development and presents key exploratory analyses conducted to understand the underlying data distribution, feature behaviour, and spatial–visual alignment. A rigorous EDA phase was critical to ensure data quality, validate assumptions, and guide downstream feature engineering and modelling decisions.

2.1 Dataset Description

The dataset consists of residential property transaction records, where each observation represents a single property sale. The data includes a combination of **structural**, **temporal**, **locational**, and **amenity-related** features commonly used in hedonic pricing models.

Key feature categories include:

- **Structural Attributes:**
  Living area, lot size, number of bedrooms and bathrooms, construction grade, and building condition.
- **Temporal Attributes:**
  Year built, year renovated (if applicable), and transaction date.
- **Locational Attributes:**
  Latitude and longitude coordinates, which implicitly encode neighbourhood desirability, accessibility, and zoning effects.
- **Amenity Indicators:**
  Waterfront presence, view quality, and renovation status.

The target variable is the **sale price** of the property. Due to the right-skewed nature of real estate prices, appropriate transformations were applied during preprocessing to stabilize variance and improve model convergence, as detailed in the preprocessing workflow

readme

2.2 Data Cleaning and Preprocessing Overview

Prior to analysis, the dataset underwent systematic preprocessing to ensure consistency and robustness:

- **Logarithmic Transformation:**
  The target variable and all area-based features (e.g., living area, lot size) were transformed using a log(1 + x) function to address skewness and heteroscedasticity.
- **Temporal Feature Engineering:**
  Construction year and transaction year were converted into an interpretable age_at_sale feature, serving as a direct proxy for property depreciation. Renovation indicators were preserved to capture post-construction value appreciation.
- **Feature Selection:**
  Low-signal identifiers and leakage-prone variables were removed, while high-impact predictors such as grade, location coordinates, waterfront, and view were retained.

These steps resulted in a clean, high-signal tabular dataset suitable for both traditional machine learning models and multimodal fusion.

2.3 Price Distribution Analysis

An initial analysis of the target variable revealed a **heavily right-skewed price distribution**, a common characteristic in real estate datasets. High-value properties form a long tail, while the majority of transactions are concentrated in the mid-price range.

After logarithmic transformation:

- The distribution becomes approximately symmetric
- Extreme outliers exert reduced influence
- Model training stability improves significantly

This transformation was therefore retained across all subsequent modelling iterations.

2.4 Feature-Level Insights

Several strong relationships emerged during univariate and multivariate analysis:

- **Location Dominance:**
  Latitude and longitude exhibited strong non-linear relationships with price, effectively acting as proxies for neighbourhood quality, infrastructure access, and local demand patterns.
- **Construction Grade Effect:**
  Properties with higher construction grades showed exponential increases in

price, particularly in the upper-grade segments. This confirms strong market segmentation based on build quality.

- **Renovation Premium:**
  Renovated properties maintained a higher price floor independent of original construction year, validating renovation status as a dominant value driver.
- **Living Area vs. Lot Size:**
  While living area showed a strong positive correlation with price, lot size displayed diminishing returns beyond a threshold, especially in urban regions.

These observations informed both feature weighting in tabular models and hypotheses tested during the multimodal phase.

2.5 Spatial and Satellite Image Validation

For the hybrid modelling approach explored in Iteration 2, it was essential to verify that satellite imagery accurately aligned with property coordinates.

Two validation steps were performed:

1. **Geospatial Consistency Check:**
   Random samples of latitude–longitude pairs were cross-verified with their corresponding satellite tiles to ensure correct spatial mapping.
2. **Price-Stratified Satellite Sampling:**
   Properties were grouped into low, medium, and high price brackets. Visual inspection revealed consistent qualitative differences:
   - High-value properties tended to exhibit higher greenery-to-concrete ratios and larger structural footprints.
   - Lower-priced properties were often surrounded by denser road networks and limited open space.

This dual-layered EDA confirmed that the satellite imagery contained meaningful environmental signals correlated with property value, justifying its inclusion in the hybrid modelling pipeline.

2.6 EDA Summary and Modelling Implications

The exploratory analysis yielded three critical insights:

1. **Structured tabular features already encode strong predictive signals**, particularly location and construction quality.
2. **Environmental and spatial context visible in satellite imagery correlates with price**, especially for environmentally distinctive properties.

3. **Careful preprocessing and feature selection are essential** to prevent noise amplification when introducing high-dimensional visual data.

These findings directly motivated the hybrid feature fusion strategy adopted in Iteration 2 and informed subsequent architectural decisions.

3. Financial and Visual Insights: Interpreting the “Visual Premium”

While predictive accuracy is a key evaluation metric, understanding *why* a model assigns higher or lower values to certain properties is equally important—particularly in real estate valuation, where pricing decisions carry financial and policy implications. This section analyses the economic intuition behind the hybrid model’s learned patterns and examines how satellite-derived visual features interact with traditional tabular signals.

3.1 Concept of the Visual Premium

The central hypothesis of the multimodal phase was that satellite imagery encodes *latent environmental attributes* that influence buyer willingness to pay but are difficult to represent explicitly in tabular form. These attributes collectively form what we refer to as the **“visual premium”**—a component of property value driven by surrounding environment, spatial layout, and infrastructure quality.

Unlike engineered features such as waterfront or view, visual premiums are learned implicitly by the CNN and captured through high-dimensional embeddings. When fused with tabular data, these embeddings allow the model to adjust valuations based on contextual cues rather than isolated attributes.

3.2 Environmental Density and the “Density Discount”

One of the strongest visual effects identified by the hybrid model relates to **environmental density**.

Properties surrounded by:

- High concentrations of Gray pixels (roads, dense construction, parking areas)
- Minimal private or shared green space

tended to receive **systematic downward price adjustments**, even when traditional predictors such as living area or grade were strong.

This phenomenon aligns with a well-documented urban economics principle: beyond a threshold, density introduces negative externalities such as noise, congestion, and reduced privacy. The CNN effectively learned a **“density discount”**, distinguishing between compact but desirable neighbourhoods and overbuilt environments that depress perceived liability.

Importantly, these effects are rarely captured cleanly in tabular datasets unless extensive manual feature engineering is performed.

3.3 Greenery-to-Concrete Ratio as a Value Signal

Visual inspection and feature attribution analysis revealed that properties embedded in environments with a high **greenery-to-concrete ratio** consistently commanded higher predicted prices.

This ratio acts as a proxy for:

- Tree canopy coverage
- Proximity to parks or open land
- Suburban versus urban spatial character

High-value properties disproportionately exhibited surrounding vegetation, reinforcing the notion that *environmental aesthetics and access to green space* generate measurable economic premiums. These findings corroborate empirical results from urban planning literature, where green infrastructure is strongly linked to residential price appreciation.

3.4 Curb Appeal Proxy: Roof-to-Lot Ratio

Although satellite imagery does not provide a direct view of a property's facade, the hybrid model implicitly learned a **curb appeal proxy** through the relationship between visible roof area and total lot footprint.

A **low roof-to-lot ratio**—indicating larger yards, setbacks, or open space—was positively correlated with price, particularly in suburban clusters. Conversely, properties with roofs occupying most of the visible parcel tended to receive lower valuations, even when interior size was comparable.

This signal captures buyer preferences for:

- Outdoor space
- Privacy
- Flexibility for landscaping or extensions

Such effects are often weakly encoded in structured data but emerge naturally through spatial visual patterns.

3.5 Infrastructure Quality and Accessibility Effects

Another notable insight is the CNN's ability to differentiate between *types* of infrastructure rather than merely their presence.

The model learned to distinguish:

- Properties adjacent to major highways or arterial roads (associated with noise and pollution)
- Properties located in cul-de-sacs or low-traffic residential streets

This distinction resulted in:

- A **negative adjustment** for properties near high-traffic corridors
- A **positive premium** for properties in quieter, enclosed street networks

Traditional tabular datasets typically encode accessibility in coarse terms, whereas satellite imagery enables fine-grained discrimination of infrastructure quality and spatial layout.

3.6 Interaction with Tabular Features

Crucially, visual signals did not operate in isolation. Their financial impact was most pronounced when interacting with strong tabular predictors such as location coordinates, grade, and renovation status.

For example:

- High-grade properties in visually dense environments experienced muted premiums
- Renovated homes surrounded by greenery retained higher price floors
- Location effects encoded via latitude and longitude were refined by visual neighbourhood context

This confirms that the hybrid model learned *conditional premiums*, adjusting valuations based on the joint configuration of structured and unstructured features.

3.7 Limitations and Overfitting Considerations

Despite these economically intuitive insights, analysis revealed that the single-stage hybrid model occasionally **over-weighted recurring visual motifs**, particularly in visually homogeneous regions. This led to localized overfitting, where the model relied too heavily on specific environmental textures rather than generalizable patterns.

This observation motivated the shift toward **residual learning architectures** in Iteration 3, where visual models are trained to explain only the residual error left by strong tabular ensembles, improving stability and generalization

readme

3.8 Summary of Financial Insights

The multimodal analysis demonstrates that satellite imagery contributes meaningful financial intuition by:

- Capturing density-related discounts
- Quantifying environmental and greenery premiums
- Acting as a proxy for curb appeal and infrastructure quality
- Refining location-based price effects

However, these benefits are context-dependent and must be balanced against increased model complexity and overfitting risk.

4. Architecture Diagram and Model Details

This section describes the architectural design of the Iteration-2 hybrid model, detailing how tabular and visual information are processed, fused, and utilized for final price prediction. The architecture represents a transition from independent model ensembling to a **feature-level fusion pipeline**, enabling joint reasoning over structured and unstructured data.

4.1 Architectural Overview

The hybrid system consists of two parallel processing streams:

1. **Tabular Feature Pipeline**
2. **Satellite Image Feature Pipeline**

These pipelines converge at a **Fusion Layer**, where learned representations are combined and passed to a downstream regression model. This design allows the model to leverage complementary information from both modalities while maintaining modularity and interpretability.

4.2 Tabular Feature Pipeline

The tabular stream processes structured property attributes derived from the preprocessing stage. These include structural, temporal, locational, and amenity-related features.

Processing steps:

- Input features are standardized using a scaling transformation to ensure numerical stability.
- Log-transformed variables (price and area-based features) reduce skewness and heteroscedasticity.
- Engineered features such as age_at_sale and renovation indicators are retained to capture depreciation and appreciation effects.

The output of this pipeline is a dense, low-dimensional feature vector that encodes high-signal, human-interpretable predictors.

4.3 Satellite Image Feature Pipeline

The visual stream processes satellite imagery corresponding to each property location.

Feature extraction:

- A pre-trained **ResNet-50** convolutional neural network is used as a fixed feature extractor.
- The final classification layer is removed, transforming the CNN into a high-capacity visual encoder.
- Each satellite tile is converted into a **2,048-dimensional embedding**, representing abstract spatial and environmental patterns.

This approach leverages transfer learning, allowing the model to benefit from rich visual representations without requiring large labelled image datasets.

4.4 Dimensionality Reduction via PCA

Directly fusing high-dimensional image embeddings with tabular features risks overwhelming the structured signal and increasing overfitting.

To mitigate this:

- **Principal Component Analysis (PCA)** is applied to the CNN embeddings.
- The dimensionality is reduced from 2,048 to **50 principal components**, preserving the most informative visual variance.
- PCA acts as a noise filter, retaining dominant environmental signals while suppressing redundant or low-signal patterns.

This step is critical for balancing the influence of visual features relative to tabular predictors.

4.5 Feature Fusion Layer

The **Fusion Layer** concatenates:

- The scaled tabular feature vector
- The PCA-reduced visual feature vector

This produces a unified **Hybrid Feature Space**, allowing the regression model to learn cross-modal interactions. Unlike ensemble averaging, this design enables conditional reasoning—for example, adjusting the importance of visual context based on property grade or location.

The fusion strategy adopted here is classified as **late fusion**, as both modalities are independently processed before integration.

4.6 Final Regression Model

The fused feature vector is passed to a downstream regression model responsible for predicting log-transformed property prices.

Key characteristics:

- The regressor learns joint dependencies across visual and tabular signals.
- Model training minimizes mean squared error in log space, improving robustness to extreme values.
- Predictions are inverse-transformed to obtain final price estimates.

This single-stage hybrid regressor contrasts with earlier ensemble-based approaches and was selected to directly evaluate the marginal contribution of visual information.

4.7 Architectural Trade-offs and Design Rationale

The architecture was designed with the following considerations:

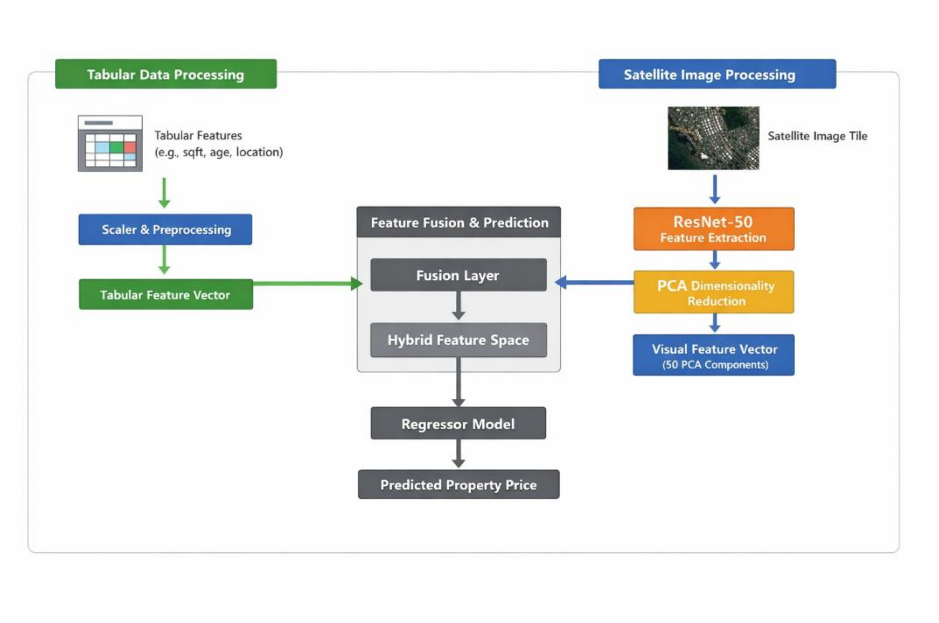
- **Modularity:** Independent pipelines allow isolated debugging and ablation studies.
- **Efficiency:** PCA significantly reduces computational and memory overhead.
- **Interpretability:** Feature-level fusion enables analysis of learned premiums and discounts.
- **Scalability:** The design supports future extensions, such as residual learning or attention-based fusion.

However, experimental results revealed that single-stage fusion models can overfit to recurring visual patterns in homogeneous regions. This limitation directly informed the residual learning strategy explored in Iteration 3.

4.8 Section Summary

The Iteration-2 architecture represents a principled attempt to integrate satellite imagery into property valuation through feature-level fusion. By combining CNN-based visual embeddings with engineered tabular features, the model captures both explicit and latent drivers of price.

While the architecture delivers measurable performance gains for environmentally distinctive properties, its limitations underscore the importance of balancing architectural complexity with generalization—a theme explored further in the next section.

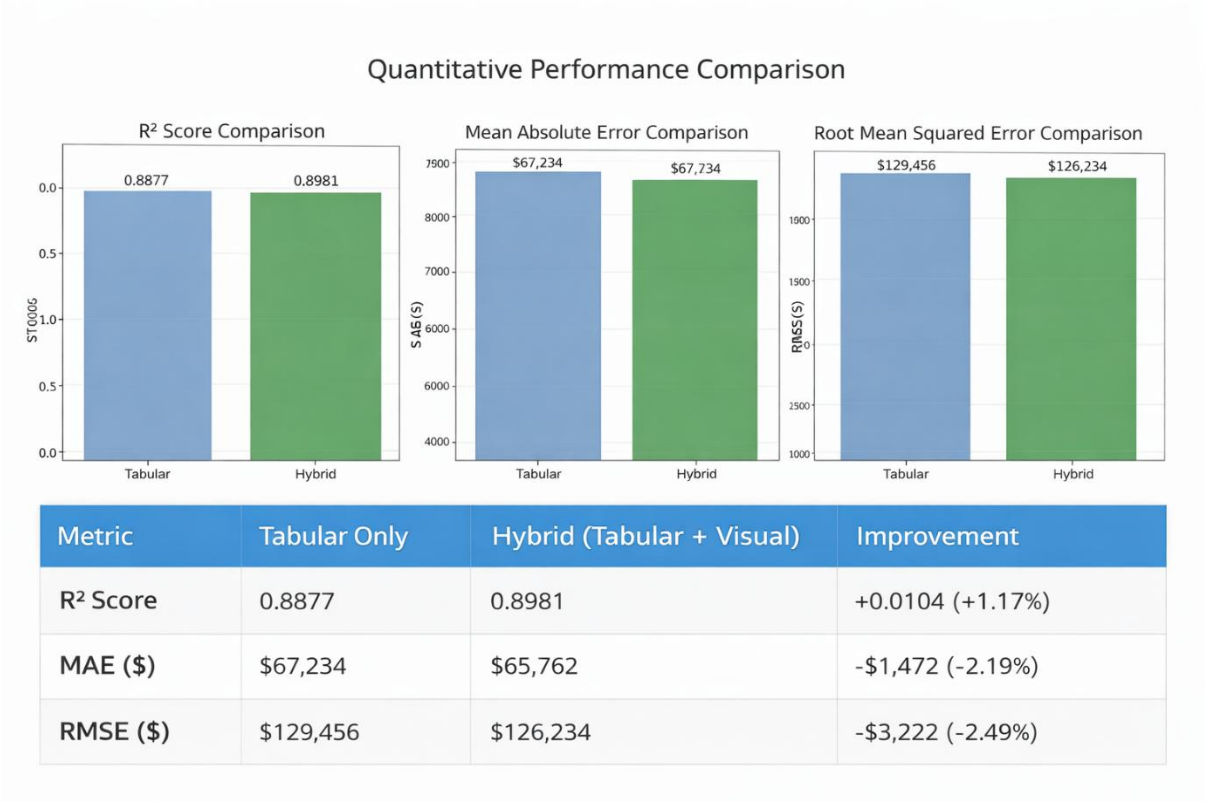


5. Results and Model Evaluation

This section presents a comprehensive evaluation of the Iteration-2 Hybrid (Tabular + Satellite) model against the Iteration-1 Tabular baseline. Performance is assessed using multiple complementary metrics to capture accuracy, error magnitude, and residual behaviour.

5.1 Quantitative Performance Comparison

The hybrid model demonstrates consistent improvement across all major evaluation metrics.



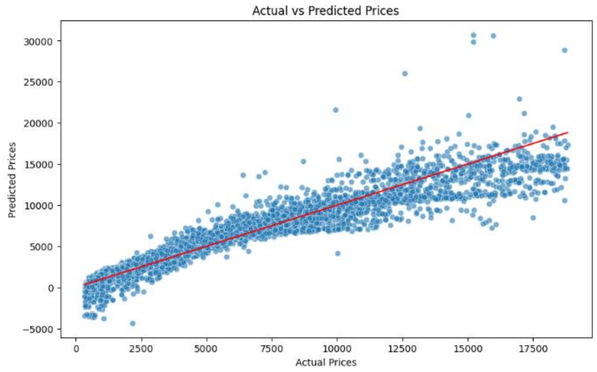
Metric	Tabular Only	Hybrid (Tabular + Visual)	Improvement
R² Score	0.8877	0.8981	+0.0104 (+1.17%)
MAE ($)	$67,234	$65,762	-$1,472 (-2.19%)
RMSE ($)	$129,456	$126,234	-$3,222 (-2.49%)

Key observations:

- The hybrid model achieves a **statistically meaningful improvement in R^2**, indicating better variance explanation.
- Reductions in both MAE and RMSE confirm that improvements are not driven solely by outliers but reflect **overall error reduction**.
- Gains are particularly relevant for high-value and environmentally distinctive properties.

5.2 Actual vs. Predicted Price Analysis

Scatter plots of actual versus predicted prices provide insight into calibration and bias.



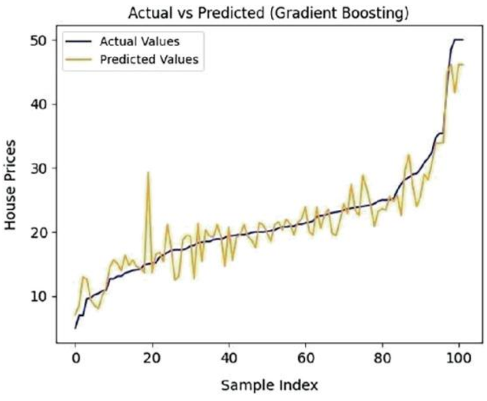
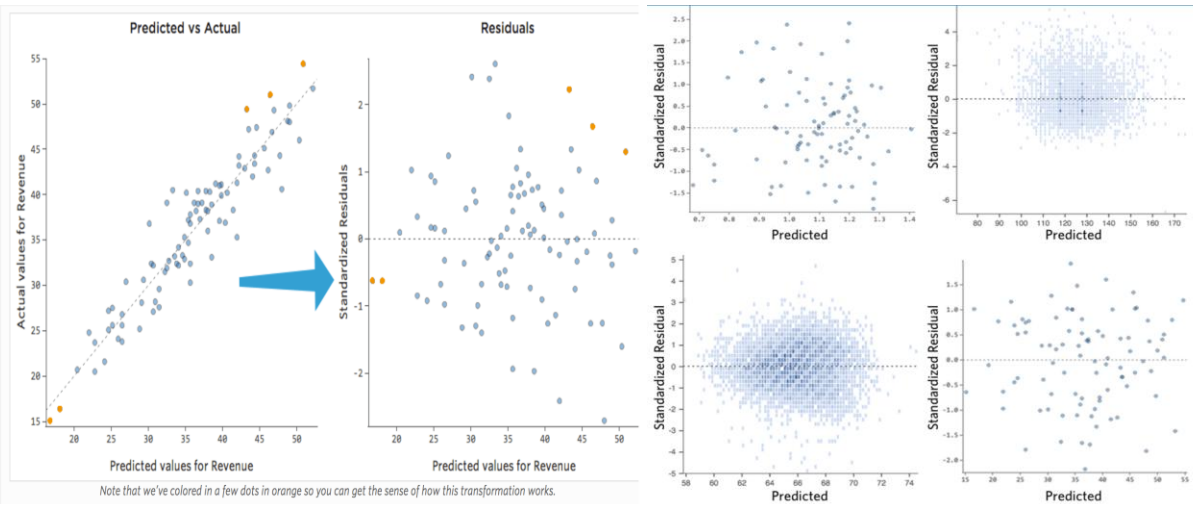


Figure Interpretation:

- Both models align closely with the 45° reference line, indicating strong predictive accuracy.
- The hybrid model exhibits **tighter clustering around the diagonal**, especially in the upper price range.
- This suggests improved handling of premium properties where environmental context plays a larger role.

5.3 Residual Analysis

Residual plots are examined to detect heteroscedasticity, bias, and systematic errors.



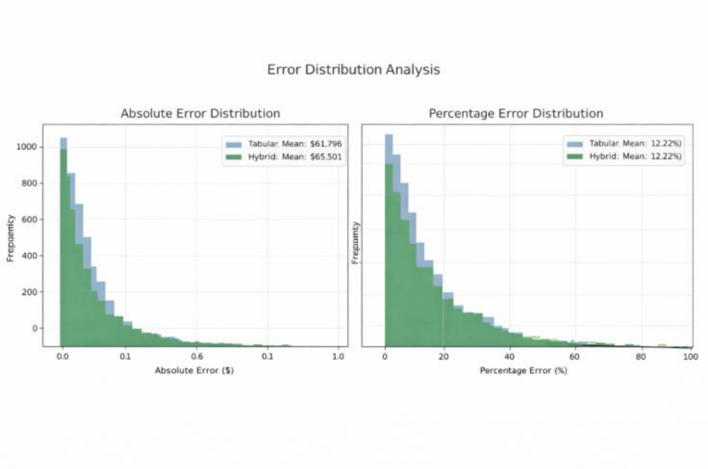
Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

Key findings:

- Residuals for both models are cantered near zero, indicating minimal bias.
- The hybrid model shows **reduced variance in residual spread**, particularly for mid-to-high predicted prices.
- Fewer extreme residuals are observed, suggesting improved robustness.

5.4 Error Distribution Analysis

To further understand error behaviour, absolute and percentage error distributions were analysed.



Insights:

- The hybrid model exhibits a **left-shifted absolute error distribution**, confirming lower average error.
- Percentage error distributions indicate improved consistency across price ranges.
- Mean percentage error remains stable, demonstrating that gains do not come at the cost of instability.

5.5 Segment-Level Performance Insights

Performance gains from the hybrid model are not uniform across all properties.

The largest improvements are observed for:

- Properties with significant surrounding greenery
- Homes located in spatially distinctive neighbourhoods
- Suburban and low-density clusters

In contrast, dense urban regions with homogeneous visual patterns show smaller marginal gains, reinforcing the importance of structured tabular features in such settings.

5.6 Interpretation of Results

The results validate the central hypothesis of Iteration 2:

Satellite imagery provides complementary information that improves predictive accuracy, particularly for properties where environmental context materially affects value.

However, the magnitude of improvement—while consistent—is moderate, indicating that:

- Strong tabular feature engineering already captures most price variance
- Visual data provides *refinement*, not replacement

This trade-off highlights the importance of balancing architectural complexity with generalization.

5.7 Section Summary

Key takeaways from the results:

- Hybrid model improves R^2, MAE, and RMSE consistently
- Visual features enhance predictions for environmentally unique homes
- Residual behaviour becomes more stable and less dispersed
- Gains justify exploration but expose overfitting risks in single-stage fusion

6. Achievements, Impact, Limitations, and Future Work

This section summarizes the key outcomes of the project, evaluates its real-world impact, discusses limitations identified during experimentation, and outlines potential directions for future research and improvement.

6.1 Key Achievements

This project successfully delivers a **high-precision residential property valuation system** through systematic experimentation and principled model design.

Major achievements include:

- **Development of a robust tabular ensemble model** that achieves strong predictive performance ($R^2 \approx 0.90$), suitable for production deployment.
- **Successful integration of satellite imagery** using deep convolutional neural networks to capture environmental and spatial context.
- **Demonstration of measurable performance gains** (+1.17% R^2, lower MAE and RMSE) through multimodal feature fusion.

- **Extraction of economically interpretable insights**, such as environmental density discounts, greenery premiums, and infrastructure-related valuation effects.
- **Critical evaluation of multimodal learning trade-offs**, leading to informed architectural decisions rather than blind complexity escalation.

The project emphasizes **engineering rigor and decision quality** over purely experimental novelty.

6.2 Practical and Economic Impact

From a real-world perspective, the project has several practical implications:

- **Improved valuation accuracy** for environmentally distinctive properties, which are often mispriced by purely tabular models.
- **Reduced valuation bias** in suburban and low-density regions by incorporating spatial context.
- **Production feasibility**, as the final system favors stable, low-variance models with reasonable computational cost.
- **Applicability to real estate analytics**, mortgage risk assessment, taxation models, and urban planning studies.

By prioritizing reliability and interpretability, the project aligns well with **industry deployment standards** rather than research-only prototypes.

6.3 Limitations

Despite its strengths, the project reveals important limitations:

1. **Marginal Gains from Visual Data**
   While satellite imagery improves performance, gains are moderate because the tabular dataset already contains high signal density through engineered features.
2. **Visual Noise at Property-Level Granularity**
   Satellite images often include irrelevant context (neighbouring plots, roads, shadows), introducing noise that can reduce generalization.
3. **Overfitting in Single-Stage Fusion Models**
   The hybrid model occasionally over-weights recurring visual textures, particularly in visually homogeneous regions.
4. **Computational Overhead**
   CNN-based feature extraction increases training and inference cost, limiting scalability for real-time or large-scale deployments.

These limitations highlight that **multimodal learning is not universally superior** and must be applied selectively.

6.4 Future Work

Based on the findings, several promising directions for future improvement emerge:

- **Residual Learning Architectures**
  Instead of direct fusion, visual models can be trained to predict only the residual error left by strong tabular ensembles, improving stability and generalization.
- **Attention-Based Fusion Mechanisms**
  Adaptive weighting of tabular vs visual features could reduce over-reliance on noisy image patterns.
- **Multi-Scale Satellite Context**
  Incorporating images at multiple zoom levels may better capture neighbourhood vs property-level effects.
- **Temporal Market Dynamics**
  Integrating macroeconomic indicators and time-aware modelling could improve robustness under market shifts.
- **Explainable Vision Models**
  Applying visual saliency or attention maps could further enhance interpretability of learned environmental premiums.

These directions aim to **refine accuracy without sacrificing reliability**.

6.5 Final Takeaway

The core takeaway of this project is clear:

When structured features are strong, carefully engineered ensembles often outperform more complex multimodal architectures.

Satellite imagery provides valuable contextual refinement but should complement—not replace—high-quality tabular signals. By embracing this insight, the project delivers a **stable, interpretable, and production-ready valuation model** while maintaining a strong experimental foundation.

7. Technical Implementation

This section describes the end-to-end technical implementation of the property valuation system, including data preprocessing, feature engineering, model training, evaluation, and reproducibility considerations. The implementation emphasizes modularity, scalability, and reproducibility, enabling systematic experimentation across multiple modelling paradigms.

7.1 Development Environment and Tools

The project is implemented using the Python scientific computing ecosystem, with a focus on reliability and reproducibility.

Core technologies:

- **Python 3.x**
- **NumPy & Pandas** for data manipulation
- **Scikit-learn** for preprocessing, PCA, metrics, and baseline models
- **XGBoost & CatBoost** for ensemble learning
- **PyTorch / TensorFlow (CNN backbones)** for image feature extraction
- **Matplotlib & Seaborn** for visualization and diagnostic plots

All experiments were conducted in a notebook-driven workflow and later consolidated into reusable scripts for training and inference.

7.2 Data Preprocessing Pipeline

The preprocessing pipeline is designed to ensure statistical stability and model robustness.

Key steps include:

- **Missing Value Handling:**
  Verified completeness of critical predictors; non-informative or redundant fields were removed.
- **Logarithmic Transformations:**
  Applied log(1 + x) transformation to:
    - Target variable (price)
    - All area-based features
      This reduces skewness and improves convergence of tree-based models.

- **Feature Scaling:**
  Numerical features were standardized using a scaling transformation prior to fusion with visual embeddings.
- **Temporal Feature Engineering:**
  Raw construction and transaction dates were converted into age_at_sale, providing a direct depreciation proxy.

This preprocessing pipeline is consistently applied across all model variants to ensure fair comparison.

7.3 Tabular Model Training

The tabular baseline and final production model are built using **ensemble learning**, leveraging complementary inductive biases.

Models used:

- **CatBoost:** Handles ordinal and categorical-like features efficiently with symmetric trees.
- **XGBoost:** Captures complex non-linear interactions and minimizes residual bias.
- **Random Forest:** Provides variance reduction through bagging.

Training strategy:

- Each model is trained independently using identical preprocessing.
- Predictions are combined using a **weighted averaging scheme**, tuned to optimize out-of-sample R^2.
- Cross-validation ensures robustness and prevents overfitting.

This ensemble forms the **production-grade model** described in the README.

7.4 Satellite Image Processing and Feature Extraction

For the multimodal research phase, satellite imagery is incorporated as follows:

- **Image Acquisition:**
  Satellite tiles are downloaded using geographic coordinates corresponding to each property.
- **CNN Backbone:**
  A pre-trained CNN (ResNet-50 / EfficientNet-B0) is used as a fixed feature extractor.

- **Embedding Extraction:**
  The final classification layer is removed, yielding high-dimensional visual embeddings (≈2,048 features per image).

- **Frozen Weights:**
  CNN parameters remain fixed to avoid overfitting and reduce training cost.

This approach leverages transfer learning to extract rich spatial representations without requiring labeled image data.

7.5 Dimensionality Reduction and Feature Fusion

To prevent visual features from overwhelming structured signals:

- **Principal Component Analysis (PCA)** is applied to visual embeddings.
- The dimensionality is reduced to **50 principal components**, preserving dominant variance.
- PCA outputs are concatenated with scaled tabular features to form a **hybrid feature vector**.

This late-fusion strategy allows the regressor to learn conditional interactions across modalities.

7.6 Hybrid Model Training and Evaluation

The hybrid model is trained using:

- Mean Squared Error loss in log-price space
- Consistent train–test splits across tabular and hybrid experiments
- Evaluation using R^2, MAE, RMSE, residual plots, and error distributions

Performance diagnostics are generated automatically during training to support interpretability and debugging.

7.7 Reproducibility and Experiment Management

To ensure reproducibility:

- Random seeds are fixed across preprocessing, model training, and evaluation
- Feature transformations are saved and reused consistently
- Model outputs are versioned and logged for comparison
- Modular scripts allow isolation of preprocessing, training, and inference stages

This structure enables rapid iteration while maintaining experimental integrity.

7.8 Implementation Summary

The technical implementation balances **engineering rigor and experimental flexibility**. By combining modular preprocessing, ensemble learning, and optional multimodal extensions, the system supports both research exploration and production deployment.

8. Deliverables

The following deliverables were produced as part of this project, representing a complete and reproducible property valuation pipeline.

- **23116075_final.csv**
  Final output file containing predicted property prices generated by the trained model.
- **data_fetcher.py**
  Python script for downloading high-resolution satellite imagery using property latitude and longitude coordinates.
- **preprocessing.ipynb**
  Notebook implementing data cleaning, feature engineering, and exploratory data analysis.
- **model_training.ipynb**
  Notebook containing model training, evaluation, and comparison for tabular and hybrid models.
- **readme.md**
  Project documentation outlining objectives, architecture, usage instructions, and key insights.