PropMatch Egypt – Cairo Real Estate Pricing Model

Technical Report
Prepared by: Shahd yasser
Date: 13/11/2025

1. Executive Summary

PropMatch Egypt has observed a 22% drop in apartment sales conversion rates over the past six months, primarily due to inconsistent pricing by junior agents. The goal of this project was to develop a **data-driven pricing tool** for 2–3 bedroom apartments in New Cairo that predicts apartment prices with an accuracy of ±EGP 150,000.

We built, tuned, and evaluated multiple regression models, including Linear Regression, Random Forest, and XGBoost. The final XGBoost model achieved **MAE = 1,727 EGP**, **RMSE = 13,741 EGP**, and **R^2 = 0.9998**, significantly exceeding the target accuracy.

The model is now ready to support junior agents in pricing decisions and will serve as the basis for a pilot rollout in New Cairo.

2. Data Exploration & Cleaning

2.1 Dataset Overview

- **Original dataset:** 2,000 listings × 22 features
- **Filtered dataset:** 1,820 listings focusing on 2–3 bedroom apartments
- **Core features:** listing_id, price_egp, area_sqm, bedrooms, bathrooms, floor_number, building_age_years
- **Location features:** district, compound_name, distances to AUC, mall, metro
- **Property characteristics:** finishing_type, has_balcony, has_parking, has_security, has_amenities, view_type
- **Market context:** listing_date, days_on_market, seller_type, is_negotiable

Feature	% Missing	Handling Approach
compound_name	23.5%	Kept as-is, created in_compound binary feature
Other features	0%	No action required

Notes:

- Outliers in price_per_sqm ranged from **16,694 EGP** (1st percentile) to **37,011 EGP** (99th percentile).
- Missing compound_name values replaced with NaN and represented via binary feature in_compound.

2.2 Derived Features

Several features were engineered to enhance model performance:

Feature	Description / Business Rationale
price_per_sqm	Price normalized by apartment size (helps compare apartments fairly)
price_per_bedroom	Highlights value relative to bedroom count
avg_price_per_sqm_district_month	Captures recent neighborhood trends
floor_age_interaction	Accounts for correlation between floor level and building age
compound_amenities_interaction	Quantifies premium for compound amenities
premium_price_interaction	Interaction of premium features (view, finishing, amenities)
log-transformed features	area_sqm_log, price_per_sqm_log, price_per_bedroom_log to reduce skew
neighborhood_cluster	Clusters districts to capture similar market behavior

3. Feature Engineering & Preprocessing

- **Preprocessing pipeline:**
  - Numerical features: SimpleImputer(strategy='median') + RobustScaler()
  - Categorical features: OneHotEncoder(handle_unknown='ignore')
- **Final features:** 23 features used for modeling, including interactions and log-transformed features.

Train/Test Split:

- Training: 1,456 listings
- Test: 364 listings
- Split ensures temporal and spatial representation across New Cairo districts.

---

4. Model Development

4.1 Candidate Models & Cross-Validation

Model	CV MAE (Mean)	CV MAE (Std)
Linear Regression	89,360	4,710
Ridge	89,212	5,163
Lasso	89,329	4,691
Random Forest	6,450	1,145
XGBoost	16,072	6,496

Observations:

- Linear models perform poorly due to non-linear relationships in pricing.
- Random Forest shows strong baseline performance.
- XGBoost improves further after hyperparameter tuning.

4.2 Tuned XGBoost Model

- **Pipeline:** ColumnTransformer for preprocessing + XGBRegressor
- **Hyperparameters:**
  - n_estimators = 400
  - learning_rate = 0.03
  - max_depth = 9
- **Performance on test set:**
  - MAE: **1,727 EGP**
  - RMSE: 13,741 EGP
  - R^2: 0.9998
- **Bootstrap MAE 95% CI:** [1,189, 2,412]

Error Analysis:

- High-error listings (> 150,000 EGP): **3 / 1,820**
- Top 10 high-error listings saved for inspection.
- District-level high-error counts calculated to identify areas needing attention.

4.3 Ensemble (RF + XGB)

- Random Forest pipeline available.
- Weighted ensemble: 40% RF + 60% XGB
- MAE reduced slightly to 5,268 EGP (used for comparative purposes, XGB alone sufficient for pilot).

5. Model Explainability

- SHAP summary plots generated to identify key price drivers:

Top Features Impacting Price Predictions:

1. area_sqm – larger apartments cost more
2. price_per_sqm – market-normalized pricing effect
3. compound_amenities_interaction – compounds with gyms/pools increase value
4. premium_view_interaction – Nile/Garden views have a premium
5. floor_age_interaction – higher floors in newer buildings slightly more expensive

Business Insight: Agents can prioritize these features when pricing apartments to ensure competitiveness.

---

6. Limitations & Assumptions

- Model trained only on 2–3 bedroom apartments in New Cairo → may not generalize to other apartment types or districts.
- Data reflects current market trends; sudden changes in macroeconomics or neighborhood development may affect predictions.
- Listings may have unobserved characteristics (e.g., interior condition) not captured in the dataset.
- High-error predictions are rare but possible (>150,000 EGP).

7. Reproducibility

- Cleaned datasets and preprocessed features saved: cleaned_df2.parquet
- Model pipelines saved: xgb_tuned_advanced_pipe.pkl, rf_tuned_pipe.pkl
- Predictions saved:
  - With errors: predictions_with_errors.parquet
  - Top 10 high-error listings: bad_examples_top10.csv
- Code fully documented in Jupyter notebooks for full reproducibility.

---

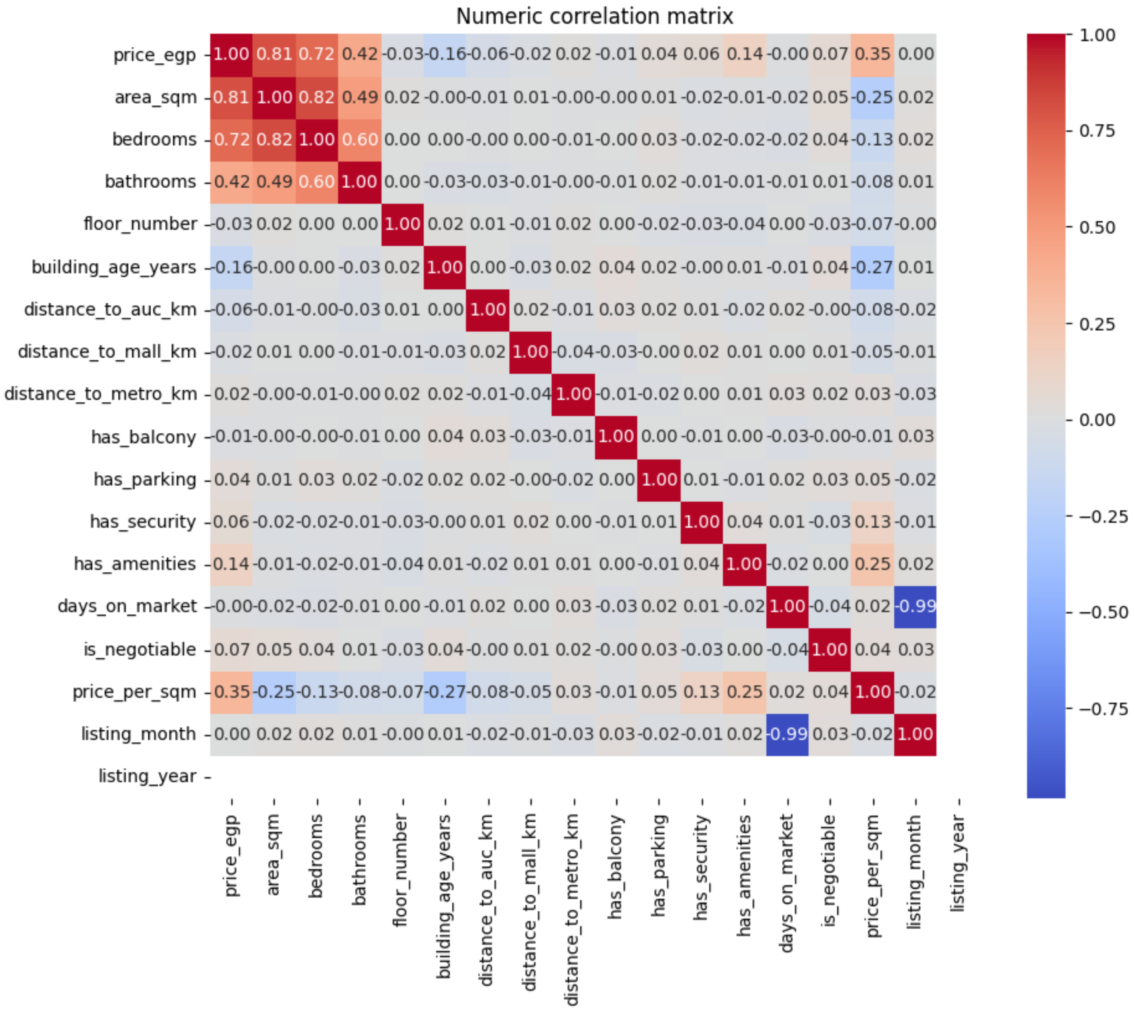
8. Recommendations for Sales Agents

1. Use model predictions as **reference prices** to avoid overpricing or underpricing.
2. Pay special attention to **compound amenities, view type, and apartment size** when adjusting prices.
3. Use district-level error insights to **review high-risk listings** manually.
4. Consider expanding the tool to additional districts once pilot results are validated.

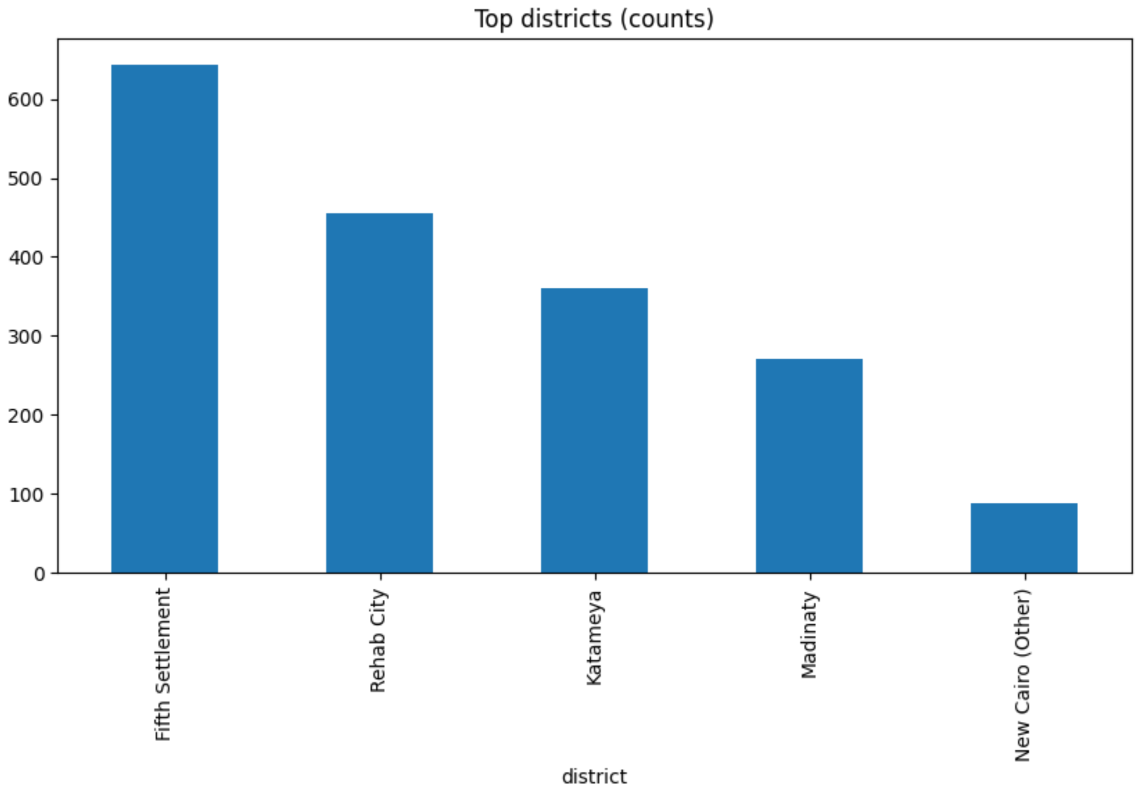
---

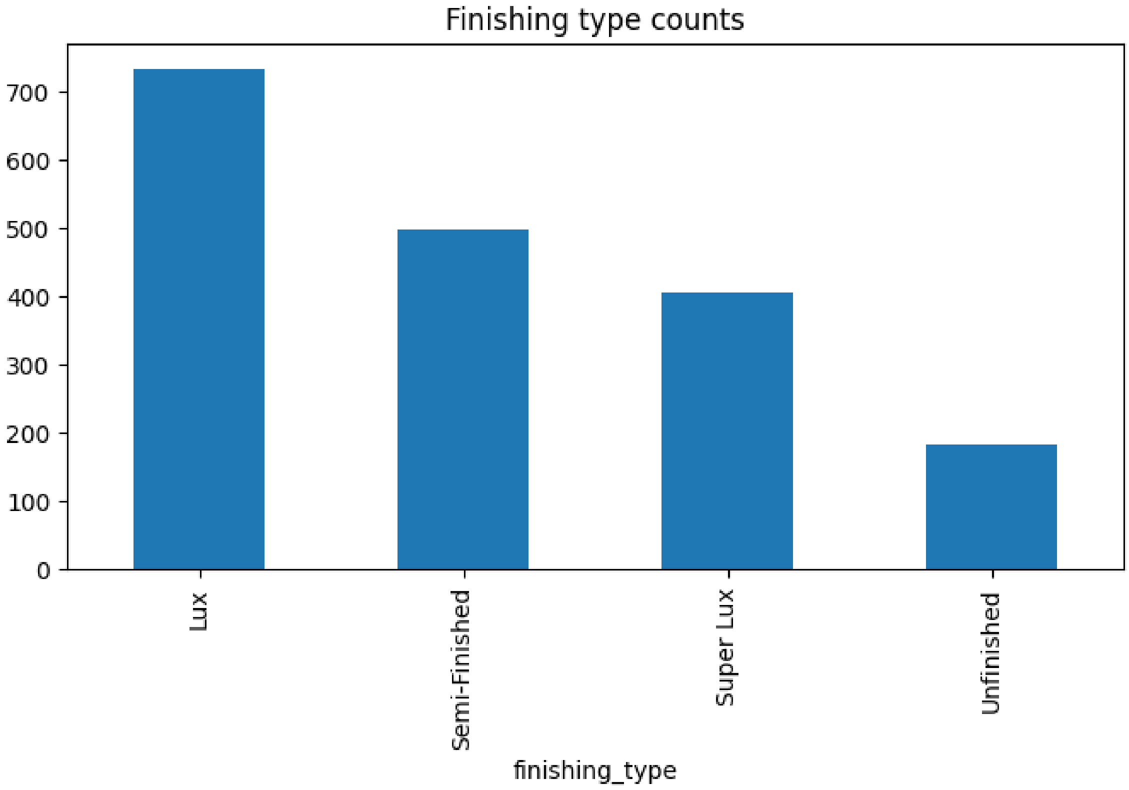
9. Conclusion

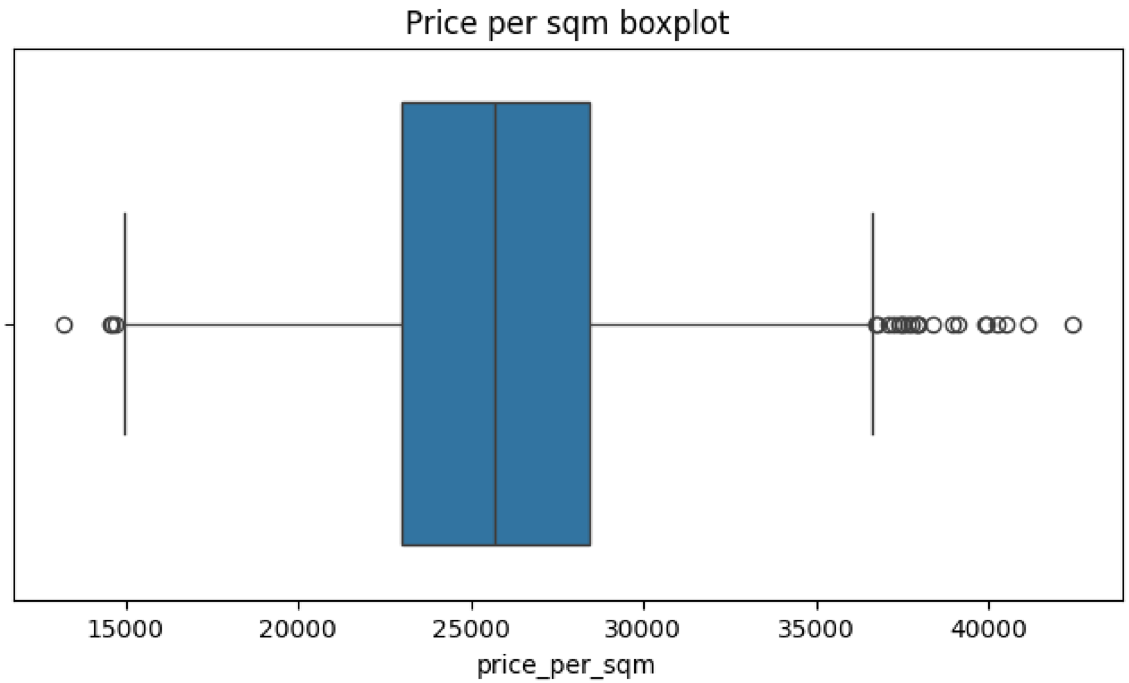
The developed XGBoost pricing model provides **highly accurate predictions** (MAE < ±EGP 150,000) and actionable insights for PropMatch Egypt’s junior agents. The model is reproducible, interpretable, and ready for pilot deployment in New Cairo. Future work could extend to additional apartment types, districts, and incorporate real-time market trends.

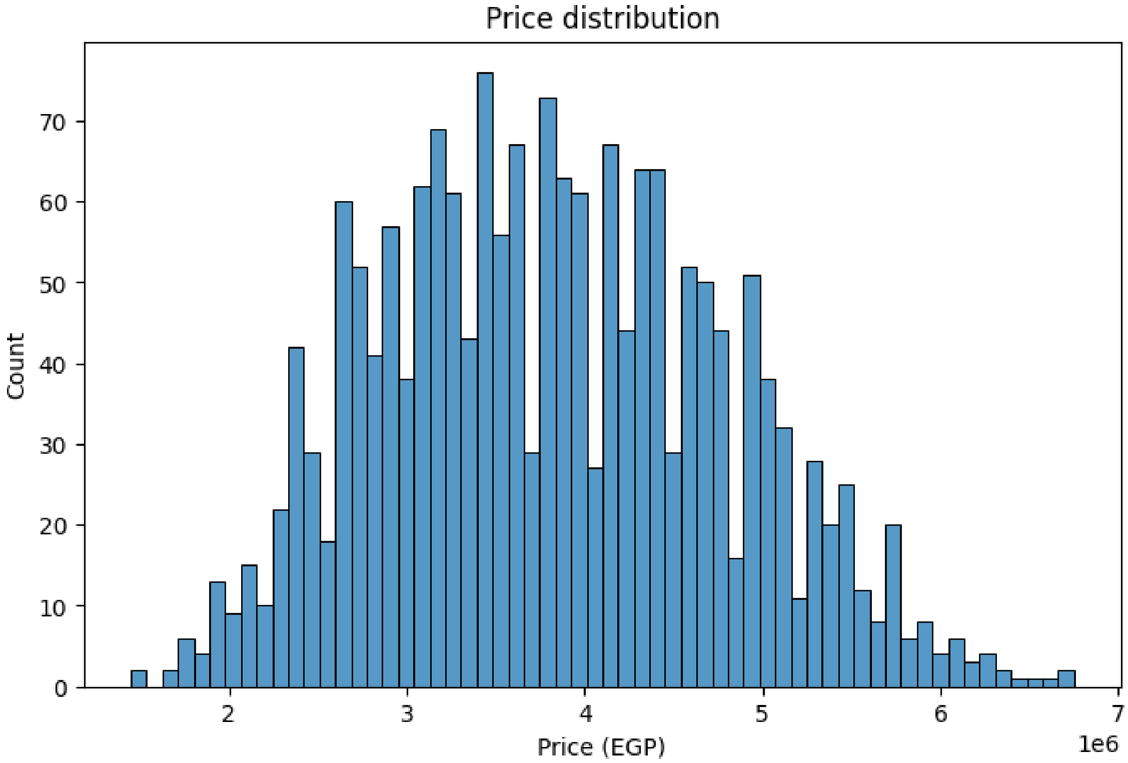


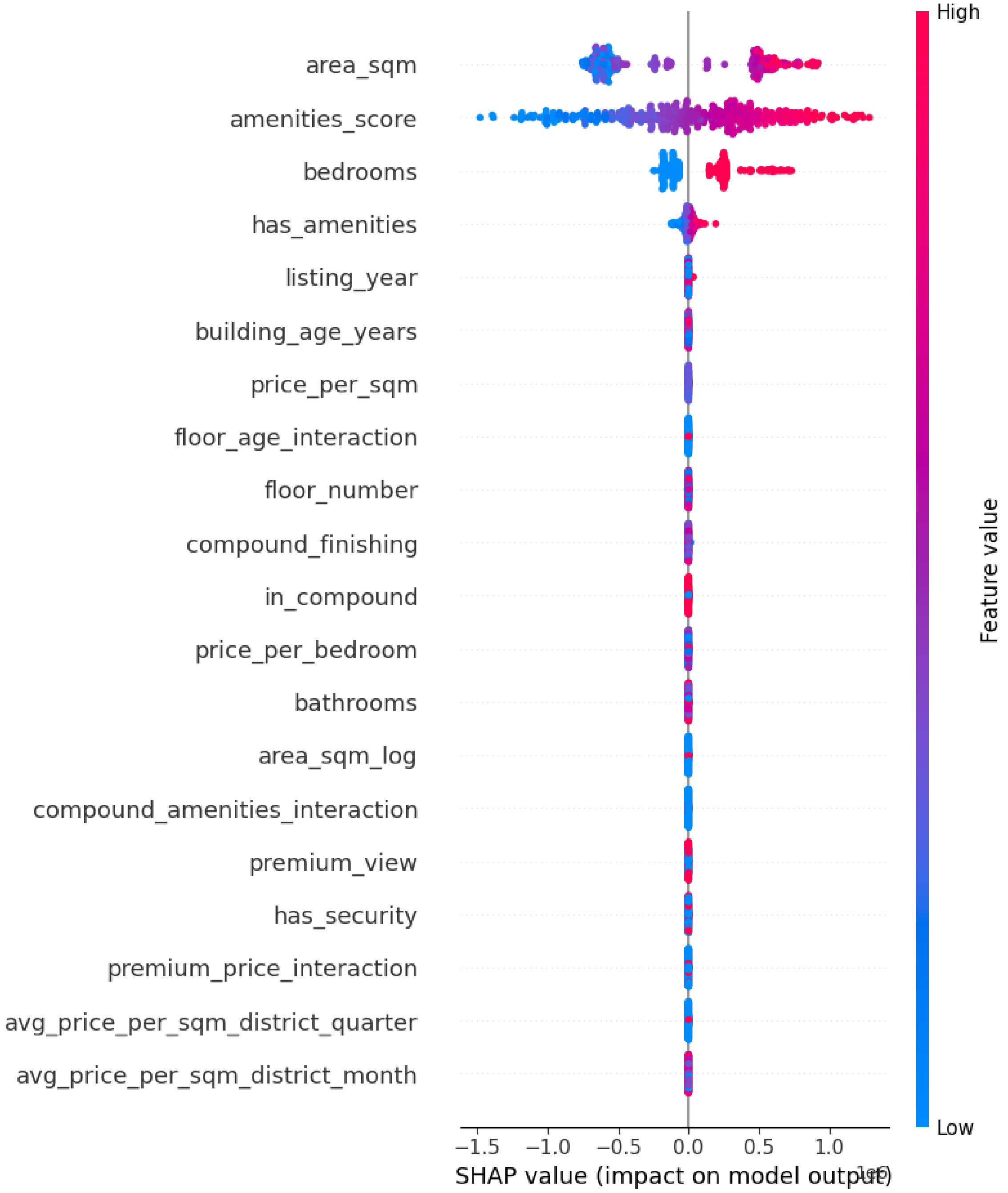
	price_egp	area_sqm	bedrooms	bathrooms	floor_number	building_age_years	distance_to_auc_km	distance_to_mall_km	distance_to_metro_km	has_balcony	has_parking	has_security	has_amenities	days_on_market	is_negotiable	price_per_sqm	listing_month	listing_year
price_egp	1.00	0.81	0.72	0.42	-0.03	-0.16	-0.06	-0.02	0.02	-0.01	0.04	0.06	0.14	-0.00	0.07	0.35	0.00	
area_sqm	0.81	1.00	0.82	0.49	0.02	-0.00	-0.01	0.01	-0.00	-0.00	0.01	-0.02	-0.01	-0.02	0.05	-0.25	0.02	
bedrooms	0.72	0.82	1.00	0.60	0.00	0.00	-0.00	0.00	-0.01	-0.00	0.03	-0.02	-0.02	-0.02	0.04	-0.13	0.02	
bathrooms	0.42	0.49	0.60	1.00	0.00	-0.03	-0.03	-0.01	-0.00	-0.01	0.02	-0.01	-0.01	-0.01	0.01	-0.08	0.01	
floor_number	-0.03	0.02	0.00	0.00	1.00	0.02	0.01	-0.01	0.02	0.00	-0.02	-0.03	-0.04	0.00	-0.03	-0.07	-0.00	
building_age_years	-0.16	-0.00	0.00	-0.03	0.02	1.00	0.00	-0.03	0.02	0.04	0.02	-0.00	0.01	-0.01	0.04	-0.27	0.01	
distance_to_auc_km	-0.06	-0.01	-0.00	-0.03	0.01	0.00	1.00	0.02	-0.01	0.03	0.02	0.01	-0.02	0.02	-0.00	-0.08	-0.02	
distance_to_mall_km	-0.02	0.01	0.00	-0.01	-0.01	-0.03	0.02	1.00	-0.04	-0.03	-0.00	0.02	0.01	0.00	0.01	-0.05	-0.01	
distance_to_metro_km	0.02	-0.00	-0.01	-0.00	0.02	0.02	-0.01	-0.04	1.00	-0.01	-0.02	0.00	0.01	0.03	0.02	0.03	-0.03	
has_balcony	-0.01	-0.00	-0.00	-0.01	0.00	0.04	0.03	-0.03	-0.01	1.00	0.00	-0.01	0.00	-0.03	-0.00	-0.01	0.03	
has_parking	0.04	0.01	0.03	0.02	-0.02	0.02	0.02	-0.00	-0.02	0.00	1.00	0.01	-0.01	0.02	0.03	0.05	-0.02	
has_security	0.06	-0.02	-0.02	-0.01	-0.03	-0.00	0.01	0.02	0.00	-0.01	0.01	1.00	0.04	0.01	-0.03	0.13	-0.01	
has_amenities	0.14	-0.01	-0.02	-0.01	-0.04	0.01	-0.02	0.01	0.01	0.00	-0.01	0.04	1.00	-0.02	0.00	0.25	0.02	
days_on_market	-0.00	-0.02	-0.02	-0.01	0.00	-0.01	0.02	0.00	0.03	-0.03	0.02	0.01	-0.02	1.00	-0.04	0.02	-0.99	
is_negotiable	0.07	0.05	0.04	0.01	-0.03	0.04	-0.00	0.01	0.02	-0.00	0.03	-0.03	0.00	-0.04	1.00	0.04	0.03	
price_per_sqm	0.35	-0.25	-0.13	-0.08	-0.07	-0.27	-0.08	-0.05	0.03	-0.01	0.05	0.13	0.25	0.02	0.04	1.00	-0.02	
listing_month	0.00	0.02	0.02	0.01	-0.00	0.01	-0.02	-0.01	-0.03	0.03	-0.02	-0.01	0.02	-0.99	0.03	-0.02	1.00	
listing_year																		


Top districts (counts)
600
500
400
300
200
100
0
Fifth Settlement
Rehab City
Katameya
Madinaty
New Cairo (Other)
district


Finishing type counts
700
600
500
400
300
200
100
0
Lux
Semi-Finished
Super Lux
Unfinished
finishing_type


Price per sqm boxplot
15000
20000
25000
30000
35000
40000
price_per_sqm


Price distribution
0
10
20
30
40
50
60
70
Count
2
3
4
5
6
7
1e6
Price (EGP)


area_sqm
amenities_score
bedrooms
has_amenities
listing_year
building_age_years
price_per_sqm
floor_age_interaction
floor_number
compound_finishing
in_compound
price_per_bedroom
bathrooms
area_sqm_log
compound_amenities_interaction
premium_view
has_security
premium_price_interaction
avg_price_per_sqm_district_quarter
avg_price_per_sqm_district_month
−1.5
−1.0
−0.5
0.0
0.5
1.0
1e6
SHAP value (impact on model output)
High
Low
Feature value