MORINGA
Discover · Grow · Transform

REAL ESTATE PRICE PREDICTOR AND ANOMALY DETECTOR

GROUP 1

Fatuma Tari · Faith Njuguna · Jeremiah Bii

Joanne Kariuki · Kelvin Omina · Michelle Ngunya

Michelle Maina

AGENDA

BUSINESS UNDERSTANDING

Business Problem: Information Asymmetry in Real Estate

Buyers and **renters** struggle to determine if property prices are fair and **sellers** often overprice or underprice property due to incomplete market data. This results in:

- Prolonged time on market
- Failed negotiations
- Financial losses
- Reduced trust in the housing market

The Question: *Is this property priced fairly compared to similar listings?*

BUSINESS UNDERSTANDING

Objectives

The project aims to improve pricing transparency in the Kenyan real estate market by:

- Classifying property listings into underpriced, fair, or overpriced categories
- Supporting better pricing decisions for buyers, renters, sellers, and agents
- Enabling scalable pricing analysis across cities and property types
- Providing interpretable outputs suitable for non-technical users (platform operators, real estate professionals)

The Question: *Is this property priced fairly compared to similar listings?*

OUR SOLUTION

A Smart Classification System

This project proposes a supervised machine learning classification system that learns pricing patterns from historical real estate data and property features such as:

- Location
- Property type
- Bedrooms, bathrooms, and toilets
- Amenities and listing characteristics
- Listing category (for rent / for sale)

The model will classify each property into one of three pricing categories:

🟢 **Underpriced** – Listed significantly below comparable market listings

🟡 **Fairly Priced** – Listed within a reasonable range of comparable market listings

🔴 **Overpriced** – Listed significantly above comparable market listings

WHO THIS HELPS

Key Stakeholders

Buyers & Renters

- Identify good deal and avoid overpriced listings
- Stronger negotiation position
- Reduced financial risk

Sellers & Property Owners

- Set competitive, evidence-based prices
- Reduce time on market
- Increase transaction success

Sellers & Property Owners

- Add data-backed pricing insights
- Enhance user trust and platform credibility

DATA OVERVIEW

Data Source

16,000+ Property Listings from Kenya Property Centre

Key Features

- Location: State, locality (focused on Nairobi, Kiambu, Kajiado, Mombasa)
- Property characteristics: Bedrooms, bathrooms, toilets, parking
- Property type: Houses, apartments
- Amenities: Furnished, serviced, shared status
- Price information: Rental (per month) and sale prices

Final Clean Dataset: 9,091 properties across 4 major counties and 30+ localities

DATA CLEANING

Major Issues Addressed

1. **Missing Data Handled**
2. **Structural Errors Fixed**
3. **Irrelevant Data Removed**
4. **Outlier Management**

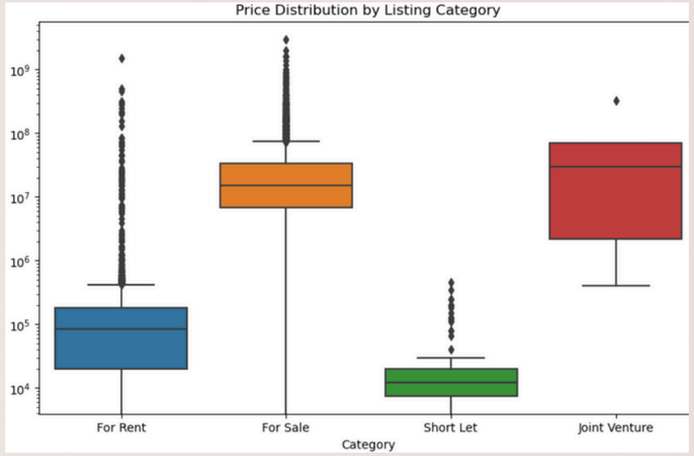
   Extreme bedroom counts: Capped at 10, Extreme bathroom counts.

Cleaning Results

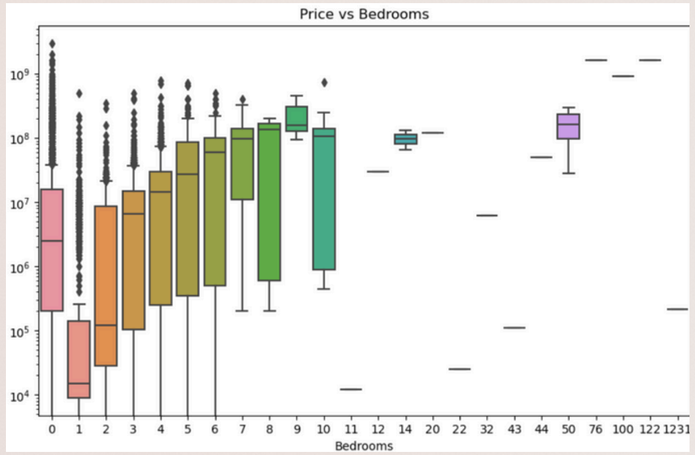
- **Starting dataset:** 16,117 properties
- **Final clean dataset:** 9,091 properties (56% retained)
- **Quality improvement:** Focused, consistent, residential-only data

EXPOLORATORY DATA ANALYSIS

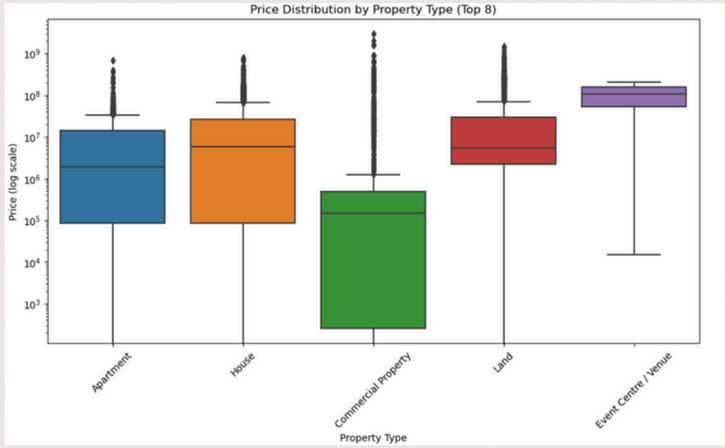
Price by listing category (rent vs sale)



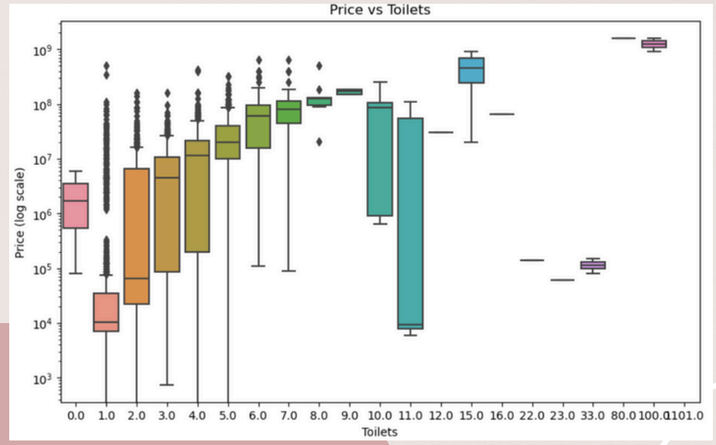
Price by Core Features(Price vs Bedrooms)



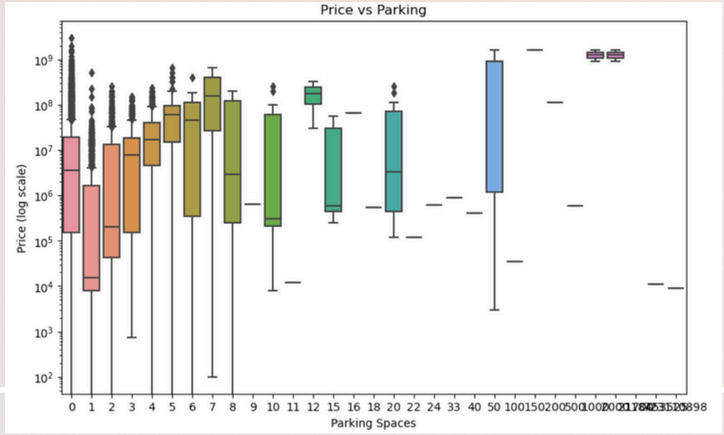
Price by Core Features(Price vs Property Types)



Price by listing category (rent vs Toilets)



Price by listing category (rent vs Parking)



Price by listing category (rent vs Location)

EXPOLORATORY DATA ANALYSIS

KEY FINDINGS FROM EDA

1. **Price Distribution Insights:**

- Highly right-skewed: Most properties priced low-to-mid range, few luxury outliers
- Sale vs. Rent: Sale prices 10-100x higher than monthly rents (as expected)

2. **Feature-Price Relationships:**

- Toilets: 0.124 correlation with price (strongest numeric predictor)
- Bathrooms: 0.122 correlation
- Bedrooms: 0.036 correlation (weaker than expected)
- Parking: -0.003 correlation (negligible)

3. **Property Size Effects:**

- More bedrooms - Higher prices (non-linear relationship)
- Bathrooms/toilets better predictors than bedrooms
- Parking shows minimal direct price impact

4. **Amenity Effects:**

- Furnished properties: 15-30% price premium
- Serviced properties: Moderate premium (10-20%)
- Shared occupancy: Slight discount (5-10%)
- Clear value signals: Amenities provide interpretable pricing signals

5. **Location Impact (Critical Finding)**

- Locality matters most: Westlands, Kilimani, Lavington command 200-400% premiums
- State-level variation: Nairobi significantly higher than other counties
- Price dispersion: Same-sized properties vary dramatically by location
- Implication: Location features will dominate model predictions

6. **Property Type Effects**

- Apartments vs. Houses: Apartments show wider price range
- Sub-type matters: Townhouses, Detached Duplexes command premiums
- Bedsitters: Distinct low-price cluster (studio apartments)

7. **Multicollinearity Detected**

- Bedrooms, bathrooms, toilets moderately correlated (0.3-0.5) *Not severe enough to require removal*

FEATURE ENGINEERING

Key transformations included:

- Creation of a price fairness label (price_label) based on market-relative pricing logic
- Extraction of temporal features - year, month, day_of_week
- One-hot encoding of categorical variables such as - Property type and subtype, State and other location indicators
- Boolean normalization for features like - Furnished, serviced, shared, parking

The final feature matrix excludes the target variable (price_label) to prevent leakage.

MODELLING

This project treats price fairness as a multiclass classification problem with three classes: Underpriced, Fairly priced and overpriced.

Models

Baseline model

Used as a baseline with class weighting to handle imbalance.

Tree based Models

Random Forest Classifier Chosen for:

- Nonlinear decision boundaries
- Robustness to outliers
- Interpretability via feature importance

XGBoost Classifier

Configured for multiclass classification using multi:softprob with
:

- Moderate tree depth
- Learning rate control for stability
- Subsampling and column sampling to reduce overfitting

Model Evaluation

Models were evaluated using: Accuracy, Precision, Recall, and F1-score per class

Results Summary

Logistic Regression

- Served as a strong baseline
- Struggled with nonlinear relationships and complex interactions

Random Forest

- Improved performance across all classes
- Demonstrated strong generalization
- Provided stable feature importance rankings

XGBoost (Best Performing Model)

- Achieved the highest overall F1-score
- Showed the best balance between precision and recall
- Handled class imbalance effectively

The final production candidate is the XGBoost multiclass classifier.

	precision	recall	f1-score	support
Underpriced	0.84	0.77	0.80	614
Fair	0.75	0.81	0.78	1077
Overpriced	0.80	0.75	0.77	617
accuracy			0.78	2308
macro avg	0.79	0.78	0.78	2308
weighted avg	0.78	0.78	0.78	2308

RECOMMENDATIONS

Based on the model outputs and exploratory analysis, the following business recommendations are proposed:

- Adopt FairPrice Check as a decision-support tool
- Real estate platforms, agencies, and property developers can integrate the model to flag listings that are likely overpriced or underpriced before publication.
- Use classification labels to guide pricing strategy
- Overpriced listings can be reviewed and adjusted to reduce time-on-market
- Underpriced listings can be repriced to capture lost revenue
- Fairly priced listings can be fast-tracked for marketing and promotion

Limitations

- Price labels are based on past market behavior
- Short-term rental dynamics are excluded
- Very high-end luxury properties are less common in the data
- Location is captured at an area level, not exact coordinates
- The model relies on listing data, not final transaction prices

CONCLUSION

Summary

FairPrice Check demonstrates how supervised machine learning can be used to transform noisy real estate listing data into a practical, interpretable pricing fairness tool. By framing price assessment as a classification task rather than a regression problem, the system provides actionable guidance that is easier for end users to trust and apply in real-world decision-making.

Access our tool here