Predicting Apartment Prices from Arabic Online Listings in Saudi Arabia: Gradient Boosting with Structural, Locational and Text-Derived Features

Abdullah M. Baqasah
Department of Information Technology, College of Computers and Information Technology,
Taif University, 21974, Taif, Saudi Arabia

***Abstract*—Reliable price estimation supports housing policy and market transparency in Saudi Arabia, yet the online listings that form the main public source of price data are noisy, heavily duplicated, and describe many attributes only in Arabic free text. This study develops and evaluates a reproducible pipeline for predicting apartment prices from 19,971 for-sale listings scraped from the Aqar.fm platform. A documented cleaning procedure removed 25.2% of the raw records, chiefly repeated advertisements, and retained 14,934 listings concentrated in Jeddah, Riyadh, and Dammam. Fifty binary features were derived from the Arabic descriptions with spelling-variant-aware keyword rules. Multiple linear regression, ridge regression, random forest, LightGBM, and XGBoost were compared on log price under five-fold grouped cross-validation, which keeps listings sharing a description group out of the opposite partition, with hyperparameters selected over widened grids. LightGBM was selected by the cross-validation criterion and outperformed the linear baselines on every test-set metric at both the log-price and price scale: a mean absolute error of 69,985 SAR (95% bootstrap CI 65,005–75,683) and MAPE of 9.3%, against 93,161 SAR and 12.9% for multiple regression. Location features accounted for most of the gain over structural attributes alone (MAE 134,218 to 73,645 SAR), Arabic text describing the building added a further 4.9% reduction (to 70,016 SAR), and neighbourhood-service text added no further reduction in error. A model trained on only two of the three cities performed poorly when tested on the held-out city (MAPE 20.2% versus 7.0% under a random split; R^2 −0.28), a result attributable mainly to location-identity encodings that are undefined for an unseen city rather than to a general failure of transfer. Permutation importance, grouped permutation importance, and TreeSHAP analyses are reported to explain the selected model's predictions, which are interpreted as predictive associations rather than causal effects.**

Keywords*—House price prediction; Saudi Arabia; Arabic text features; LightGBM; XGBoost; random forest; SHAP; data quality; real estate listings

I. Introduction

Housing occupies a central place in Saudi economic and social policy. Vision 2030 sets a target of raising the homeownership rate to 70% by 2030 [1], and the share of owner-occupied housing has been reported to shift from a continuous decline to accelerated growth after the programme entered into force [2]. Access nevertheless remains constrained. Housing experts rank the high price of residential land and the high cost of construction among the principal obstacles to affordable housing [3], and reviews of the sector identify housing shortage, cost, demographic change and land scarcity as persistent problems [4]. In such a market, transparent and accurate price estimates assist buyers in judging offers, lenders in assessing collateral and planners in monitoring market conditions.

Machine learning (ML) has become a standard instrument for property valuation. A survey of 70 studies observed that the data sets employed are often small and that simple methods predominate, and it called for larger digital data sources and for models whose predictions can be explained [5]. Tree-based ensembles frequently perform best: random forests and gradient boosting outperformed support vector machines on about 40,000 Hong Kong transactions [6], and boosted trees were the strongest candidates among the models compared for Taiwan [7]. The evidence is nevertheless not uniform. Least-squares support vector regression performed best on Taiwanese transaction data [8], and support vector regression outperformed gradient boosting on data from Amman [9]. No algorithm can therefore be presumed superior without a direct comparison on the data of interest.

Several studies apply ML to Saudi real-estate data. Artificial neural networks were trained on Aqar.fm listings from Riyadh, Jeddah, Dammam and Al-Khobar [10]; XGBoost was compared with ridge regression and CatBoost on Aqar.fm data [11]; random forests outperformed decision trees and linear regression on 59,846 transaction records [12] and on land parcels in northern Riyadh [13]; and random forests also led a comparison with hedonic regression and a neural network on 2,997 Riyadh units [14]. Three aspects of this literature warrant further attention. First, Aqar.fm-type listings describe many property attributes (private parking, a driver's room, central air conditioning, nearby services) exclusively in Arabic free text, yet none of these studies report using that text as model input, although descriptions have improved valuation accuracy in other languages [15], [16]. Second, scraped listings contain repeated and near-repeated advertisements, and the extent of this problem and its consequences for train–test leakage are rarely quantified, even though leakage is a recognised cause of over-optimistic results across ML-based science [17]. Third, generalisation to a city that is absent from the training data is seldom examined, although prices differ considerably between Saudi cities. A recent systematic review of real-estate valuation likewise identified the geographic isolation of samples, the reliance on structured rather than unstructured data, and limited interpretability as principal shortcomings [18].

This study addresses these points using 19,971 Aqar.fm apartment listings, through documented data preparation, a comparison of two classical and three tree-based regression models, and a structured error analysis. The contributions are as follows:

- A documented, reproducible data-quality pipeline for Arabic listing data, in which every cleaning step is reported together with the number of records it removes;
- A set of 50 Arabic keyword features, written to accommodate spelling variants and dialect words, together with their prevalence;
- A leakage-aware evaluation, comprising a split and cross-validation grouped by listing description and a leave-one-city-out test;
- An ablation that separates the contributions of structural, locational, building-text and neighbourhood-text features, complemented by permutation importance and SHAP values;
- Regression-appropriate metrics (MAE, RMSE, MAPE, R^2) reported on the SAR scale.

The remainder of the study is organised as follows. Section II situates the study within the literature on machine learning for property valuation, with particular attention to prior work on house price prediction in Saudi Arabia. The data and methods are then set out in Section III, covering the cleaning of the raw listings, the construction of the Arabic text features, and the tuning of the five compared models. Section IV presents the results, in turn reporting model performance, the ablation study, and the error analysis, after which Section V interprets these findings in light of the literature and notes the limitations of the study. The study closes with Section VI, which summarises the contribution and points to directions for future work.

II. Related Work

A. Machine Learning for Property Valuation

Hedonic regression has long been the standard tool for explaining property prices, but it presupposes a fixed functional form. ML models relax this assumption. XGBoost reached an accuracy of 84.1% against 42% for hedonic regression when both used the same 13 inputs [19], and it achieved the highest accuracy at every sample size in a study of Japanese apartment rents [20]. Tree ensembles also outperformed linear models on 428,000 Adelaide transactions [21], and random forests were the most precise method in an analysis of asking prices in Alicante, whereas hedonic models were less precise but more robust [22]. XGBoost [23] and LightGBM [24] are the gradient-boosting systems most frequently applied to tabular problems of this kind, and a systematic comparison found that XGBoost outperformed several recent deep-learning models on tabular data while requiring less tuning [25]. A survey by the authors of the Kuala Lumpur study concluded that locational and structural attributes are the most prominent determinants of house prices [26].

B. Saudi and Gulf Studies

Table I summarises the Saudi and Gulf studies reviewed. Because data sets, targets and metrics differ, the figures are not comparable across studies or with the present results.

TABLE I. Saudi and Gulf Studies of Real-Estate Price Prediction

Study	Data	Models	Reported result
Alzain et al. [10]	Aqar-sourced listings via Kaggle; 4 separate city datasets (Riyadh 959, Jeddah 430, Dammam 656, Al-Khobar 1,673), Jan. 2021	ANN (separate model per city; 70/30 split; min-max normalisation only)	Testing-phase R = 100% (Dammam), 96.5% (Jeddah), 95.0% (Al-Khobar), 84.0% (Riyadh); RMSE on min-max-scaled price as low as 0.0011 (Dammam); no duplicate handling, text features or cross-city test described
Alsulamy [11]	Aqar.fm website; Z-score-normalised and log-transformed	XGBoost, ridge, CatBoost	XGBoost best: R^2 0.98 (training), 0.77 (validation), 0.82 (test)
Alshammari [12]	59,846 transaction records, 13 columns reduced to 6; Ministry of Justice open data, all Saudi cities, 3 months (Jan.–Apr. 2022)	RF, decision tree, linear regression (70/30 split)	RF best: R^2 0.65, MAE 81,000 SAR (DT: R^2 0.30, MAE 502,000; LR: R^2 0.06, MAE 670,000); price-per-m² and area dominate feature importance; no duplicate handling, text features or cross-city test described
Louati et al. [13]	5,946 land parcels, northern Riyadh; Ministry of Justice "daal" portal, scraped Jan. 2018–Dec. 2020; mode imputation for missing values, outliers excluded	Decision tree, RF, linear regression (70/30 split; grid-searched hyper-parameters)	RF best: MAE 233,182 SAR, MSE 1.95×10^{11}, median SE 110,526 (DT: MAE 282,774; LR: MAE 306,315); authors explicitly flag the single-region (north Riyadh) scope as a limit on generalisation to other areas
Ezzeldin et al. [14]	2,997 units, five Riyadh districts, Q3 2024; 12 features, 70/30 split	RF, ANN, hedonic regression	RF best: R^2 0.9065, RMSE 0.1911, MAE 0.1353 (ANN $R^2 \approx 0.90$; hedonic regression $R^2 \approx 0.80$); leading RF predictors were north/west-Riyadh indicators, building type, entrances and size
Chikhmous et al. [27]	592 Riyadh apartments from online sale advertisements (2019)	Hedonic regression (linear and semi-log)	Semi-log form more explanatory; floor area, furnishings, yards and elevators positive, building age and floor level negative (internal attributes); proximity to expressways, colleges, hospitals and future metro stations positive, downtown proximity unexpectedly negative, Olaya district strongest positive CBD effect (external, GIS-derived attributes)
Khrais and Shidwan [28]	4,600 listings from Bayut and Property Finder	ANN (12 features; location limited to distance from the capital; 80/10/10 split)	R = 0.658, R^2 = 0.433; weakest for high-priced homes. District-level location could not be used because the source listings lacked it, and single-region sampling is flagged as a limitation

Albursha id et al. [29]	Land (910), apartments (2,539), villas (482); 3 Bahrain listing sites scraped with ParseHub/Octoparse, geocoded with Geopy/ArcGIS Pro	Linear regression, RF, XGBoost (separate model per property type)	XGBoost best for apartments (R^2 0.82, MAE 12,856 BHD) and land (R^2 0.98, MAE 7,646); for villas all three models performed poorly, and XGBoost was worst (R^2 −0.43), which the authors attribute to data scarcity for that property type
Khunji et al. [30]	3,084 Bahrain rental listings; 14 attributes, 3 websites, March 2024; 80/20 split	Linear regression, RF, SVM	RF best: MAE 0.1385, RMSE 0.2394 (linear regression MAE 0.1746, RMSE 0.2935; SVM MAE 0.1681, RMSE 0.2976)

Two regularities emerge. Random forest is the reported winner in four of the Saudi and Gulf studies that include it in a comparison [12], [13], [14], [30], whereas the one Saudi study that compares XGBoost on Aqar.fm data reports it as best [11]. This is consistent with the general superiority of non-linear ensembles over linear baselines, but it leaves open which ensemble is preferable for Aqar.fm listings: the Bahrain pipeline, for instance, has XGBoost beating random forest for apartments and land but performing worst of the three models for villas, with a negative R^2. None of these studies uses description text, documents the treatment of repeated advertisements, or tests generalisation to a city excluded from training. The Khrais and Shidwan ANN is a direct illustration: its method section [28] shows that district-level location could not be used at all because the scraped source (Bayut and Property Finder listings) did not include it, a limitation the authors state explicitly rather than merely omit. The Riyadh land-parcel study is a second: its authors restrict their own conclusions to the single region studied, noting that other areas may need different features. The hedonic study of Chikhmous et al. [27] is informative as a benchmark for signs: a negative association with building age and a positive association with floor area are the expected directions against which the present results can be checked.

C. Scraped Listings and Data Quality

Web-scraped listings are an accepted data source. Nearly 19,000 Vilnius listings scraped during the pandemic were modelled with 15 ML methods, of which XGBoost was the most accurate, albeit by a negligible margin [31]; a Madrid application estimated market prices from public listings after a feature-engineering stage [32]; and a Bahrain pipeline compared linear regression, random forest and XGBoost on scraped listings [29]. Data quality is, however, a recurrent concern. The quality of acquisition has been described as decisive for prediction quality [33]; a review of 124 big-data hedonic studies concluded that uneven data quality reduces appraisal accuracy and has received insufficient attention [34]; and a Berlin rent study explicitly detected offers that appeared on two portals and removed the duplicates before modelling [35]. Scraping additionally raises legal, ethical and statistical issues, including incompleteness, inconsistency and bias [36].

D. Text as a Source of Price Information

Textual descriptions contain information that structured attributes omit. Including text from multiple-listing-service remarks reduced pricing error by more than 25% in an Atlanta hedonic model [16], and the "soft" information in descriptions was associated with sale prices and with unobserved quality [37]. A comparison of description-only, numeric-only, and combined inputs found that description text alone was a strong predictor and that the best performance was obtained by combining both feature types [38]. Evidence from Berlin and Los Angeles indicates that adding description representations reduced the MAE by up to 17.09% relative to baselines, with a smaller benefit for low-priced rentals [15]. The benefit is not uniform across statements: some qualitative statements were redundant given quantitative attributes, whereas those describing structural quality carried unique information [39]. A systematic review reports that research on text mining in real estate has grown markedly since 2015 but faces problems of data quality, standardisation and computational efficiency [40]. For Arabic, text mining of classified advertisements for used cars and construction equipment has been shown to improve price prediction significantly relative to structured-only baselines [41], which supports the premise adopted here, although that study concerns a different product category.

E. Arabic Text Processing

Arabic poses specific challenges for text processing, including rich morphology, complex grammar, and several character sets, and the usual pre-processing steps comprise normalisation, tokenisation, stemming, stop-word removal, and noise reduction [42]. Classical, modern standard and dialectal varieties coexist [43]. Gulf and Saudi dialects have received dedicated attention, with a transformer model exceeding 90% macro-F1 on Gulf and Saudi dialect corpora [44] and a language model pretrained on Saudi dialect text [45]. Real-estate advertisements mix Modern Standard Arabic and dialect words (for instance, several spellings for "driver's room" "غرفة سائق" and "maid's room" "غرفة خادمة"), which motivates the spelling-variant rules described in Section III-C. A complete Arabic NLP pipeline is left for future work.

F. Validation, Leakage and Transferability

When related observations are split across training and test sets, performance estimates become optimistic. Leakage of this kind has affected hundreds of papers in 17 fields [17]; repeated observations of the same subject drastically inflated prediction performance in neuroimaging models [46]; and improper splitting inflated test accuracy by 5–30% in an image-classification study [47]. In real-estate data, spatial dependence has the same effect: error estimates from non-spatial resampling were found to be overly optimistic for tree-based hedonic models [48]. Grouped validation, as used here, is a direct response to the first of these problems.

Transferability across geographies has produced mixed evidence. Machine learning models for street-frontage classification and house price regression did not generalise well across geographies [49]. In contrast, hedonic spatial-structure parameters estimated in one Norwegian region gave equivalent predictions in another [50], a Vancouver model could be adapted to Kelowna with careful calibration [51], and a transfer-learning model based on spatial features reached a MAPE as low as 2.08% when moved to another city under complete spatial-feature coverage [52]. The outcome thus appears to depend on which features carry the location signal, a point examined in Section V.

G. Explainable Valuation Models

SHAP values assign each feature an additive contribution to an individual prediction and unify several earlier attribution methods [53]. For tree ensembles, exact values can be computed in polynomial time [54], and the algorithm has been implemented in XGBoost and LightGBM [55]. In valuation, permutation importance and SHAP were combined for tree ensembles compared with multiple regression [56]; SHAP clarified the influence of open-data predictors on Lisbon prices, where the added features reduced the MAE by 8.24% [57]; and SHAP was applied to 1.5 million French transactions [58]. Detailed location information can also raise accuracy considerably: geocoded location features improved some error measures by more than 50% in French cities [59], and optimised accessibility indices improved random-forest accuracy by 21.6% in Madrid [60].

H. A Prior Four-Model Comparison

A comparison of multiple regression, ridge regression, LightGBM and XGBoost on Kuala Lumpur housing listings, using a 70/30 split and grid search, found XGBoost best on MAE, RMSE and adjusted R^2 [61]. That study examined the skewness of price and applied a log transform, inspected correlations, one-hot encoded categorical variables, and applied the model to a fresh sample, noting that errors were largest in high-end locations; these results are used as a benchmark in Sections IV and V.

III. Data and Methods

A. Data Source

Fig. 1 summarises the overall pipeline, from the raw listings to the evaluation reported in Section IV. The data set comprises 19,971 apartment-for-sale listings from Aqar.fm (21 columns, Table II; all textual fields in Arabic). Jeddah (8,641), Riyadh (5,348), and Dammam (3,351) contribute 87% of the records, with the remainder from more than 30 other cities.

TABLE II. Variables in the Raw Data Set

Type	Variables
Metadata (not used)	id, ad_number, title, last_update, views
Structural	rooms, living_rooms, bath_rooms, area (m²), level, age, kitchen, garage, elevator, dimensions
Locational	city, district, front (facing direction), street_width (m)
Text	description (Arabic free text)
Target	price (SAR)

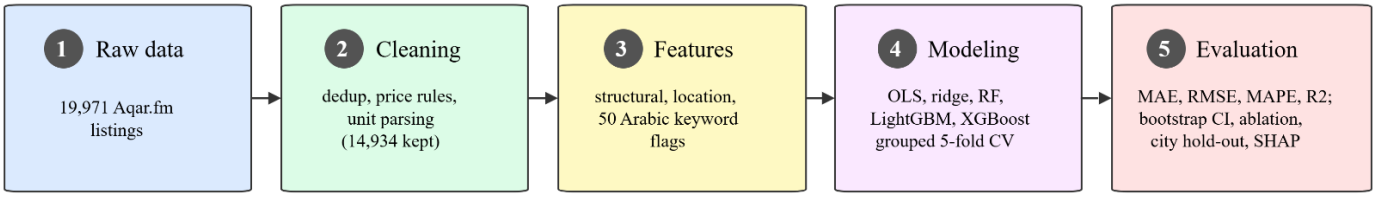

Fig. 1. Study pipeline, from raw Aqar.fm listings to evaluation.

B. Data Cleaning and Preparation

Table III lists every preparation step in the order applied. The first block parses Arabic labels and units into numbers; the second removes unreliable records, such as duplicate advertisements and implausible prices; the third creates encoding and validation helpers used by the models in Section III-D. Each step reports the number of records it affects, so the cumulative effect of cleaning on the sample size can be verified rather than assumed. Codes 0 (level, street width) and −1 (age) denote unknown values in the parsed data, kept distinct from legitimate zero or low values so that imputation (Section III-D) can be applied deliberately rather than silently. Duplicate removal follows the practice of the Berlin study [35], and the emphasis on documenting quality steps responds to the concern of [34].

TABLE III. Data Preparation Steps Applied to the Raw Data

#	Step	Detail	Effect
1	Missing price	Records without a price removed	15 records
2	Price parsing	Currency word and separators removed; 27 prices carry an annual or monthly suffix, but their values (400,000–1,600,000 SAR) lie on the sale-price scale, so they are retained as sale prices	27 values retained
3	Metadata columns	id, title, ad_number, last_update and views dropped	5 columns
4	front	Arabic direction names mapped to N, NE, E, SE, S, SW, W, NW, plus 3S and 4S (three or four street frontages)	10 categories
5	rooms, living_rooms, bath_rooms	Nulls set to 0; top-coded labels “7+” and “5+” set to 7 and 5	integers
6	street_width	Unit letter removed; null set to 0 (unknown)	1.4% unknown
7	level	Ground = 1, first = 2, numeric levels retained, “20+” = 20; null set to 0 (unknown)	11.6% unknown
8	age	“New” = 0, “35+” = 35; null set to −1 (unknown)	8.4% unknown
9	kitchen, garage, elevator	Null treated as absent	binary

10	area	First token of the string parsed as integer m²	20–1,000 m²
11	Text features	50 keyword flags extracted from description (Section III-C); description then dropped	+50 columns
12	Repeated advertisements	First record retained for each ad_number	4,669 records
13	Near-duplicates	Same city, district, area, rooms, bathrooms, price, level, age, street width, and description under a different ad number	315 records
14	Implausible prices	Price < 50,000 SAR removed (e.g. 430 or 700 SAR for 170–270 m² apartments, probably entered in thousands)	8 records
15	Price per m²	Records outside the 0.1%–99.9% quantiles removed (611–33,689 SAR/m²)	30 records
16	Unknown values	Indicators age_missing, level_missing, street_width_missing added; codes replaced by the median of known values	3 columns
17	Rare cities	Cities with fewer than 100 listings pooled as "other" (city_grp)	13 groups
18	dimensions	Dropped (26% missing; list-valued; area already captures size)	1 column
19	Group identifier	Listings with identical descriptions share a group ID used in splitting and cross-validation	14,093 groups

After steps 1 and 12–15, 14,934 of the 19,971 raw records (74.8%) remain.

C. Arabic Text Features

The description field was converted into 50 binary indicators. An indicator equals 1 if any listed phrase occurs in the description (case-insensitive substring matching). Each rule enumerates the spelling variants encountered in listings, such as alternative forms of the feminine ending and of hamza, singular and plural forms, and several dialect words for "driver" and "maid". The flags were divided into 30 building features (general, external, internal, structure and utilities) and 20 neighbourhood-service features (Table IV) so that their contributions could be measured separately. Arabic classifieds have been mined for price signals before [41], and standard Arabic pre-processing steps (normalisation, tokenisation, stemming) are described in [42]; the keyword approach used here is a deliberately simple, auditable alternative to a full NLP pipeline.

TABLE IV. DESCRIPTION-DERIVED BINARY FEATURES

Group	Flags (count)	Example Arabic phrase (meaning)
General (3)	owners_association, directly_by_owner, immediate_emptying	اتحاد ملاك (owners' association)
External (4)	car_parks, shaded_car_parks, backyard, entrances	موقف خاص (private parking)
Internal (8)	driver_room, maid_room, roof, laundry_room, storage_room, balcony, jacuzzi, american_kitchen	غرفة سائق (driver's room)
Structure (3)	water_tank_insulation, thermal_insulation, modern_design	عزل حراري (thermal insulation)
Utilities (12)	cctv, central_gas, central_cooling, central_satellite, independent_meter, separate_water_tank, smart_lock, smart_key, security_man, central_vacuuming, fire_fighting_system, smart_control	تكييف مركزي (central air conditioning)
Neighbourhood services (20)	mosque, park, school, university, health_care, pharmacy, field, airport, mall, playground, coffee, pool, nursery, supermarket, laundry, barber, atm, restaurant, sport, path	مسجد (mosque)

Certain rules were narrowed deliberately. The sports-field flag matches only the plural because the singular also matches a stadium name, and the shopping flag omits the singular "market" because it occurs in the phrase "property market". The flags are keyword indicators and do not resolve negation or context.

D. Transformation and Encoding

1) Target: Models predict log(price), which reduces the pronounced right skew of price. Predictions are exponentiated for the SAR-scale metrics. Back-transformation of log predictions is biased in general, and corrections such as Duan's smearing estimator exist [62]; the bias of ensemble-tree predictions can additionally depend on the transformation [63]. A smearing correction was computed and compared on the held-out test set (the factor is the mean of exp(training residuals), here 1.0153). Applying it changed the test MAE from 69,985 to 71,029 SAR, the RMSE from 166,781 to 167,049 SAR and the price-scale R^2 from 0.7124 to 0.7115 — a small adverse change on every metric here. The reported SAR-scale metrics therefore use direct exponentiation without correction. Under this choice, and assuming approximately symmetric errors in log space, the quantity predicted is the conditional median of price given the inputs rather than a bias-corrected conditional mean; the smearing estimator is reported only as a check on that choice.

2) Categorical variables: City_grp and front are one-hot encoded. The city–district combination (about 450 levels) is encoded with a smoothed mean target encoder [64] (smoothing 20), implemented in pandas/numpy. For a categorical level c with training log-price observations y_i, the encoded value is $(\Sigma\, y_i + 20 \cdot \bar{y}) / (n_c + 20)$, where n_c is the number of training rows at level c, and $\bar{y}$ is the training-fold mean of log price (the prior); an unseen level receives $\bar{y}$ directly. The encoder is fit and applied out-of-fold to avoid a listing leaking into its own encoding: the training data are split into five GroupKFold folds, using the same description-derived groups as the outer train/test

split and cross-validation (Section III-G), rather than a plain shuffled K-fold. Each level's statistics are computed from the four folds a listing is not in, and that encoding is applied to the held-out fold; the fold-wise encoded values are then used for training the downstream model. On the final test set, each level is instead encoded using statistics from the full training set, with no further folding. Grouping the internal folds this way closes a leakage path that a plain shuffled K-fold would leave open. Without it, two near-duplicate listings sharing a description group — and therefore always kept together across the outer split and cross-validation folds — could still land in opposite folds of the encoder's own internal refit, letting one listing's price leak into its near-duplicate's district encoding within the training data. Regularised target encoding has performed best in benchmarks with high-cardinality features [65]. Unseen districts receive the training mean.

3) Scaling: Standardisation is applied to ridge regression only; tree models use raw values.

4) Unknown values: Age, level and street_width carry sentinel codes for unknown values (Table III, step 16) alongside binary missingness indicators. Descriptive statistics and figures in Section IV use sentinels replaced by the median of known values, computed on the full cleaned data set before splitting. For modelling, medians are instead computed from the training fold only, after the train/test split, and applied to both training and test sentinel rows, so test-set values never influence the imputed training values or vice versa. The indicators carry a small amount of independent signal: by permutation importance, level_missing ranked 14th overall (ΔRMSE 0.0008), while age_missing and street_width_missing ranked near zero (Section IV-I).

E. Feature Sets

Four nested feature sets support the ablation: A, structural attributes (area, rooms, living rooms, bathrooms, level, age, kitchen, garage, elevator and the age and level indicators); B = A + location (street width and its indicator, city group, facing direction and the encoded city–district); C = B + the 30 building text features; and D = C + the 20 neighbourhood-service text features.

F. Models, Tuning and Metrics

Five regression models were compared: multiple linear regression, ridge regression, LightGBM[24], XGBoost[23], and random forest. An earlier version of this pipeline could not execute random forest or a wide grid search because scikit-learn's compiled extension modules were blocked in the computing environment first used. Both are included here following a change of environment (Section III-I). Hyperparameters were tuned by grid search over five-fold GroupKFold, with the description group as the grouping variable and the RMSE on log price as the criterion (Table V). The model with the lowest cross-validated RMSE was selected; ties within 0.001 on this criterion were broken using the test-set MAE. The test set was otherwise used for reporting only.

TABLE V. HYPER-PARAMETER SEARCH SPACES

Model	Search space	Fixed
Multiple regression	none	OLS
Ridge regression	$\alpha \in \{0.1, 1, 10, 30, 100, 300, 1000, 3000\}$	standardised inputs
LightGBM	num_leaves ∈ {15, 31, 63, 127}; learning_rate ∈ {0.02, 0.05, 0.1, 0.2}	500 trees
XGBoost	max_depth ∈ {3, 4, 6, 8, 10}; learning_rate ∈ {0.02, 0.05, 0.1, 0.2}	500 trees
Random Forest	n_estimators ∈ {300, 500}; max_depth ∈ {None, 10, 20, 30}; min_samples_leaf ∈ {1, 2, 5}	n/a

Because the target is a continuous price, model accuracy is reported using five regression metrics, computed on the SAR scale unless stated otherwise. The mean absolute error is

$$\mathrm{MAE} = (1/n)\,\Sigma_i|y_i - \hat{y}_i|, \tag{1}$$

The root mean squared error is

$$\mathrm{RMSE} = \sqrt{[(1/n)\,\Sigma_i(y_i - \hat{y}_i)^2]}, \tag{2}$$

The mean absolute percentage error is

$$\mathrm{MAPE} = (100/n)\,\Sigma_i|y_i - \hat{y}_i| / y_i, \tag{3}$$

The coefficient of determination is

$$R^2 = 1 - \Sigma_i(y_i - \hat{y}_i)^2 / \Sigma_i(y_i - \bar{y})^2, \tag{4}$$

And, following Eq. (4), the adjusted coefficient of determination is

$$\text{adjusted } R^2 = 1 - (1 - R^2)(n - 1) / (n - p - 1), \tag{5}$$

where, n is the number of test observations, y_i and $\hat{y}_i$ are respectively the actual and predicted price of listing i, $\bar{y}$ is the mean actual price over the test set, and p is the number of model inputs after encoding. RMSE and R^2 on the log scale are also reported in Section IV. Classification metrics such as AUC, precision, and sensitivity do not apply to this regression task.

G. Validation Design

1) Grouped hold-out. 80% of listings were used for training and 20% for testing, with all listings sharing a description assigned to the same side (random state 42); the test set contains 2,994 listings and the training set 11,940. Splitting related observations across sets inflates performance estimates, as shown by [17], [46], and [47].

2) Grouped cross-validation. Five-fold GroupKFold on the training set for tuning and model selection.

3) Leave-one-city-out. The selected model was trained on Jeddah and Riyadh only and tested on Dammam, and compared with the same model evaluated on Dammam listings from the random test split.

Grouping by description does not remove spatial dependence, which can make random resampling optimistic in real-estate data [48]; the city hold-out is included to expose this limitation.

H. Interpretation

Permutation importance measures the increase in RMSE (log price) when a single input column is shuffled in the test set (five repetitions). SHAP values [53] were computed with the exact tree algorithm [54] implemented in the boosting library itself [55], on a random sample of up to 2,000 test listings, and their additivity was verified against the model output. Both approaches are associational: they describe what the model uses and not what causes price differences [56].

I. Implementation

All processing was performed in Python 3.11 (pandas 3.0, numpy 2.4, scikit-learn 1.9, LightGBM 4.7, XGBoost 2.1 CPU build, matplotlib) in the accompanying notebook, with a fixed random seed (random_state = 42) for the train/test split, all cross-validation folds, the target encoder's internal folds, and every stochastic model. Every table and figure in Section IV is generated by it, and the code and data supporting this study are available from the corresponding author upon reasonable request. As a deployment scope statement, the model is intended to screen or benchmark asking prices for apartments in Jeddah, Riyadh and Dammam (the three cities with sufficient training data; Section IV-H), using inputs available at listing time (structural attributes, location and a free-text description), not to replace formal valuation (Section V-F). Given a typical test-set error of about 9–11% of price (Section IV-F, IV-K), an estimate more than roughly 20–25% from the model's prediction may warrant closer review. The model reflects market conditions in the October–November 2022 scraping window; it was not validated on cities or periods outside this window, and the data would need to be refreshed and the model retrained to track subsequent price movements.

IV. Results

A. Data Quality

Cleaning reduced the 19,971 raw records to 14,934 (Table III). The largest single effect was the removal of 4,669 repeated advertisements (identical ad numbers), which amounts to 23.4% of the raw data. A further 315 near-duplicates (2.1% of the 15,287 records remaining after step 12) shared attributes, price, and description with another advertisement under a different ad number. By comparison, only 15 records lacked a price, eight prices were below 50,000 SAR, and 30 records were price-per-m² outliers. Since 5,969 raw records (including 188 with an empty description) repeat the description of an earlier record, a random split would have placed near-copies on both sides of the train–test boundary; grouped splitting was therefore used.

The 0.1st/99.9th-percentile price-per-m² filter (Table III, step 15) removes 30 records. Their characteristics support reading them as data errors rather than legitimate extremes: 15 have a price per m² of 50–610 SAR/m² (for example 1,000 m² for 50,000 SAR), implausibly cheap for an apartment and consistent with a missing digit in the price, while the other 15 have 33,700–77,600 SAR/m² (for example 190 m² for 9,000,000 SAR), 7–16 times the city median and consistent with an extra digit. Refitting the selected model on the data set built without this filter (otherwise identical; n = 14,964 versus 14,934) confirms the filter's value rather than merely assuming it. Every test-set metric is worse without it (MAE 86,703 against 69,985 SAR; RMSE 394,446 against 166,781 SAR; MAPE 10.6% against 9.3%; R^2 0.260 against 0.712 on the price scale), with the RMSE and price-scale R^2 most affected — consistent with a small number of extreme, erroneous values dominating squared-error metrics.

B. The Cleaned Data Set

Table VI presents the composition by city. Riyadh has the highest median price and price per square metre (800,000 SAR; 4,737 SAR/m²), followed by Jeddah (650,000; 3,510) and Dammam (580,000; 2,973); apartments are smaller in Riyadh and Jeddah (167, 173 m²) than in Dammam (190 m²), so the two rankings differ in detail. Table VII gives descriptive statistics.

TABLE VI. Composition of the Cleaned Data Set By City Group

City group	n	Median price	Median m²	SAR/m²
Jeddah	6,474	650,000	173	3,510
Riyadh	4,007	800,000	167	4,737
Dammam	2,488	580,000	190	2,973
Makkah	275	670,000	192	3,512
Dhahran	241	610,000	180	3,351
Jazan	208	560,000	204	2,725
Madinah	204	700,000	210	3,409
Khobar	202	750,000	199	3,667
Khamis Mushait	151	685,000	280	2,429
Abha	137	700,000	250	2,833
Buraydah	123	530,000	231	2,059
Taif	112	750,000	240	3,058
Other (<100 each)	312	600,000	220	2,673
All	14,934	650,000	180	3,438

C. Distribution of Price

After cleaning, price is markedly right-skewed (skewness 9.22; excess kurtosis 166.4). The mean price (702,565 SAR) exceeds the median (650,000 SAR) by 8%, consistent with a small number of very expensive listings pulling the distribution's upper tail. The logarithmic transformation brings both statistics close to normality (skewness 0.58; excess kurtosis 3.8; Fig. 2), which supports modelling log(price) rather than price directly (Section III-D). The central half of prices lies between 535,250 and 799,000 SAR (1st/99th percentiles 300,000/1,500,000 SAR); the upper tail is thin but extreme (maximum 10,000,000 SAR), more than 14 times the median.

TABLE VII. Descriptive Statistics of the Cleaned Data Set

Variable	Mean	SD	Min	Med.	Max
Price (SAR)	702,565	347,153	100,000	650,000	10,000,000
Area (m²)	189.6	64.8	20	180	1,000
Rooms	4.3	1.1	1	4	7
Living rooms	1.1	0.5	0	1	6
Bathrooms	3.4	0.8	1	3	5
Level*	2.3	1.4	1	2	20
Age, years*	0.9	2.8	0	0	35
Street width, m*	24.2	10.1	1	23	100

*Known values only. Kitchen, garage, and elevator are present in 87.8%, 62.2%, and 81.6% of listings; 78.7% of apartments are new.

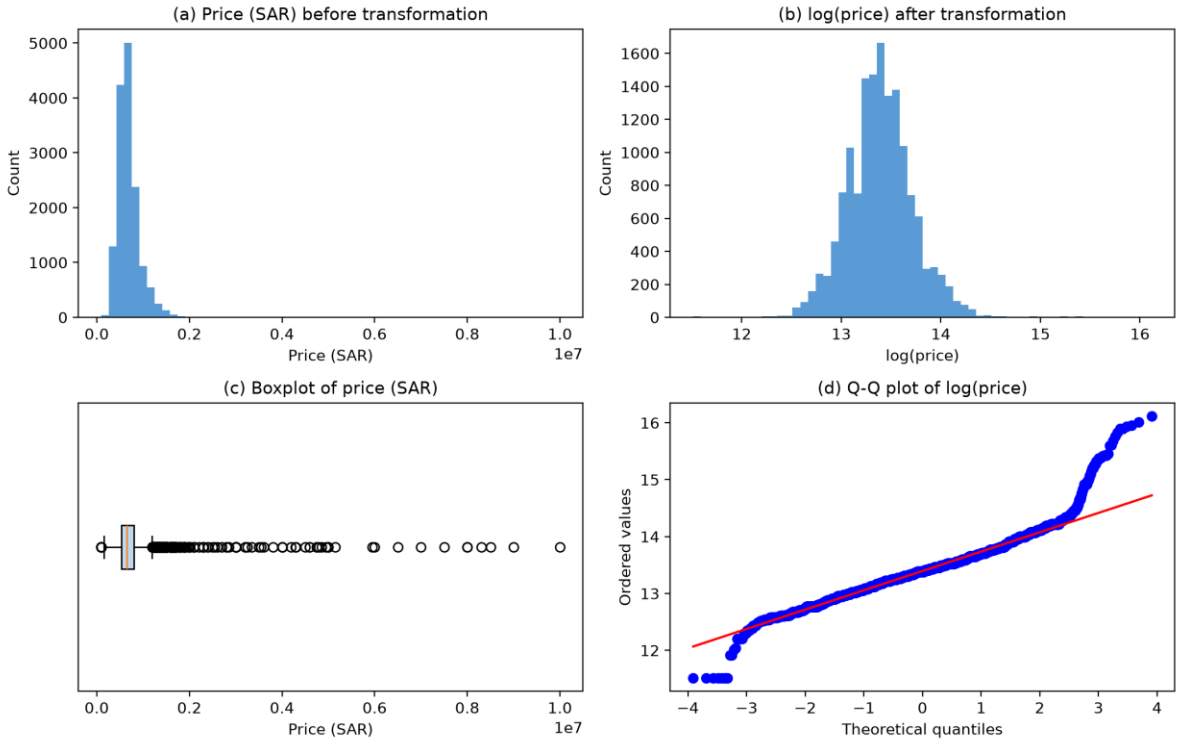

Fig. 2. Distribution of price before and after the logarithmic transformation: (a) Price (SAR) before transformation; (b) Log(price) after transformation; (c) Boxplot of price (SAR); (d) Q-Q plot of log(price) against a normal distribution.

D. Correlations

Among numeric attributes, area has the strongest linear correlation with log price (r = 0.39), followed by bathrooms (0.32), living rooms (0.17), street width (0.12 among known values), and rooms (0.10). Level (0.01), age (−0.06), and the kitchen, elevator, and garage indicators (−0.06 to −0.07) are close to zero (Fig. 3). The negative sign for age is consistent with the hedonic evidence for Riyadh [27]. The modest magnitudes indicate that structural attributes leave most of the price variation unexplained. The small negative correlations for garage and elevator should not be read as a price penalty; they partly reflect city composition, since a garage is reported in 91% of Jeddah listings but only 27% of Riyadh listings, and within Riyadh the correlation between garage and log price is positive (r = 0.31).

E. Text-Derived Features

At least one building feature is mentioned in 71.4% of descriptions and at least one neighbourhood-service feature in 30.8%; the average listing triggers 2.6 of the 50 flags. The most frequent flags are private parking (32.2%), multiple entrances (22.3%), driver's room (20.3%), maid's room (16.5%), nearby parks (15.6%), nearby mosques (15.5%), and independent meters and separate water tanks (14.4% each). Prevalence is highly uneven: 22 of the 50 flags occur in fewer than 1% of listings, and eight occur in fewer than 20 listings (jacuzzi, smart key, security man, central vacuuming, smart control, playground, nursery and barber). Rare flags offer tree models little to learn from.

As a purely descriptive check, the median price per m² of listings with and without each flag occurring in at least 150 listings was compared. The largest positive gaps were for balconies (5,774 versus 3,421 SAR/m²; n = 394), American-style kitchens (4,248 versus 3,421; n = 347) and nearby universities (4,200 versus 3,424; n = 199), and the largest negative gaps were for private roofs (−9.7%), storage rooms (−7.3%) and independent meters (−7.0%). These are unadjusted differences that confound features with city and building type and are not estimates of the value of any feature.

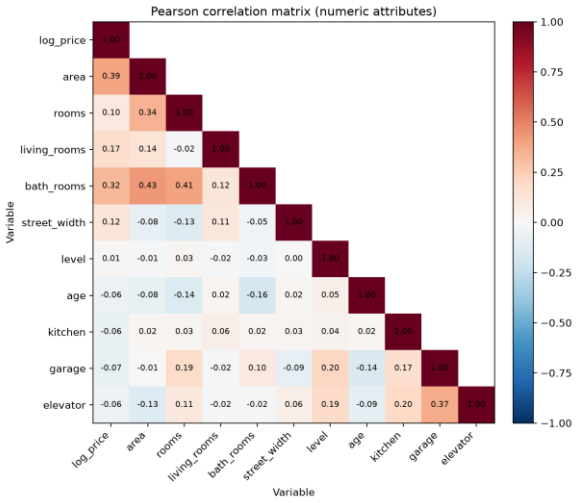

Fig. 3. Pearson correlations among numeric attributes and log price.

F. Model Comparison

Table VIII compares the five models on the held-out test set. LightGBM and XGBoost were statistically tied on the cross-

validation criterion (RMSE on log price, 0.1625 for both, a difference smaller than the grid's resolution); applying the tie-break rule above (lower test-set MAE) selected LightGBM. On the held-out test set, LightGBM yielded an MAE of 69,985 SAR, an RMSE of 166,781 SAR, a MAPE of 9.3% and an R^2 of 0.712 on the price scale (0.801 on the log scale).

LightGBM, the model selected by cross-validation, outperformed both linear baselines on every reported metric at both the log-price and price scale: the test MAE was 69,985 SAR against 93,161 SAR for multiple regression and 93,003 SAR for ridge regression (a 24.9% reduction); the MAPE was 9.3% against 12.9% for both linear models; the price-scale R^2 was 0.712 against 0.596 and 0.591; and the log-scale R^2 was 0.801 against 0.689 for both. Ridge regression (best $\alpha = 300$) was close to ordinary least squares, so regularisation added little for the linear models. XGBoost tied LightGBM on the cross-validation criterion but trailed it on every test-set metric (MAE 71,842 SAR, RMSE 176,599 SAR, R^2 0.678 on the price scale), which is why the tie-break rule selected LightGBM.

Random forest, newly included in this comparison, had a higher cross-validated RMSE (0.1659) than the two boosting models and was not selected, although it recorded the lowest test-set MAE of the five models (68,761 SAR); its RMSE (178,306 SAR) and price-scale R^2 (0.671) were the weakest of the three tree-based models, illustrating that selecting a model by one criterion can rank it differently from a comparison on another metric computed on the same split. The 95% bootstrap confidence intervals in Table VIII (2,000 resamples) show that the MAE intervals of the three tree-based models overlap (LightGBM 65,005–75,683; XGBoost 66,637–78,061; random forest 63,333–74,961 SAR) but are clearly separated from both linear models (87,194–99,858 SAR): the advantage of tree-based models over linear regression is well supported by this evidence, while the ranking among the three tree-based models is not resolved by a single split and should be read as approximate.

TABLE VIII. MODEL COMPARISON WITH ALL FEATURES (SET D) ON THE HELD-OUT TEST SET

Model	CV RMSE (log) [a]	MAE (SAR) [b]	RMSE (SAR) [b]	MAPE (%) [b]	R^2 (price) [b]	RMSE (log)	R^2 (log)	Adj. R^2 (log)
LightGBM	0.1625	69,985 [65,005–75,683]	166,781	9.3	0.712	0.149	0.801	0.795
XGBoost	0.1625	71,842 [66,637–78,061]	176,599	9.6	0.678	0.156	0.784	0.777
Random Forest	0.1659	68,761 [63,333–74,961]	178,306	9.1	0.671	0.157	0.780	0.774
Ridge	0.1994	93,003 [87,194–99,762]	198,812	12.9	0.591	0.187	0.690	0.680
Multiple regression	0.1995	93,161 [87,315–99,858]	197,615	12.9	0.596	0.187	0.689	0.680

Models are listed in increasing order of cross-validated RMSE on log price.

Computed after converting the model's log-price predictions back to SAR; RMSE (log), R^2 (log), and Adj. R^2 (log) are computed directly on log price, without conversion.

Bracketed values after MAE are 95% bootstrap confidence intervals (2,000 resamples of the test set, percentile method).

G. Contribution of Feature Groups

Table IX summarises the ablation across the four nested feature sets, re-estimated with LightGBM (the selected model) under the corrected, grouped-fold target encoder and train-only median imputation (Section III). The ± figure after each MAE is the half-width of its 95% bootstrap confidence interval (2,000 resamples). From set A to set B, the test MAE changed by −60,573 SAR and the price-scale R^2 by +0.407; from set B to set C by −3,629 SAR and +0.038; from set C to set D by −31 SAR and −0.018. Overall, the test MAE moved from 134,218 SAR (structural attributes only) to 69,985 SAR (all features), with nearly all of this gain realised by the time location is added in set B. As a lightweight alternative to the 50 keyword flags, a TF-IDF representation of the description (200 terms, min_df = 5, reduced by truncated SVD to 50 components) was substituted for the building- and neighbourhood-text columns of set D and fit with the same model class and hyper-parameters (row E). It performed slightly worse than the keyword flags on every metric (MAE 72,277 against 69,985 SAR; R^2 0.689 against 0.712). This suggests that the hand-written rules are not an obvious weak point relative to this alternative, despite their simplicity and their inability to represent negation, and since the two approaches are close in accuracy, the interpretability of named keyword flags over opaque SVD components is a practical reason to prefer them here.

Location accounted for most of the improvement (Fig. 4): it lowered the test MAE by 60,573 SAR (45% of the set-A error and 98% of the total reduction from A to D) and raised the log-scale R^2 from 0.415 to 0.786. Building-level text features gave a further, smaller gain (MAE 73,645 to 70,016 SAR, a 4.9% reduction; log-scale R^2 0.786 to 0.799). Neighbourhood-service text features moved the price-scale metrics negligibly, and in the opposite direction to the MAE change (MAE 70,016 to 69,985 SAR; RMSE 161,492 to 166,781 SAR; log-scale R^2 0.799 to 0.801). The bootstrap confidence intervals for sets C and D overlap almost completely (Table IX), so this last step is better read as a null result than as a genuine improvement or deterioration. Grouping train and test by identical description text, used throughout this study, keeps an exact repost out of the opposite partition. It does not, however, guarantee independence between near-identical descriptions — an edited repost, or two listings for different units in the same building that copy the same boilerplate text — nor between listings in the same building or district that use different wording. Test-set metrics should therefore be read as an upper bound on real-world accuracy, and the gap is likely largest for the text-derived features, since these are the features most directly tied to the description text used for grouping.

TABLE IX. ABLATION WITH THE SELECTED MODEL (TEST SET; MAE AND RMSE IN SAR, MAPE IN %)

Feature set	Inputs	MAE	RMSE	MAPE	R^2	R^2 log
A. Structural	11	134,218 ± 8,086	262,922	18.2	0.285	0.415
B. + Location	37	73,645 ± 5,481	172,597	9.8	0.692	0.786
C. + Building text	67	70,016 ± 5,225	161,492	9.4	0.730	0.799
D. + Neighbourhood text	87	69,985 ± 5,339	166,781	9.3	0.712	0.801
E. TF-IDF text (replaces C+D) c	87	72,277	173,460	9.6	0.689	0.792

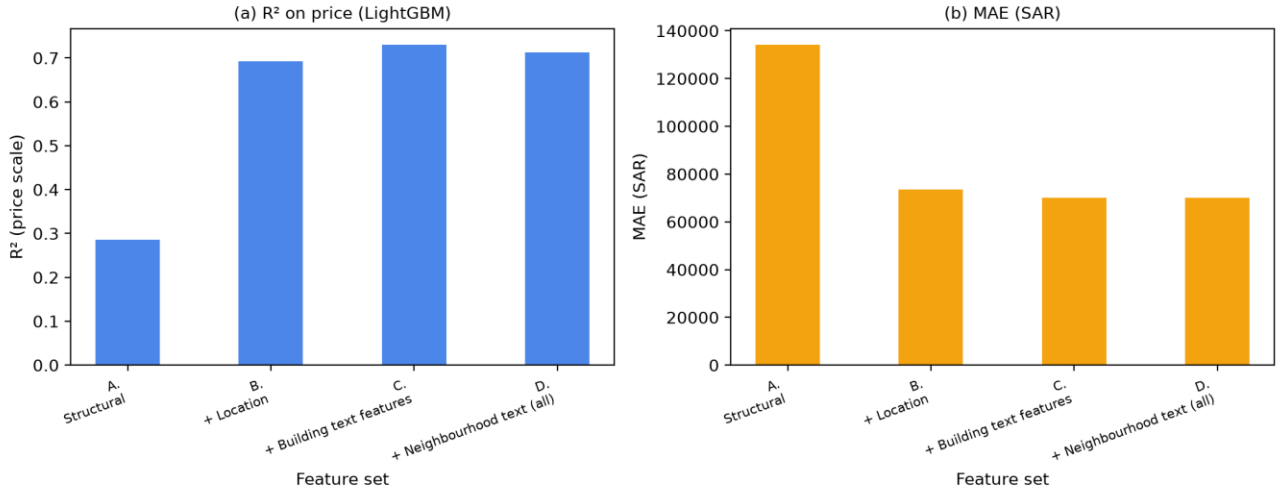

Fig. 4. Effect of adding feature groups on test accuracy: (a) R^2 on the price scale; (b) Mean absolute error (SAR).

H. Generalisation to an Unseen City

Table X reports accuracy on Dammam listings with and without Dammam in the training data, using the selected LightGBM model and the corrected, grouped-fold target encoder. Excluding Dammam from the training data yielded an MAE of 115,748 SAR, a MAPE of 20.2% and an R^2 of −0.28 on Dammam listings, against an MAE of 39,943 SAR, a MAPE of 7.0% and an R^2 of 0.66 when Dammam was represented in the training data.

TABLE X. ACCURACY ON DAMMAM LISTINGS WITH AND WITHOUT DAMMAM IN TRAINING

Training data	MAE (SAR)	MAPE (%)	R^2
WITHOUT Dammam	115,748	20.2	-0.28
WITH Dammam (random split)	39,943	7.0	0.66

The model did not transfer well to the unseen city: without Dammam in training, R^2 was −0.28 on the price scale, worse than predicting the Dammam mean, with a MAPE of 20.2% against 7.0% under the random split. The main structural reason is that two of the model's three leading inputs, the city and city-district target encodings (Section IV-I), are identity features: every level of the city-district encoding is a specific place name. A district actually located in Dammam is therefore categorically unseen by a model trained only on Jeddah and Riyadh, so the encoder can only fall back on its global prior for every Dammam listing. This is a property of the encoding rather than evidence that location information in general cannot transfer across Saudi cities. Training on only two cities cannot by itself distinguish a transferable location signal, which this encoding does not expose, from a genuinely non-transferable one. A secondary factor is that Dammam prices are lower than in the training cities (median 580,000 SAR against 650,000 in Jeddah and 800,000 in Riyadh); a model with no Dammam-specific identity information has no basis for learning this offset. The "with Dammam" figures are themselves optimistic, since listings from the same district or building can appear in both training and test data with different descriptions (Section IV-G).

I. Feature Importance

By permutation importance (ten repetitions), the top five inputs were city_district (0.1824 ± 0.0028), area (0.1252 ± 0.0025), city_grp (0.0478 ± 0.0019), age (0.0056 ± 0.0008), and rooms (0.0055 ± 0.0012).

Table XI and Fig. 5 report permutation importance (mean increase in RMSE on log price, ± one standard deviation across ten repetitions) for the top ten and top twenty inputs. Location and size dominated: the district encoding (0.1824) and area (0.1252) were far ahead of the rest, followed by city group (0.0478), age (0.0056) and rooms (0.0055); their standard deviations (0.0003–0.0028) are an order of magnitude smaller than the gaps between them, so this ranking is stable under repeated shuffling. Text flags were individually minor — driver_room (0.0015), maid_room (0.0011), independent_meter (0.0008) and level_missing (0.0008) ranked highest among them — and several of the 50 flags had importance indistinguishable from zero once their standard deviation is taken into account. Individually shuffling one of several correlated columns (for example, city, district and street width)

can understate its true contribution. To check this, a grouped permutation test additionally shuffled four blocks of related columns jointly: location (city, district, front, street width), structural attributes, building-text flags and neighbourhood-service flags. The grouped increase in RMSE (log price) was 0.224 ± 0.003 for location and 0.165 ± 0.004 for structural attributes, both above the sum of their members' individual scores — confirming that within-group correlation depresses single-column importance for these two groups. Building text and neighbourhood-service text scored only 0.0081 ± 0.0006 and 0.0025 ± 0.0005 respectively, remaining small even when shuffled as a block. This indicates that the small individual-flag importances in Table XI are not primarily an artefact of correlation among the text flags themselves, although correlation among the location columns does inflate the apparent gap between location and the other groups when importance is read column by column.

TABLE XI. TOP TEN INPUTS BY PERMUTATION IMPORTANCE (TEST SET)

Rank	Feature	Δ RMSE (log)
1	city_district	0.1824 ± 0.0028
2	area	0.1252 ± 0.0025
3	city_grp	0.0478 ± 0.0019
4	age	0.0056 ± 0.0008
5	rooms	0.0055 ± 0.0012
6	street_width	0.0042 ± 0.0005
7	bath_rooms	0.0033 ± 0.0004
8	front	0.0020 ± 0.0003
9	level	0.0020 ± 0.0004
10	garage	0.0020 ± 0.0003

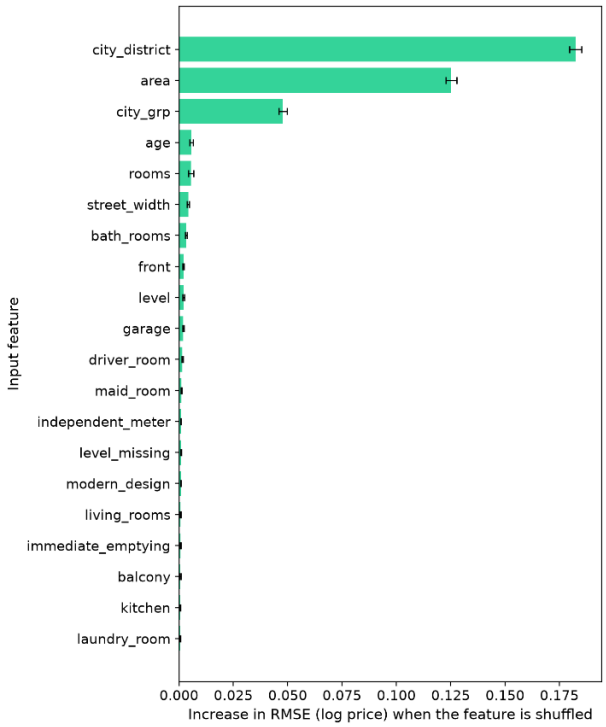

Fig. 5. Permutation importance of the 20 highest-ranked inputs.

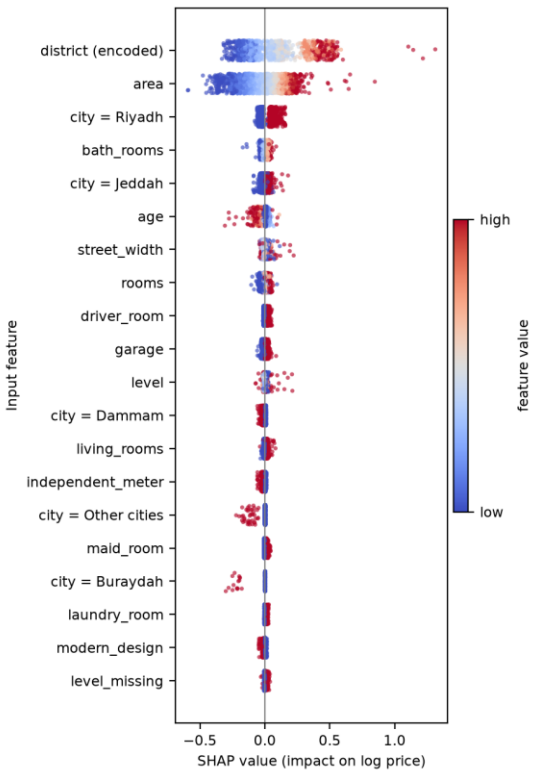

Fig. 6. SHAP summary plot for the selected model: features ranked by mean absolute SHAP value; colour indicates the feature value.

J. SHAP Analysis

SHAP values were computed for the LightGBM model on a random sample of up to 2,000 test listings using its native tree-SHAP implementation. The five leading inputs by mean absolute SHAP value were city-district encoding (0.167), area (0.133), city = Riyadh (0.035), bath_rooms (0.016), and city = Jeddah (0.016). Grouped shares of total SHAP value were Location 51.1%, Structural 38.5%, Building text 8.4%, and Neighbourhood-service text 2.0%, closely matching the permutation-importance ranking above. The summary plot (Fig. 6) shows the expected directions: higher district values, larger area, and more bathrooms/rooms raise predictions, while older buildings lower them; Riyadh and Jeddah indicators raise predictions, Dammam and other cities lower them; a driver's room raises predictions, while independent meters lower them. As with the permutation results, these are predictive associations, not causal effects — amenity flags correlate with price partly because they co-occur with newer, larger, or better-located properties, whose attributes already carry most of the model's signal.

K. Error Analysis

Beyond the absolute error reported previously, Tables XII and XIII also report the mean signed error (predicted minus actual), which shows whether a group's error is systematic rather than symmetric noise. By price quintile (Table XIII), the model is not unbiased: it overpredicts the cheapest quintile by 33,627 SAR on average (+8.4% of actual price) and underpredicts the most expensive quintile by 81,408 SAR (−5.2%), with the three middle quintiles close to zero bias

(−1.3% to +2.4%). This is a regression-to-the-mean pattern consistent with shrinkage toward the conditional mean under squared-error training, and it means the U-shaped MAPE noted below is not purely symmetric variance but partly a directional bias at both ends of the price distribution.

TABLE XII. TEST-SET ERROR BY CITY GROUP

City group	n	MAE (SAR)	MAPE (%)	Mean signed error (SAR)	Mean signed error (%)
Other (pooled)	53	190,511	18.7	−94,897	2.6
Abha	29	82,369	13.0	40,277	7.6
Khobar	36	102,576	12.3	−48,329	−4.6
Dammam	526	39,944	7.0	1,449	1.2
Riyadh	837	94,152	11.1	−14,567	0.1
Taif	23	87,797	13.3	−15,489	4.5
Dhahran	50	51,221	8.8	−12,927	−0.6
Madinah	41	73,607	12.4	27,692	5.9
Buraydah	21	56,741	10.1	−39,514	−6.8
Jazan	46	38,153	6.8	8,489	2.4
Jeddah	1249	61,289	8.3	−5,572	1.0
Khamis Mushait	25	50,316	8.2	4,641	2.3
Makkah	58	89,436	13.5	−6,498	1.2

The error also differed by city (Table XII): the MAPE was 11.1% in Riyadh (MAE 94,152 SAR, mean signed error −14,567 SAR), 8.3% in Jeddah (MAE 61,289 SAR, −5,572 SAR) and 7.0% in Dammam (MAE 39,944 SAR, essentially unbiased at +1,449 SAR). Riyadh and Jeddah are both predicted slightly low on average, consistent with the top-quintile underprediction above, since both cities contain most of the highest-priced listings. The small city groups have only 21–58 test listings each, so their error rates (for example, 13.5% for Makkah and 18.7% for the pooled small cities) are imprecise and their signed-error estimates more so. The top price quintile had by far the largest absolute error (MAE 158,238 SAR against 44,204–64,040 SAR in quintiles 2–4), as the Kuala Lumpur study found for high-end locations [61]. The relative error (MAPE) was U-shaped, highest in the cheapest quintile (11.8%) and in the top quintile (11.6%). The signed error shows that this U-shape reflects two different mechanisms: over-prediction of cheap listings and under-prediction of expensive ones, as shown in Fig. 7.

TABLE XIII. TEST-SET ERROR BY PRICE QUINTILE

Price quintile	n	MAE (SAR)	MAPE (%)	Mean signed error (SAR)	Mean signed error (%)
Q1 lowest	647	48,242	11.8	33,627	8.4
Q2	585	44,204	7.9	13,016	2.4
Q3	651	48,940	7.4	−8,276	−1.3
Q4	586	64,040	8.2	−9,902	−1.2
Q5 highest	525	158,238	11.6	−81,408	−5.2

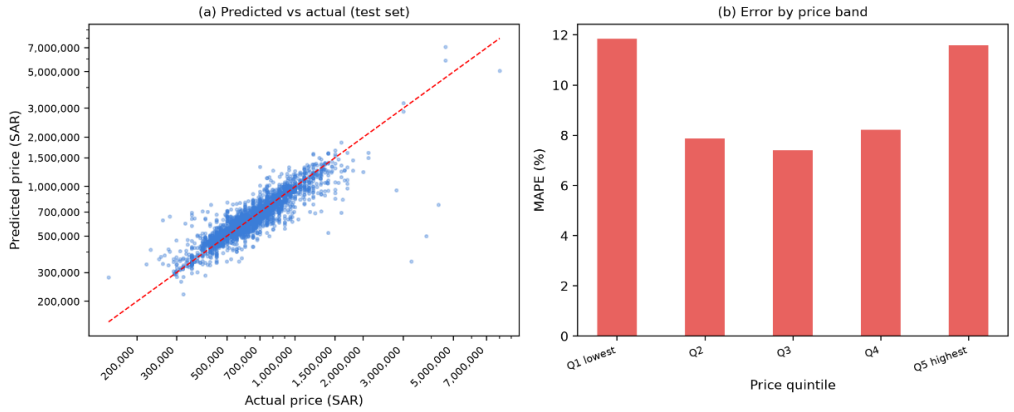

Fig. 7. Predicted versus actual price (test set) and error by price band: (a) predicted versus actual price (test set); (b) mean absolute percentage error by price quintile.

V. DISCUSSION

A. Data Quality as a Substantive Finding

Nearly a quarter of the raw records were repeated advertisements and a further 2% were near-copies. A model trained and evaluated on the raw file would encounter many test listings whose twin lies in the training set, which would inflate accuracy in the manner documented for leakage in other fields [17] and [46]. Studies that scrape Aqar.fm-type platforms [10], [11] should therefore report deduplication and the way training and test sets were separated. This observation is not a criticism of any specific study: even the Khrais and Shidwan paper does not describe a deduplication step (Section II-B).

B. Model Choice in Context

Boosted trees outperformed the linear baselines by a wide margin (Section IV-F). This agrees with the Kuala Lumpur study, in which LightGBM and XGBoost also outperformed multiple and ridge regression [61], and with Saudi studies in which non-linear ensembles outperformed linear models [11], [12], [13]. LightGBM and XGBoost were statistically tied on the cross-validation criterion. Applying the stated tie-break rule selected LightGBM, which also led on every test-set metric, so a reasonably clear choice between the two boosting models was possible here, though the margin between them is still small relative to a single split. The ordering reported for Kuala

Lumpur is therefore broadly reproduced. Absolute levels are not comparable across papers. The log-scale R^2 obtained here (0.78–0.80) is below the test R^2 of 0.82 reported for XGBoost on Aqar.fm data [11] and the adjusted R^2 of 0.91 reported for Kuala Lumpur [61], but those studies differ in data, target scale, cleaning, and validation. A design without grouped splitting could yield higher scores; this was not tested. Random forest, the best model in several Saudi studies, was included here (Section III-F) but was not selected: it had a higher cross-validated RMSE than the two boosting models, although it recorded the lowest test-set MAE of the five models (Section IV-F). This underlines that model selection by a single criterion and a single split is not fully conclusive.

C. What Location and Text Add

Two conclusions follow from the ablation and importance results. First, location is the principal source of signal. Adding city, facing direction, street width, and district to the structural attributes reduced the test MAE by 45% (134,218 to 73,645 SAR), and district and city identity were the two highest-ranked inputs. This agrees with the weight given to location in the literature in [26] and [59], and is consistent with the region-specific findings reported for Kuala Lumpur [61]. Second, Arabic text adds measurable but modest value. Building-level flags reduced the MAE by a further 4.9% beyond location, with driver's room, independent meters and maid's room among the most used flags. Neighbourhood-service flags changed accuracy negligibly — a null result given the overlapping confidence intervals noted in Section IV-G, rather than clear evidence that this information has no value. The latter finding is nonetheless broadly consistent with the finding that some textual statements are redundant given other information [39]: mentions of nearby mosques, parks, or schools are seller claims that district identity already captures. Richer text representations reduced the MAE by up to 17% in Berlin and Los Angeles data [15], so simple keyword flags probably capture only part of the information in Arabic descriptions.

The SHAP decomposition agrees with the ablation: location-related inputs account for 51.1% of the total mean absolute SHAP value, structural attributes for 38.5%, building text features for 8.4% and neighbourhood-service text features for 2.0%.

A caution applies irrespective of the numbers. Text flags are written by sellers, who may mention amenities more often when the asking price is higher or the property is newer or larger. A flag that improves prediction therefore does not necessarily add value to the apartment; it may proxy seller behaviour or unobserved quality.

D. Transferability Across Cities

The limited transfer to Dammam without Dammam in the training data (MAPE 20.2% against 7.0% under a random split, Section IV-H) contrasts with the successful transfers reported elsewhere [50] [51] and [52]. The main reconciliation, established directly from how the encoding is constructed (Section IV-H), is that the location signal here is carried by identity features (city and district) that are categorically undefined for an unseen city, whereas the successful transfers relied on continuous spatial or accessibility features and, in some cases, calibration [51], [52]. It also mirrors the generally poor geographic generalisation observed for ML models [49]. For practice, a model of this type should be trained and validated within each market, or extended with geocoded distance features that can travel between cities [60]. This identity-feature explanation follows from the encoding mechanism itself rather than from a controlled test; whether continuous location features would transfer better in this setting remains untested.

E. Arabic Text Processing

Keyword rules are transparent, inexpensive, and easy to audit, and they encode Gulf spelling variants explicitly. Their weaknesses are evident in Section IV-E: many flags are rare, and rules cannot handle negation or context (a description stating that a mosque is distant still triggers the mosque flag). Richer alternatives include normalisation and stemming pipelines [42], dialect-aware processing [44], and pretrained Saudi-dialect language models such as SaudiBERT [45], whose embeddings could be evaluated in the same ablation. A lightweight step in this direction was taken here: a TF-IDF representation reduced by truncated SVD was substituted for the keyword flags (Section IV-G) and did not outperform them. Whether a larger pretrained embedding would do better, as description embeddings did in [15] and [38], remains open.

F. Implications

For buyers and lenders, an interpretable model that shows which features raise or lower an estimate can support negotiation and collateral checks. For planners, city-level error patterns show where asking prices are hardest to anchor. Given the policy emphasis on affordability and transparency reported in [3] and [4], such tools complement, and do not replace, formal valuation. The test MAE of the selected model (69,985 SAR) is about 11% of the median price of 650,000 SAR, which suits the screening of asking prices rather than exact valuation.

G. Limitations

Several limitations qualify these findings and should guide their interpretation.

- Asking prices, not transaction prices, were modelled; they include negotiation margins and seller strategy.
- Three cities supply 87% of the data, and the model relies on city and district identity rather than portable location features; it did not transfer to Dammam when excluded from training (Section IV-H), and geocoded distances to services, which improved forecasts elsewhere [59] and [60], were not available.
- Text features are simple keyword rules that cannot resolve context, and many are rare (Section IV-E); a TF-IDF alternative performed slightly worse rather than better (Section IV-G), but larger pretrained Arabic language models were not tried.
- The bootstrap confidence intervals reported throughout (Sections IV-F, IV-G) quantify sampling uncertainty given the test set drawn, but not uncertainty from the choice of train/test split itself, which was fixed by a single random seed; a repeated-split or nested cross-validation design was not run.

- Grouping train and test by identical description text keeps an exact repost out of the opposite partition but not an edited repost or other near-identical listings (Section IV-G); reported accuracy should be read as an upper bound for this reason as well as for the spatial dependence noted above.

VI. Conclusion and Future Work

This study developed and documented an end-to-end pipeline for predicting apartment prices from Arabic online listings in Saudi Arabia. The data-quality findings are firm: about a quarter of the raw records had to be removed, mainly repeated advertisements, and the remaining data are concentrated in three cities. The pipeline converts Arabic descriptions into 50 interpretable features, compares five regression models, including random forest, under grouped validation with bootstrap confidence intervals, and tests generalisation to an unseen city. LightGBM, selected by cross-validation, clearly outperformed linear regression on every test-set metric (MAPE 9.3% against 12.9% for multiple regression). Location supplied most of the predictive signal, and building-related Arabic text features added a small but measurable gain, while neighbourhood-service text features added negligibly more. A model trained without a given city predicted that city poorly (MAPE 20.2% against 7.0% under a random split), mainly because its location encoding is undefined for an unseen city rather than because location signal cannot transfer in general. Future work should add transaction prices, geocoded distances to services that could transfer across cities, and listing dates. It should also use repeated-split or nested cross-validation to quantify split-selection uncertainty, and evaluate dialect-aware Arabic language models as a further alternative to the keyword flags and the TF-IDF baseline evaluated here.

Acknowledgment

The authors would like to acknowledge the Deanship of Graduate Studies and Scientific Research, Taif University, for funding this work.

References

[1] A. M. Alhamoudi, ‘Enabling Industry 4.0 Technologies to Drive Sustainable Housing Delivery across the Housing Supply Value Chain in Saudi Arabia: Challenges and Prospects’, Sustainability, vol. 16, no. 13, p. 5413, Jun. 2024, doi: 10.3390/su16135413.

[2] M. Hariri, ‘Influence of Saudi Vision 2030 on housing-related macroeconomic variables’, International Journal of Housing Markets and Analysis, vol. 16, no. 4, pp. 828–845, Aug. 2023, doi: 10.1108/IJHMA-03-2022-0046.

[3] M. F. Alhajri, ‘Housing challenges and programs to enhance access to affordable housing in the Kingdom of Saudi Arabia’, Ain Shams Engineering Journal, vol. 13, no. 6, p. 101798, Nov. 2022, doi: 10.1016/j.asej.2022.101798.

[4] A. Alqahtany, ‘Affordable housing in Saudi Arabia’s vision 2030: new developments and new challenges’, International Journal of Housing Markets and Analysis, vol. 14, no. 1, pp. 243–256, Feb. 2021, doi: 10.1108/IJHMA-04-2020-0035.

[5] S. C. K. Tekouabou, Ş. C. Gherghina, E. D. Kameni, Y. Filali, and K. Idrissi Gartoumi, ‘AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey’, Archives of Computational Methods in Engineering, vol. 31, no. 2, pp. 1079–1095, Mar. 2024, doi: 10.1007/s11831-023-10010-5.

[6] W. K. O. Ho, B.-S. Tang, and S. W. Wong, ‘Predicting property prices with machine learning algorithms’, Journal of Property Research, vol. 38, no. 1, pp. 48–70, Jan. 2021, doi: 10.1080/09599916.2020.1832558.

[7] S. Lahmiri, S. Bekiros, and C. Avdoulas, ‘A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization’, Decision Analytics Journal, vol. 6, p. 100166, Mar. 2023, doi: 10.1016/j.dajour.2023.100166.

[8] P.-F. Pai and W.-C. Wang, ‘Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices’, Applied Sciences, vol. 10, no. 17, p. 5832, Aug. 2020, doi: 10.3390/app10175832.

[9] M. A. Shbool, R. Al-Dmour, B. A. Al-Shboul, N. T. Albashabsheh, and N. Almasarwah, ‘Real estate decision-making: precision in price prediction through advanced machine learning algorithms’, International Journal of Housing Markets and Analysis, vol. 19, no. 3, pp. 771–795, Apr. 2026, doi: 10.1108/IJHMA-01-2025-0004.

[10] E. Alzain, A. S. Alshebami, T. H. H. Aldhyani, and S. N. Alsubari, ‘Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia’, Electronics (Basel), vol. 11, no. 21, p. 3448, Oct. 2022, doi: 10.3390/electronics11213448.

[11] S. Alsulamy, ‘Artificial intelligence models in real estate forecasting: a machine learning approach for sustainable urban construction in Saudi Arabia’, Engineering, Construction and Architectural Management, vol. 33, no. 6, pp. 4833–4852, May 2026, doi: 10.1108/ECAM-06-2024-0796.

[12] T. Alshammari, ‘Evaluating machine learning algorithms for predicting house prices in Saudi Arabia’, in 2023 International Conference on Smart Computing and Application (ICSCA), IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICSCA57840.2023.10087486.

[13] A. Louati, R. Lahyani, A. Aldaej, A. Aldumaykhi, and S. Otai, ‘Price forecasting for real estate using machine learning: A case study on Riyadh city’, Concurr. Comput., vol. 34, no. 6, Mar. 2022, doi: 10.1002/cpe.6748.

[14] M. Ezzeldin and M. Abouagwa, ‘Predicting Housing Prices in Riyadh: A Comparative Analysis on Machine Learning and Hedonic Regression’, Innovation in Statistics and Probability, vol. 1, no. 2, pp. 26–44, Nov. 2025, doi: 10.64389/isp.2025.01220.

[15] K. Baur, M. Rosenfelder, and B. Lutz, ‘Automated real estate valuation with machine learning models using property descriptions’, Expert Syst. Appl., vol. 213, p. 119147, Mar. 2023, doi: 10.1016/j.eswa.2022.119147.

[16] A. Nowak and P. Smith, ‘Textual Analysis in Real Estate’, Journal of Applied Econometrics, vol. 32, no. 4, pp. 896–918, Jun. 2017, doi: 10.1002/jae.2550.

[17] S. Kapoor and A. Narayanan, ‘Leakage and the reproducibility crisis in machine-learning-based science’, Patterns, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[18] Q. H. Mohammed, Sadeer Sadeq, Mohammed Hamid Alkubaisi, and Mina Faris Alnaimy, ‘Data Mining and Machine Learning Approaches for Real Estate Valuation: A Systematic Review of Predictive Accuracy’, Journal of Telecommunication and Computer Technology, vol. 1, no. 1, pp. 57–81, May 2026, doi: 10.25195/jtct.v1i1.3.

[19] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, ‘House price prediction using hedonic pricing model and machine learning techniques’, Concurr. Comput., vol. 34, no. 27, Dec. 2022, doi: 10.1002/cpe.7342.

[20] T. Yoshida, D. Murakami, and H. Seya, ‘Spatial Prediction of Apartment Rent using Regression-Based and Machine Learning-Based Approaches with a Large Dataset’, The Journal of Real Estate Finance and Economics, vol. 69, no. 1, pp. 1–28, Jul. 2024, doi: 10.1007/s11146-022-09929-6.

[21] A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit, ‘Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms’, Cities, vol. 131, p. 103941, Dec. 2022, doi: 10.1016/j.cities.2022.103941.

[22] J. R. Rico-Juan and P. Taltavull de La Paz, ‘Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain’, Expert Syst. Appl., vol. 171, p. 114590, Jun. 2021, doi: 10.1016/j.eswa.2021.114590.

[23] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[24] G. Ke et al., ‘LightGBM: a highly efficient gradient boosting decision tree’, in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 3149–3157.

[25] R. Shwartz-Ziv and A. Armon, 'Tabular data: Deep learning is not all you need', Information Fusion, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[26] N. Hamizah Zulkifley, S. Abdul Rahman, N. H. Ubaidullah, and I. Ibrahim, 'House Price Prediction using a Machine Learning Model: A Survey of Literature', International Journal of Modern Education and Computer Science, vol. 12, no. 6, pp. 46–54, Dec. 2020, doi: 10.5815/ijmecs.2020.06.04.

[27] A. Chikhmous and M. T. Rahman, 'Examining the effect of apartment attributes on their sale prices in Riyadh, Saudi Arabia', Spatial Information Research, vol. 32, no. 4, pp. 411–424, Aug. 2024, doi: 10.1007/s41324-023-00565-7.

[28] L. T. Khrais and O. S. Shidwan, 'The role of neural network for estimating real estate prices value in post COVID-19: a case of the middle east market', International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no. 4, p. 4516, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4516-4525.

[29] E. A. Alburshaid and R. Ksantini, 'Real Estate Predictive Analysis for the Kingdom of Bahrain Using Web Scraping and Machine Learning', in 2024 Arab ICT Conference (AICTC), IEEE, Feb. 2024, pp. 101–107. doi: 10.1109/AICTC58357.2024.10735042.

[30] H. A. Khunji and A. M. Zeki, 'Comparing Linear Regression, Random Forest and Support Vector Machine for Rental House Price Forecasting in Bahrain', in 2024 5th International Conference on Data Analytics for Business and Industry (ICDABI), IEEE, Oct. 2024, pp. 169–172. doi: 10.1109/ICDABI63787.2024.10800514.

[31] A. Grybauskas, V. Pilinkienė, and A. Stundžienė, 'Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic', J. Big Data, vol. 8, no. 1, p. 105, Dec. 2021, doi: 10.1186/s40537-021-00476-0.

[32] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, 'Identifying Real Estate Opportunities Using Machine Learning', Applied Sciences, vol. 8, no. 11, p. 2321, Nov. 2018, doi: 10.3390/app8112321.

[33] T. Jach, 'Web Scraping Methods Used in Predicting Real Estate Prices', in International Conference on Computational Collective Intelligence, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238424806

[34] C. Wei, M. Fu, L. Wang, H. Yang, F. Tang, and Y. Xiong, 'The Research Development of Hedonic Price Model-Based Real Estate Appraisal in the Era of Big Data', Land (Basel)., vol. 11, no. 3, p. 334, Feb. 2022, doi: 10.3390/land11030334.

[35] C. Meyberg, U. Rendtel, and H. Leerhoff, 'Flat rent price prediction in Berlin with web scraping', AStA Wirtschafts- und Sozialstatistisches Archiv, vol. 18, no. 2, pp. 245–278, Jun. 2024, doi: 10.1007/s11943-024-00340-6.

[36] A. Brenning and S. Henn, 'Web scraping: a promising tool for geographic data acquisition', ArXiv, vol. abs/2305.19893, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:258987409

[37] L. Shen and S. Ross, 'Information value of property description: A Machine learning approach', J. Urban Econ., vol. 121, p. 103299, Jan. 2021, doi: 10.1016/j.jue.2020.103299.

[38] H. Zhang, Y. Li, and P. Branco, 'Describe the house and I will tell you the price: House price prediction with textual description data', Nat. Lang. Eng., vol. 30, no. 4, pp. 661–695, Jul. 2024, doi: 10.1017/S1351324923000360.

[39] Y. Seo, J. Im, and B. Mikelbank, 'Does the Written Word Matter? The Role of Uncovering and Utilizing Information from Written Comments in Housing Ads', Journal of Housing Research, vol. 29, no. 2, pp. 133–155, Jul. 2020, doi: 10.1080/10527001.2020.1849929.

[40] J. Węgrzyn, 'Mining Real Estate Data: A Systematic Review of Text-Based Approaches', Real Estate Management and Valuation, vol. 34, no. 1, pp. 1–15, Mar. 2026, doi: 10.2478/remav-2026-0001.

[41] B. Brahimi, 'Arabic Text Mining for Used Cars and Equipments Price Prediction', Computación y Sistemas, vol. 26, no. 2, Jun. 2022, doi: 10.13053/cys-26-2-3185.

[42] A. A. Nafea et al., 'A Brief Review on Preprocessing Text in Arabic Language Dataset: Techniques and Challenges', Babylonian Journal of Artificial Intelligence, vol. 2024, pp. 46–53, May 2024, doi: 10.58496/BJAI/2024/007.

[43] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, 'Arabic natural language processing: An overview', Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 5, pp. 497–507, Jun. 2021, doi: 10.1016/j.jksuci.2019.02.006.

[44] N. Alshenaifi, A. M. Azmi, and M. Hosny, 'Comparative analysis of deep learning and transfer learning techniques for dialect identification in the Gulf and Saudi regions', Applied Computing and Informatics, pp. 1–14, Nov. 2025, doi: 10.1108/ACI-02-2025-0067.

[45] F. Qarah, 'SaudiBERT: A Large Language Model Pretrained on Saudi Dialect Corpora', ArXiv, vol. abs/2405.06239, p., 2024, doi: 10.48550/arxiv.2405.06239.

[46] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, 'Data leakage inflates prediction performance in connectome-based machine learning models', Nat. Commun., vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

[47] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, 'Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images', Sci. Data, vol. 9, no. 1, p. 580, Sep. 2022, doi: 10.1038/s41597-022-01618-6.

[48] J. Deppner and M. Cajias, 'Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach', The Journal of Real Estate Finance and Economics, vol. 68, no. 2, pp. 235–273, Feb. 2024, doi: 10.1007/s11146-022-09915-y.

[49] S. Law, P. Jeszenszky, and K. Yano, 'Examining geographical generalisation of machine learning models in urban analytics through street frontage classification and house price regression', Sep. 2021. doi: 10.25436/E2VC71.

[50] D. P. McArthur, L. Osland, and I. Thorsen, 'Spatial Transferability of Hedonic House Price Functions', Reg. Stud., vol. 46, no. 5, pp. 597–610, May 2012, doi: 10.1080/00343404.2010.521147.

[51] M. M. Orvin, M. Fatmi, and I. Hoque, 'Housing price dynamics within an integrated urban model: testing the spatial and temporal transferability', International Journal of Urban Sciences, vol. 30, no. 3, pp. 711–741, Jul. 2026, doi: 10.1080/12265934.2025.2517097.

[52] C. Kmen, G. Navratil, and I. Giannopoulos, 'Predicting the Unseen: Transductive Transfer Learning in Real Estate Price Prediction', Transactions in GIS, vol. 30, no. 1, Feb. 2026, doi: 10.1111/tgis.70217.

[53] S. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', 2017, pp. 4765–4774. doi: 10.4230/oasics.dx.2024.27.

[54] S. M. Lundberg et al., 'From local explanations to global understanding with explainable AI for trees', Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[55] S. M. Lundberg, G. Erion-Barner, and S.-I. Lee, 'Consistent Individualized Feature Attribution for Tree Ensembles', ArXiv, vol. abs/1802.03888, 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:3626364

[56] M. C. Iban, 'An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants', Habitat Int., vol. 128, p. 102660, Oct. 2022, doi: 10.1016/j.habitatint.2022.102660.

[57] F. Trindade Neves, M. Aparicio, and M. de Castro Neto, 'The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities', Applied Sciences, vol. 14, no. 5, p. 2209, Mar. 2024, doi: 10.3390/app14052209.

[58] D. Tchuente, 'Real Estate Automated Valuation Model with Explainable Artificial Intelligence Based on Shapley Values', The Journal of Real Estate Finance and Economics, vol. 72, no. 3, pp. 567–605, Apr. 2026, doi: 10.1007/s11146-024-09998-9.

[59] D. Tchuente and S. Nyawa, 'Real estate price estimation in French cities using geocoding and machine learning', Ann. Oper. Res., vol. 308, no. 1–2, pp. 571–608, Jan. 2022, doi: 10.1007/s10479-021-03932-5.

[60] D. Rey-Blanco, J. L. Zofío, and J. González-Arias, 'Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses', Expert Syst. Appl., vol. 235, p. 121059, Jan. 2024, doi: 10.1016/j.eswa.2023.121059.

[61] S. Abdul-Rahman, N. H. Zulkifley, I. Ibrahim, and S. Mutalib, 'Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur', International Journal of Advanced Computer Science and Applications, vol. 12, no. 12, 2021, doi: 10.14569/IJACSA.2021.0121291.

[62] N. Duan, 'Smearing Estimate: A Nonparametric Retransformation Method', J. Am. Stat. Assoc., vol. 78, no. 383, pp. 605–610, Sep. 1983, doi: 10.1080/01621459.1983.10478017.

[63] K. Belitz and P. E. Stackelberg, 'Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models', Environmental Modelling & Software, vol. 139, p. 105006, May 2021, doi: 10.1016/j.envsoft.2021.105006.

[64] D. Micci-Barreca, 'A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems', ACM SIGKDD Explorations Newsletter, vol. 3, no. 1, pp. 27–32, Jul. 2001, doi: 10.1145/507533.507538.

[65] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, 'Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features', Comput. Stat., vol. 37, no. 5, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.