Int. J. Advance Soft Compu. Appl, Vol. 18, No. 1, March 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Predicting Apartment Prices in Jordan Using Ensemble Machine Learning Algorithms to Support Investment and Urban Planning Decisions

Nisrean Thalji[1] and Mohammad Abdulaziz Alwadi[2]
[1]Faculty of Artificial Intelligence, Department of Intelligent Systems, Al-Balqa Applied University, Al-Salt, Jordan
n.thalji@bau.edu.jo
[2]Faculty of Computer Studies, Arab Open University, Amman, Jordan
m_alwadi@aou.edu.jo

Abstract

The housing market is of great significance to the development and advancement of cities, but customary forms of property valuation are frequently biased, time-consuming, and not always effective. This paper focuses on the city of Irbid in Jordan, aiming to collect all the information on apartments and houses, predict the prices of properties, and clarify the key factors influencing the prices. Following the comprehensive cleaning process of the data and exploratory analysis, three ensemble machine learning models were trained and optimized to achieve accurate price predictions. The performance of all three models demonstrated excellent and consistent predictions, highlighting the efficiency of ensemble methods in predicting property prices. SHAP analysis indicated that the size of the house, the number of bedrooms, the number of lounges as well as the location are the most significant factors influencing the prices in Irbid. This reflects the functioning of the local market.

Keywords - Irbid city real estate; Random Forest; Extreme Gradient Boosting; Categorical Boosting; Machine learning; Ensemble models.

1 Introduction

Housing market is a central determinant of economic growth and urban development, thus making an accurate valuation of property to be an indispensable quality of just transactions and well supported planning [1], [2]. However, the conventional methods of valuation are subjective in nature, time-consuming and cannot provide any reliable and uniform estimates [3]. Recent advances in machine learning have allowed the analysis of more complex real-estate data more effectively and with greater accuracy, in particular, ensemble models that can capture nonlinear relationships between covariates (size, age, and location) [4].

Irbid, the second-largest city in Jordan, is undergoing a tremendous population increase and is recording the high demand of apartments, and hence a dynamic real-estate scenario where price prediction is greatly sought after. This paper utilizes three ensemble algorithms

Received 4 December 2025; Accepted 27 February 2026

that include RFR, XGBR, and CBR to estimate the prices of apartments in Irbid. RFR goes through the reduction of variance using many decision trees [5], XGBR is boosted using gradient boosting and regularization [6], and CBR effectively works with categorical variables by using ordered boosting [7].

2 Related Work

Past studies have shown that RFR has great performance on several types of housing data sets [8]–[12], and XGBR has always shown better predictive capability in establishing complex trends [13]. CBR also works with mixed data and categorical data such as research on King County housing [14],[15]. Comparative international studies have a consistent report of the superiority of CBR and XGBR to RFR, though RFR continues to be rather robust in the context of heterogeneous data sets [16]–[18].

A recent study [19] applied an automated house price prediction framework integrating genetic algorithms (GA) for feature optimization and ANOVA for statistical analysis. Five ensemble models XGBR, RFR, CBR, Adaptive Boosting Regression (ADBR), and Gradient Boosted Decision Trees Regression (GBDTR) were compared, with Categorical Boosting combined with GA (CBRGA) identified as the best-performing model.

Moreover, the need to have larger and region-specific data as well as a better model interpretability is stated in earlier literature especially in the context of the fast urbanization [20]. Current studies in Jordan have been based on, either, the old paradigms of hedonic regressions, simple data-mining tools or a few machine-learning models that can only work with Amman on their own without comparing ensemble models [21], [22]. Furthermore, recent research in the Jordanian context has highlighted the effectiveness of the modern machine learning methods namely the RFR algorithm and the deep learning models in forecasting the price-related dynamics in ancillary economic industries, such as the fuel sales [23].

Despite global advancements in machine learning, no machine learning-based predictive studies have been conducted specifically for Irbid. Previous studies indicate the following: the lack of use of regression models (RFR, XGBR, CBR) in Irbid or any other city in Jordan; the limited scope and methodology of Jordanian research; the need for larger and more specific data for the region; the absence of a dataset specific to Irbid; and the lack of comparative evaluation of regression models within the Jordanian context. To address this gap, this research aims to: establish a dataset at the apartment and house levels in Irbid; and provide area-specific predictive analyses that can benefit real estate professionals.

3 Materials and Methods

This research predicts housing prices and explains how the model works. It also shows the factors that affect prices. The study covers the dataset, data preparation, feature selection, model training, result validation, and interpretation tools.

3.1. Dataset Description

As many apartment and house listings as possible were collected from various regions of Jordan. Specific specifications of Irbid were then obtained and applied to predict the prices of apartments and houses in Irbid. This city has particular traits that would not be found in other Jordanian cities, property patterns and prices can be quite different in those places. Information was gathered on some of the Jordanian real estate sites such as OpenSooq, Kharta, and social media advertisements regarding various regions of Irbid starting in the year 2024 up to September 2025. In the process of collecting data, emphasis was laid on the acquisition of the key attributes affecting the property price, i.e., the property size, the number of bedrooms, the number of bathrooms, the number of living rooms, the number of sitting areas, the floor, the property age, the condition of property (new or old), the existence of the balcony, furnishing, as well as the number of parking spaces, and the property price, for a total of 13 attributes. These attributes are summarized in Table 1 to present their names, data types, and descriptions clearly and transparently. Such a systematic information aids in interpreting the structure of the data set involved and guarantees its reproducibility in the course of the research process.

One should also note that the gathering of the information on the online platforms and social media advertisements can also potentially lead to sampling bias since not all the property listings are equally represented online. There may be more active advertising of certain types of property, price levels, or geographic areas, so the distribution of the data in general may be influenced. In order to reduce this impact, data were gathered through a variety of independent sources representing different regions, and types of property so as to enhance representativeness. Moreover, the relatively high amount of data and incorporation of heterogeneous property characteristics assist in decreasing the effects of such bias. Future research could include other sources of data such as official real estate records, agency database, to further improve the completeness and representativeness of the data.

The dataset supporting the findings of this study is publicly available in the Zenodo repository [24]. Providing open access to the dataset ensures full transparency, reproducibility, and accessibility for future research, benchmarking, and comparative studies in real estate price prediction. In addition, all preprocessing, feature engineering, and hyperparameter tuning steps are fully documented to ensure reproducibility. It is necessary to note that the dataset will include the period between 2024 and September 2025, which reflects the current market trends in Irbid. Although this period is a current representation of the housing market, the temporal generalizability of the model to previous or future market conditions may be limited especially when there are economic variations, policy variations, or a long-term pattern in the market. However, the dataset will include a significant number of listings in different areas and property features, which will guarantee a representative sample of the modern housing market. The research can be improved in future by adding several years of past data to the dataset in order to further develop the model generalization and time capabilities.

Table 1: Description of Attributes

Feature	Data Type	Description
Area	Categorical	Property district
Size_m2	Numerical	Property area in square meters

Bedrooms	Numerical	Number of bedrooms
Bathrooms	Numerical	Number of bathrooms
Halls	Numerical	Number of halls
Lounges	Numerical	Number of lounges
Floor	Numerical	Property floor number
Age	Categorical	Property age
New	Categorical	Property condition: new or old
Balcony	Categorical	Balcony presence
Furnished	Categorical	Furnished status
Parking	Categorical	Availability of parking
Price_kJD	Numerical	Price in thousand JD

It is worth noting that the information captured in apartment sale listings in Irbid lacks many details related to neighborhood features and local facilities, such as neighborhood quality, educational facilities, socio-economic status, and environmental factors (e.g., noise and pollution levels). These listings also do not provide site-specific features, including proximity to transport infrastructure, the city center or satellite employment hubs, and shopping centers, schools, or other amenities. In addition, such listings rarely contain information about investors, management organizations, or developers, and therefore do not include assessments of management or developer reputation. The absence of detailed neighborhood-level attributes may limit the model's ability to fully capture location-specific value drivers, as proximity to schools, commercial centers, transportation networks, and infrastructure is known to significantly influence real estate prices. These external environmental and accessibility factors often contribute to price variation beyond property-specific characteristics. Despite this limitation, the dataset includes key structural and location-related features that provide strong predictive capability. Future research may integrate geographic information system (GIS) data, distance-based features, and neighborhood socio-economic indicators to further improve prediction accuracy and enhance model comprehensiveness.

3.2 Data Cleaning

A summary of the techniques applied in this stage is presented in Table 2. The data cleaning process was carried out by handling missing values by median substitution for numerical variables and mode for categorical variables, removing outliers and duplicates while retaining the first, deleting invalid values and type errors, standardizing inconsistent text formatting, and removing contradictory, illogical, or logically anomalous entries to ensure data quality and accuracy.

Table 2: Data Cleaning Summary

Category	Examples	How we handle it
Missing Values	Size_m2 = NaN, Bedrooms = NaN	Impute numeric with median, categorical with mode
Outliers	Floor = 40	Remove
Duplicates	Identical rows	Remove duplicates, keep first occurrence

Invalid Values	Bedrooms = -1, Floor = -10	Remove
Inconsistent Formatting	north vs North	Standardize text
Type Errors	Size_m2 = 'large'	Remove
Contradictory	Furnished=True but extremely low price	Remove
Illogical Combinations	Bedrooms=5 with Bathrooms=1	Remove
Logical Outliers	1BR and 150 m²	Remove

3.3 Exploratory Data Analysis (EDA)

The dataset now contains 4928 records. A summary of the dataset after cleaning is shown in Table 3. The table presents descriptive statistics for the main property dataset characteristics. It presents the lowest, highest, average, median and standard deviation of each characteristic. There is a variation in the size of property, and the prices of apartments; to give an example, property sizes vary between 40 and 370 m 2, with an average of 165.95 m 2 and a median of 146 m2, whereas the apartment prices vary between 7,000 and 196,700 Jordanian dinars with an average of 68,510 dinars and a median of 61400 dinars. Bedrooms are between 1 and 5 with the mean of about 3, and the number of bathrooms between 1 and 5 with the mean being 2.62. The rooms will remain the same number of halls per apartment (1), the number of Lounges will be between 0 and 2, with the average of 0.65. The range of floor height is between -2 and 3 with an average of 0.52 indicating the fact that there are a few apartments which are at the lower floor and the distribution of floors. On the whole, these statistics give an initial idea of the data distribution and dispersion.

Table 3: Summary Statistics of Numerical Features

Feature	Min	Max	Mean	Median	Std
Size_m2	40.0	370.0	165.946	146.0	88.322
Price_kJD	7.0	196.7	68.510	61.4	38.384
Bedrooms	1.0	5.0	2.917	3.0	1.319
Bathrooms	1.0	5.0	2.620	3.0	1.230
Halls	1.0	1.0	1.000	1.0	0.000
Lounges	0.0	2.0	0.652	1.0	0.651
Floor	-2.0	3.0	0.521	1.0	1.719

To model the data, we first applied an extensive EDA in order to visualize and comprehend the data and then continue with further analysis and model development by revealing patterns and distributions thereby enhancing the quality and interpretability of the data. We have explored this dataset by using correlation and violin plots in this study. A correlation matrix is generated to identify linear relationships among numerical features, as shown in Fig. 1.

The values, known as Pearson's Correlation Coefficients (r), r varies between -1 to +1. A value of 1.00 (dark red) represents a perfect positive correlation (dark red) represents a perfect positive correlation as well as values towards 0 (dark blue), there is little or no statistical relationship among variables.

The statistical analysis of the housing data in Irbid demonstrates that the most significantly affecting factors on the price of a house comprise the physical size and capacity of the house: Total Size corresponds with price (price = 0.93), and the correlation is very strong. Bedrooms are associated with price (r = 0.90). There is also the strong positive correlation between Number of Lounges (r = 0.85). This shows that functional space and number of rooms are the major motivators of the Irbid housing market and the value of property increases in line with these. Also, the data reveals an almost perfect collinearity between Total Size and Number of Bedrooms (r = 0.96) that can be interpreted as a consistent trend in the standardized architectural strategy in Irbid where larger floor plans are virtually only utilized to provide more bedrooms instead of incorporating other luxury features. Interestingly, there are features that have very little effect on the valuation: Number of Bathrooms is weakly correlated with price (r = 0.24). The impact of the Floor Level on the value of the property is virtually insignificant (r = 0.02). To sum up, the real estate pricing model of the city of Irbid is family-based. Horizontal space and social spaces (lounges) are the main sources of value, whereas vertical location (floor level) and secondary amenities (lounges) have an insignificant role in the process of local valuation.

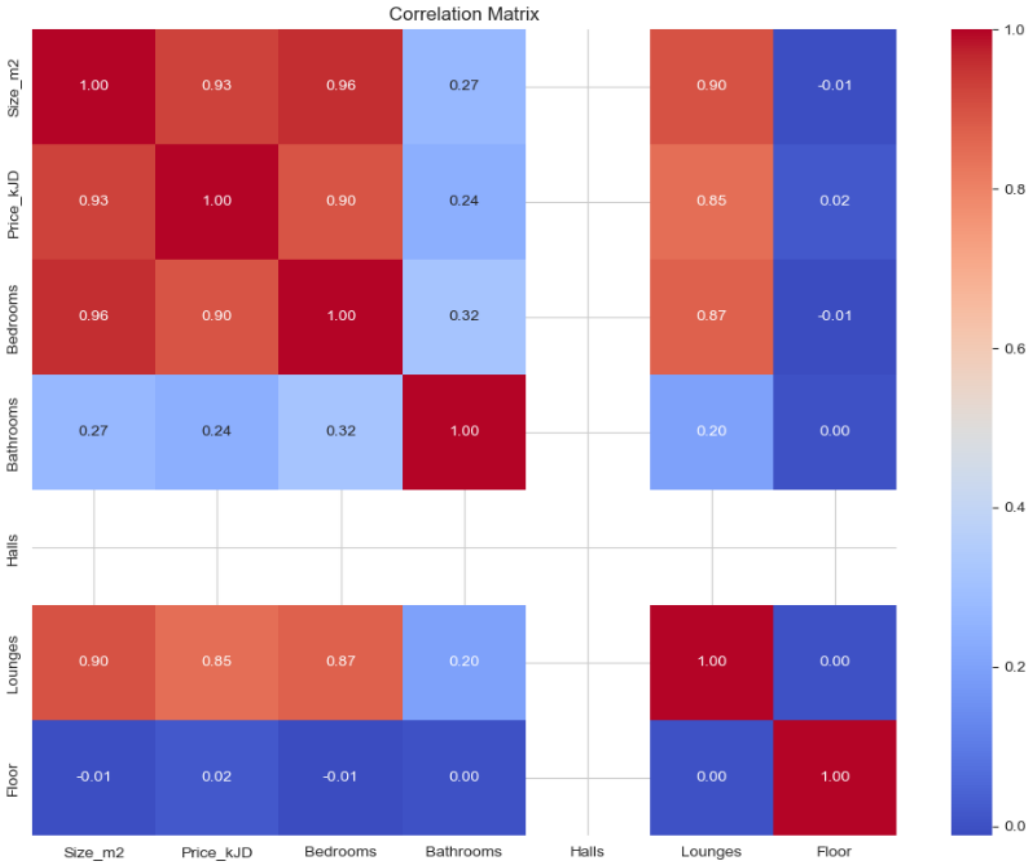

Fig. 1. Correlation Matrix for Numerical Features

Fig. 2 statistic provides a collection of violin plots, used to represent the distribution of values, concentration regions, the median, the interquartile range, and the overall distribution of data of each numerical variable in the data set.

The features were included in the figure, namely: area in square meters (Size_m2), price in thousands Jordanian Dinars (Price_kJD), bedrooms (Bedrooms), bathrooms (Bathrooms), halls (Halls), lounges (Lounges), and floor (Floor). The area plot indicates that the values are approximately less than 50 to 400 m 2 with the interquartile range lying within the range of about 100 to 220 m 2, and a median equal to 150 m 2 thus giving an indication that most of the apartments are of mid-size and fewer of large sizes. The price is spread out as 0 to 200 thousand JD with some of the values concentrated around the middle range (40 to 90 thousand JD) with a median of about 65 thousand JD with a tail to the higher end, which shows that there are a few units of high price though not very many. In the case of bedrooms, the multimodal distribution is observed where the central tendency is very clear with the highest number of bedrooms being 3. The bathrooms are also similar with the more common number of 2 to 3 and also with several peaks indicating the discrete character of the variable. The concentration is high with a limit of 1 in the halls plot which demonstrates low variability in this aspect. There are two distinct heights at 0 and 1 which have little extension to 2 and this indicates that most units have no lounge or only one. Lastly, floor variable is approximately between -2 and 3 with a larger proportion lying in the mid-rise buildings (0 to 1) as most of the apartments are located in low-rise buildings. In general, these plots show the distributions, concentration regions and possible data deviations, which will give a rich visual insight into the housing characteristics, will assist the exploratory data analysis and assist in constructing more accurate predictive models.

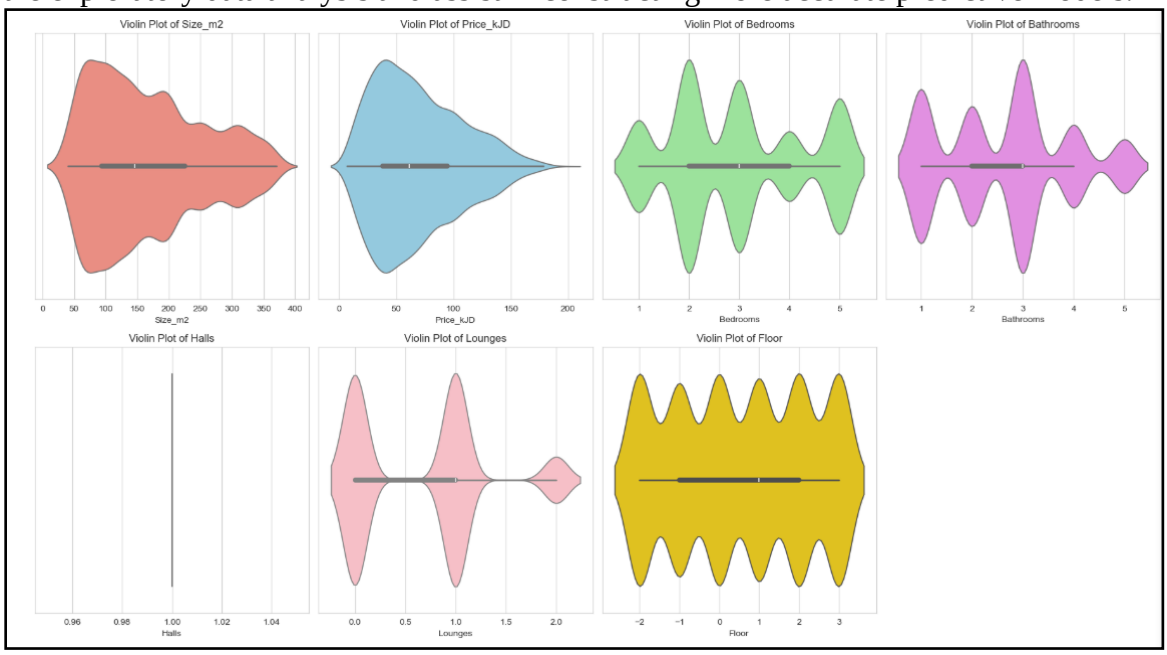

Fig. 2. Violin plots of numerical features in the Irbid housing dataset

3.4 Feature Engineering and Feature Scaling

In the feature engineering and scaling stage, several transformations were applied to prepare the dataset for modeling. First, new features were created to enhance the

representation of apartment characteristics. Total_Rooms was computed as the sum of bedrooms, bathrooms, halls, and lounges to represent the total living space. Admittedly, Total Rooms feature is directly linked with the sets of variables it comprises, which can create multicollinearity in some situations of modeling. Nevertheless, any tree-based ensemble approach (Random Forest, XGBoost, CatBoost) is also resistant to multicollinearity by default, because it does not estimate coefficients but uses hierarchical feature splitting. The presence of Total_Rooms gives a consolidated form of the total property capacity which can potentially include valuable structural data that is not reflected entirely by the counts of individual rooms. Also, the analysis of feature importance proved that the addition of this feature did not impair the stability of the model or its prediction accuracy. Thus, Total_Rooms was kept as a non-trivial engineered attribute to increase the interpretability and predictive ability of the model.

Additionally, the categorical Age feature was converted into a numeric value (Age_numeric) by mapping each age category to its midpoint, enabling quantitative analysis. Such transformation was done to maintain the ordinal character of the Age variable because categories of property ages are inherently of a logical temporal sequence. Each category can be plotted to a midpoint to offer a sensible numerical estimate of the underlying continuous age domain to enable the model to reflect the progressive impact of age on property value. By doing so, the learning algorithms will be able to understand trends associated with age much better without introducing artificial categorical fragmentation. Equivalent midpoint encoding variants have been broadly used in real estate and tabular machine learning problems in which categorical ranges reflect continuous underlying variables.

Next, categorical features including Area, Age, New, Balcony, Furnished, and Parking were transformed using One-Hot Encoding. This process converts each category into a binary column (0/1), with the first category dropped to avoid redundancy. It should be noted that One-Hot Encoding was not applied for CBR, as this algorithm can handle categorical features natively without requiring explicit encoding. The resulting dataset contained all original numeric columns, newly engineered features, and the encoded categorical variables.

After making and encoding the features, feature scaling was done on all the numeric columns, which included both the original variables as well as the new ones. Utilizing Standard Scaler standardized each numeric feature to have a mean of zero and a unit variance. This made sure that variables in larger ranges (like Size_m^2) didn't have an unfair effect on the model. After the dataset was scaled, it was checked to make sure the change worked, and descriptive statistics showed that the numeric features had been successfully standardized.

Table 4 provides an overview of the principal features of the post-preprocessing final dataset, such as the count of records, composition of features, and statistical aspects of the target variable. The data demonstrates the variety of the prices of apartments, as there are different types of properties and market segments in Irbid. The lack of any missing values

and the well-defined set of features are evidence that the dataset is in its final form ready to be used in training and evaluating a machine learning model in the strongest way.

Table 4 Final dataset summary after preprocessing and feature engineering

Item	Value
Number of records	4928
Number of input features (before encoding)	12
Numerical features	7
Categorical features	5
Engineered features	1 (Age_numeric)
Target variable	Price_kJD
Target min (kJD)	7
Target max (kJD)	196.7
Target mean (kJD)	68.51
Target median (kJD)	61.4

3.5 Machine Learning Algorithms

In our present study, we have used three ensemble regression schemes including XGBR, RFR and CBR to predict the value of residential houses. These choice of methodological options were informed by their reported excellence in more recent empirical research on predictive analytics of real-estate data, their ability to approximate complex non-linear relationships and their ability to take both numeric and categorical covariates. With the help of ensemble methods, we seek to improve the process of generalization and reduce predictive error compared to single-model methods.

RFR is a group learning algorithm that builds a set of decision trees to forecast a target variable. The trees are trained on random sub-sample of observations and random sub-sample of predictor variables (features). A final prediction is obtained as an average of all the trees hence minimizing variance and increasing model stability. RFR was commonly used to prevent overfitting and enhance predictive performance, especially on high-dimensional and heterogeneous data [25]. The average of the results of all trees is taken to get the final prediction of a given sample as shown in eq.1:

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} f_m\,(x_i) \qquad (1)$$

where: $\hat{y}_i$= predicted value for sample x_i, $f_m(x_i)$= prediction of the m-th tree, M= total number of trees in the forest. Each tree is trained independently on a bootstrap sample of the data, and at each split, a random subset of features is considered. Unlike gradient boosting methods, RFR does not use a sequential additive objective function; instead, it relies on averaging to reduce variance, providing robustness against overfitting.

XGBR is an advanced gradient-boosting model whereby trees are built in a sequential manner. A new tree is learnt at every iteration to refine the errors that were not removed by the previous predictors. XGBR also includes regularization methods that prevent over-fitting and are efficient in allowed to interact with complex features, which make it a strong algorithm to predict numerical results with complex dependencies,

including housing prices [26]. The objective function of the XGB at the t-th training step is defined as shown in eq. 2:

$$\mathcal{L}^{(t)} = \sum_i l(y_i^{\text{pred}}, y_i^{\text{truth}}) + \sum_k \Omega(f_k) \qquad (2)$$

Where $l(\cdot)$denotes the loss function measuring the difference between the predicted value and the ground truth. The regularization term $\Omega(f_k)$controls the model complexity and is defined as eq. 3:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2 \qquad (3)$$

where $T$is the number of leaves in the k-th tree, ωrepresents the leaf weights, and $\gamma$and $\lambda$are regularization parameters that control the complexity of the model and help prevent overfitting.

CBR is a gradient-boosting model that specifically supports categorical predictors and thus does not require a lot of pre-processing to be performed. It uses methods like ordered boosting and target encoding in order to convert the categorical variables into numerically appropriate representations that reduce over-fitting. CBR has been known to be very predictive and stable especially when used on dataset with both numeric and categorical variables [27]. CBR follows the standard gradient boosting framework, where the prediction for an input sample $x_i$is expressed as an additive model of decision trees, eq. 4:

$$\hat{y}_i = \sum_{t=1}^{T} f_t\,(x_i) \qquad (4)$$

where $f_t$denotes the decision tree constructed at the t-th boosting iteration, and $T$is the total number of trees. The training objective of CBR is to minimize the following regularized loss function, eq. 5:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \qquad (5)$$

where $l(y_i, \hat{y}_i)$is the loss function, $y_i$is the ground truth value, and $\Omega(f_t)$is a regularization term that penalizes model complexity.

By combining all three ensemble methods, we capitalize on the synergistic benefits of better handling of feature interaction, reduced over-fitting and greater predictive robustness, and in combination are able to improve the overall quality and reliability of our residential property-price prediction project.

3.6 Machine Learning Algorithms Assessment

To measure the model performance, a long list of metrics was computed such as R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Median Absolute Error (MedAE), Mean Absolute Percentage Error (MAPE) and a 95 percent confidence interval of the prediction errors.

These metrics are defined as eq. 6 to eq. 10:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (7)$$

$$\text{MedAE} = \text{median}(| y_i - \hat{y}_i |) \quad (8)$$

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \frac{| y_i - \hat{y}_i |}{| y_i |} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (10)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the observed values, and n is the number of samples. Using these metrics ensures a rigorous assessment of each model's accuracy, reliability, and robustness.

3.7 Models Training and Evaluation

At this phase, the regression models were trained and tested. The target variable was set to Price kJD and two sets of features were prepared, one with One-Hot Encoding with RFR and XGBR and the original data with categorical columns with CBR, it is possible to use categorical variables as they are. Each model was divided into a training (80 percent) and testing (20 percent) dataset. Cross-validation was used to conduct the hyperparameter tuning to make sure that the model evaluation was robust and to minimize the possibility of variation in performance because of single data splitting. Particularly, GridSearchCV, and RandomizedSearchCV were applied with k-fold cross-validation on the training set, which enabled the models to be tested on more than one data partition. This method enhances the results in terms of reliability and generalizability of the chosen hyperparameters and reduces the chance of overfitting. Performance of the final models was then evaluated on the independent test set to have a non-biased estimate of predictive accuracy.

3.8 Hyperparameter Tuning and Model Evaluation

Experiments were performed on the hyperparameters of each regression algorithm to enhance performance of the models, as shown in Table 5. The test data were then used to make predictions using optimized models. During hyperparameter optimization, 5-fold cross-validation was employed within GridSearchCV and RandomizedSearchCV to systematically evaluate candidate parameter combinations. This ensures that the selected hyperparameters generalize well across different data subsets and enhances the robustness of the modeling process.

The hyperparameter values were chosen using the current best practices in literature and previous empirical experiments of ensemble learning on tabular data. The ranges that were adopted were adequate to represent the most significant parameters changes and were also computational with no redundant model complexity. The preliminary experiments suggested that the extension of the ranges did not provide significant performance gains as ensemble models like Random Forest, XGBoost, and CatBoost are usually stable in moderate parameter ranges. The chosen tuning ranges, thus, gave a good compromise between the quality of optimization, cost of computation, and generalization of the models.

Table 5: Final Chosen Hyperparameters

Model: Tuning Method	Parameter Ranges Tested	Final Chosen Hyperparameters
RFR: GridSearchCV	n_estimators: [100, 200, 300] max_depth: [None, 10, 20] min_samples_split: [2, 5] min_samples_leaf: [1, 2] max_features: ['sqrt', 'log2']	n_estimators: 100 max_depth: 10, min_samples_split:2, min_samples_leaf: 1, max_features: sqrt,
XGBR: GridSearchCV	n_estimators: [100,200,300] learning_rate: [0.01,0.05,0.1] max_depth: [3,5,7] subsample: [0.7,0.8,1] colsample_bytree: [0.7,0.8,1]	n_estimators: 200 learning_rate: 0.05, max_depth: 3, subsample: 0.8 colsample_bytree: 1,
CBR: RandomizedSearchCV	iterations: [200,500,700] learning_rate: [0.01,0.05,0.1] depth: [4,6,8] l2_leaf_reg: [1, 3, 5]	iterations: 500 learning_rate: 0.05 depth: 4 l2_ leaf_reg: 1,

4. Results and discussion

Table 6 summarizes the performance metrics of the three ensemble models (RFR, XGBR, and CBR). The models were evaluated using multiple metrics, including R^2, MAE, RMSE, MedianAE, MAPE, and the 95% confidence interval of prediction errors.

R^2 is the percentage that explains the target variable as explained in the model. CBR model got the maximum R^2 value (0.9031) followed by XGBR and RFR with values 0.9013 and 0.8940 respectively, which means that CBR explains the underlying variability in the data slightly better than the rest.

MAE is used to measure the mean values of the prediction errors with low values indicating high predictive ability. CBR had the minimum MAE (10.3803), that is why its predictions are much closer to the actual ones, whereas XGBR and RFR had a slightly higher MAE of 10.4641 and 10.7181 respectively.

RMSE with its weighting of big errors showed the least value of CBR (12.2184), and thus its effectiveness in reducing large prediction errors.

MAPE also confirms this tendency: CBR has a MAPE of 24.3846%, slightly better than XGBR (24.7106%), RFR (25.5099%), and, thus, proves a more accurate relative prediction. Even though the MAPE values of about 24-25 percent are far from zero, meaning that the error of prediction does exist, the given value is still comparable with the real-life tasks of predicting real estate, as the fluctuation of the prices depends on a great number of unmeasured variables, including accurate micro-location features, quality of neighborhood, market fluctuations, as well as specific features of the buyer. The pricing of real estate is a complex phenomenon that includes both quantifiable and unquantifiable variables that may not necessarily be reflected in listing information. Moreover, the data set was obtained based on actual listing as opposed to controlled experimental data which brings in natural variation. Nonetheless, the models attained a high R 2 value of more than 0.89 and this implies that most price variation has been effectively explained. Thus, the achieved MAPE values have a realistic and acceptable predictive performance in the application to investment analysis, price estimation, and urban planning support, and also reveal the possibility of further enhancement due to the addition of more location-specific and time-related features.

The 95% confidence interval (CI) offers valuable information on how most prediction errors can occur, with an estimate of the range on which most errors will fall that can be useful information on the reliability of models and realistic uncertainty. In the case of the CatBoost, the CI lies between some -20.35 to +21.75 kJD meaning that in most instances, the error in prediction should be within this range. This implies that the actual price of a property will be within a range of say about -21 kJD to +21 kJD of the price that the model is predicting.

Table 6: Model Performance Metrics After Hyperparameter Tuning

Metric	RFR	XGBR	CBR
R2	0.8940	0.9013	0.9031
MAE	10.7181	10.4641	10.3803
RMSE	12.7774	12.3288	12.2184
MedianAE	9.9902	10.0839	9.9017
MAPE	25.5099	24.7106	24.3846
CI_Lower	-21.9712	-20.8199	-20.3490
CI_Upper	24.3649	22.2082	21.7531

In order to further evaluate the predictive quality of CatBoost model, a scatter plot involving the predicted and actual apartment prices is provided in Fig. 3. It is shown that there is a good fit between the values that are predicted and those that are observed in the plot and a majority of the points are clustered around the diagonal line of reference that indicates perfect prediction. This means that the model is able to give precise and consistent predictions with varying price ranges.

The low variance of the points around the reference line attests the strength and the ability to generalize the CatBoost model. These small deviations in the higher price ranges are anticipated because of higher variability and the effect of some external variables that are not entirely represented in the data set like the exact location features and market

fluctuations. On the whole, the given visualization averts the quantitative performance measures presented in Table 6 and indicates the validity of the model to use in real-estate price forecasting.

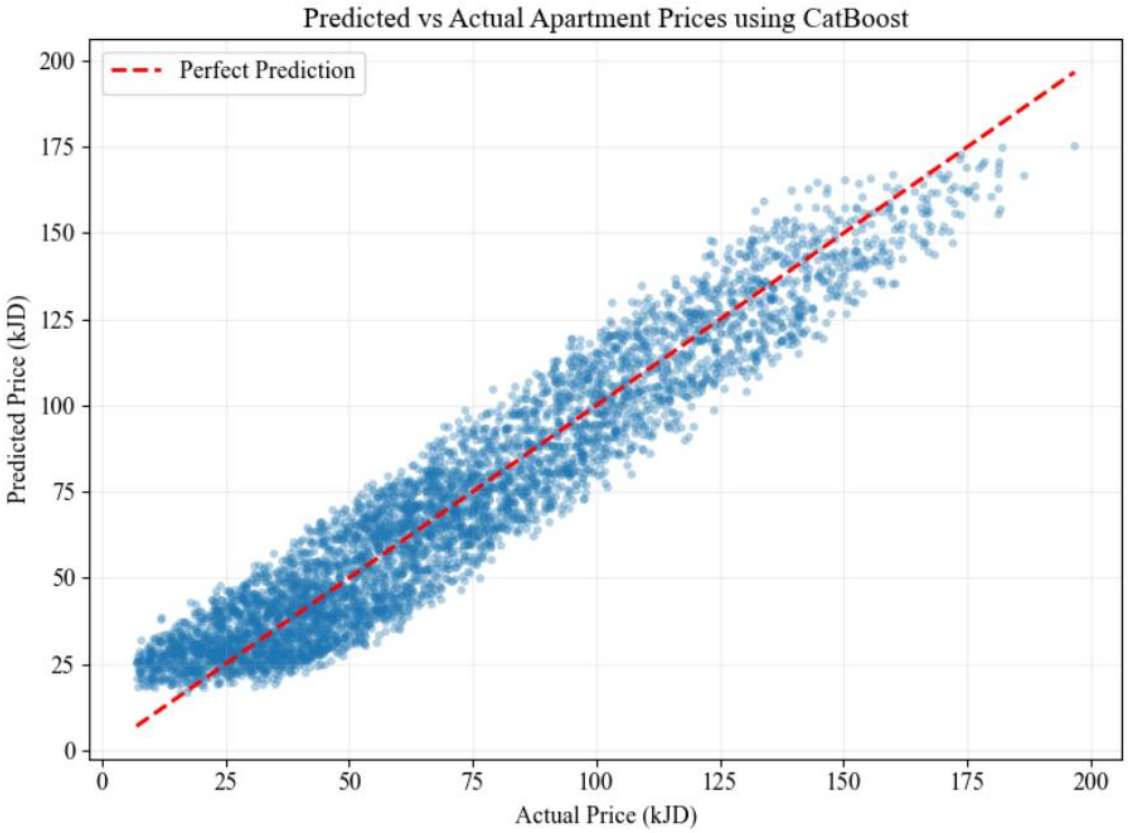

Fig. 3. Scatter plot of predicted versus actual apartment prices using the CBR regression model.

To practitioners like real estate investors, property developers, and urban planners, this period gives them a reasonable forecast of the uncertainty in predictions and can help them make sound decisions. A smaller confidence interval means that the model is more stable and reliable. The tightest interval of CatBoost is provided compared to the other models, which proves it to be more consistent and robust. These findings show that the model can be applied in real life situations where estimation of prices is needed to make planning, investment analysis and market analysis.

On the whole, all three models are effective, but CBR is more effective than the two in all measures of evaluation, which makes it acceptable to both investors who have access to the correct price estimates, and policymakers in their planning and regulatory decision-making.

To analyze the role of features in predicting the models, SHAP (SHapley Additive exPlanations) analysis was conducted on the most effective model, which was CBR. Fig. 4 shows the SHAP summary plot of the CBR model that shows the contribution of each

feature towards the predicted prices of apartments in Irbid. This visual representation gives a very clear depiction of feature significance and their impact to the model output that is easy to interpret.

The features are ranked by demerit in terms of their overall significance with the top features being the most influential features. The feature values are represented as color scale such that the larger values are represented by red color and smaller values are represented by blue color. The horizontal position indicates how great and which way the effect of the feature on the price being predicted is with a positive value pointing to an increased predicted price and a negative value pointing to a reduced predicted price. This discussion is very informative about the decision-making process of the model and enhances transparency and interpretability.

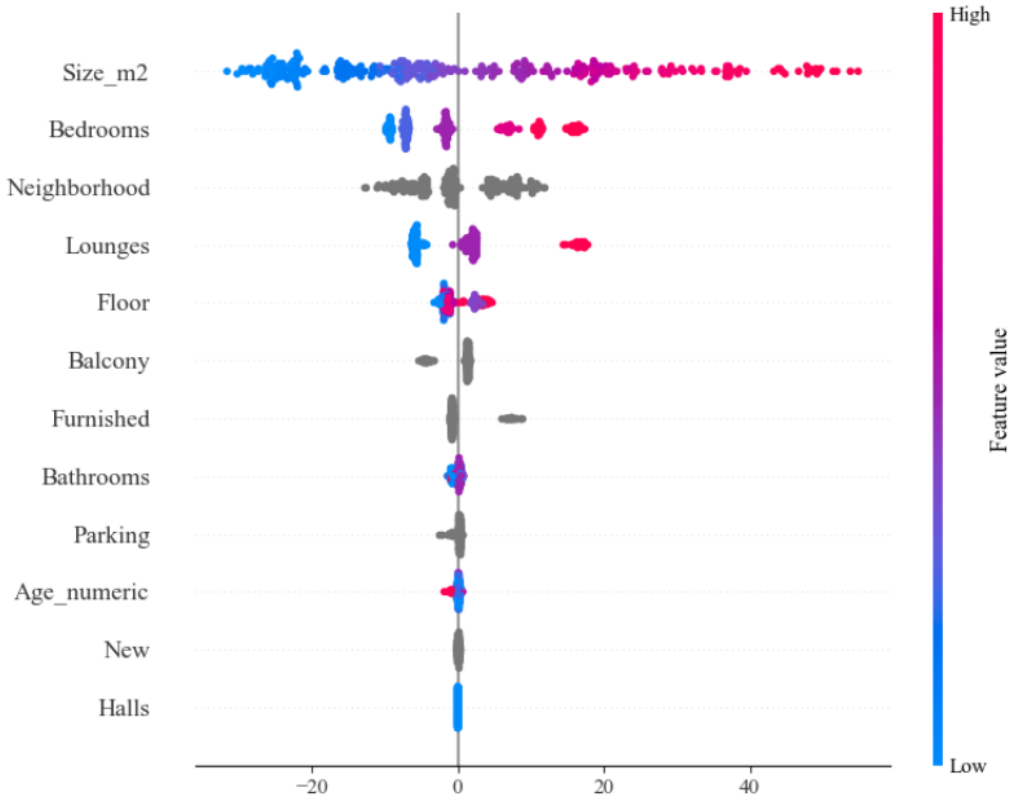

Fig. 4. SHAP Values Analysis

Through the analysis, property size (Size m 2) is the most significant attribute with the large property highly correlated with the property predicted price. The effect of Bedrooms and Lounges is also substantial which is to show that the more the functional and living space that a property has, the more market value it will have. The results illustrate the family specificities of the housing preferences and the space functional needs, which are inherent to the Irbid real estate market.

The geographical location is also a significant factor, where some places are beneficial in price forecasting as some are characterized by variations in demand, reachability and neighborhood appeal. Other characteristics like Furnished status, Age, Parking availability

and the presence of Balcony depict moderation in this relation and the attributes indicate that the structural and amenity attributes play a role in valuing properties. On the contrary, Halls and some age categories have comparatively smaller effects, which show that they have little impact on price change.

In general, the SHAP analysis proves that the apartment prices in Irbid are mostly determined by the size of structures, functional capacity, and location-related factors. These results are consistent with the actual market dynamics as well as offer valuable insights that can be used in making real estate investment decisions, urban planning, policy development, etc.

5. Conclusion and Future Work

The purpose of this paper was to gather and process real estate data in the city of Irbid in order to make a prediction of the price of apartments and houses and also narrate the situation under which price is determined. Through the assistance of extracted information on the different local real estate's websites, comprehensive data cleaning, exploratory analysis and interrelationship examination of the variables, we were in a position to come up with robust machine learning frameworks, i.e. RFR model, XGBR model and CBR model that can be employed to forecast the prices of the properties with a reasonable degree of precision. The results indicated that the CBR model would give the most decent results that could work with mixed types of data (numerical and categorical). Based on the SHAP analysis, property size, number of bedrooms, number of lounges and geographic location had the strongest effects on the price in Irbid, thereby demonstrating the dynamics of the local market and the logic of functionality that formed the prices.

The analysis might be refined in the future as additional neighborhood and environment factors, such as distances to schools, amenities, socio-economic environment and infrastructure, and time data may be included in the analysis to assess the price changes over time. Moreover, there is a possibility that the hybrid models or a combination of deep learning techniques and ensemble models, should be taken into account to make predictions more accurate and able to more thoroughly reflect the complex relationships among the properties features in a more appropriate manner. Overall, the research provides a systematic and quantitative context to the perception of the Irbid property market and offers decision-supportive tools to the investors, planners, and other stakeholders of the area.

References

[1] D. Sanfelici and L. Halbert, "Financial market actors as urban policy-makers: the case of real estate investment trusts in Brazil," Urban Geography, vol. 40, pp. 83–103, 2019, doi: 10.1080/02723638.2018.1500246.

[2] J. A. A. Numan and I. M. Yusoff, "Identifying the current status of real estate appraisal methods," Real Estate Management and Valuation, 2024, doi: 10.2478/remav-2024-0032.

[3] F. W. Hartono, Muljono, and A. Z. Fanani, "Improving the accuracy of house price prediction using CatB regression with random search hyperparameter tuning: A

comparative analysis," Advance Sustainable Science, Engineering and Technology, vol. 6, pp. 02403014–02403014, 2024, doi: 10.26877/asset.v6i3.602.

[4] H. Sharma, H. Harsora, and B. Ogunleye, "An optimal house price prediction algorithm: XGBoost," Analytics, vol. 3, pp. 30–45, 2024, doi: 10.3390/analytics3010003.

[5] M. Syahid Pebriadi and P. Negeri Banjarmasin, "House price prediction using the Random Forest algorithm on the RapidMiner application," Formosa Journal of Science and Technology, vol. 4, pp. 727–738, 2025, doi: 10.55927/fjst.v4i2.20.

[6] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş, "A novel hybrid house price prediction model," Computational Economics, vol. 62, pp. 1215–1232, 2023, doi: 10.1007/s10614-022-10298-8.

[7] Y. Wang and Q. Zhao, "House price prediction based on machine learning: A case of King County," in Proc. 7th Int. Conf. Financial Innovation and Economic Development (ICFIED), 2022, pp. 1547–1555, doi: 10.2991/aebmr.k.220307.253.

[8] P. Adzanoukpe, "Predicting house rental prices in Ghana using machine learning," Preprint, 2025, doi: 10.20944/preprints202412.1927.v1.

[9] M. Khosravi et al., "Performance evaluation of machine learning regressors for estimating real estate house prices," Preprint, 2022, doi: 10.20944/preprints202209.0341.v1.

[10] T. Quang, N. Minh, D. Hy, and M. Bo, "Housing price prediction via improved machine learning techniques," Procedia Computer Science, vol. 174, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[11] M. H. Hasan, M. A. Jahan, M. E. Ali, Y.-F. Li, and T. Sellis, "A multi-modal deep learning based approach for house price prediction," 2024. (No DOI available.)

[12] F. T. 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, p. 2209, 2024, doi: 10.3390/app14052209.

[13] Y. Fu, "A comparative study of house price prediction using linear regression and Random Forest models," Highlights in Science,

Engineering and Technology, vol. 107, pp. 96–103, 2024, doi: 10.54097/vcy5n584.

[14] W. Weng, "Research on house price forecast based on machine learning algorithm," BCP Business & Management, vol. 32, pp. 134–147, 2022, doi: 10.54691/bcpbm.v32i.2881.

[15] O. Lohith, A. Jha, and S. C. Tamboli, "Comparative analysis of Random Forest regression for house price prediction," International Journal of Engineering Research & Technology, vol. 11, pp. 2320–2882, 2023.

[16] S. C. K. Tekouabou et al., "AI-based machine learning methods for urban real estate prediction: A systematic survey," Archives of Computational Methods in Engineering, vol. 31, p. 1079, 2024, doi: 10.1007/s11831-023-10010-5.

[17] N. Yahya et al., "Predictive visual analytics for machine learning model in house price prediction: A case study," Open International Journal of Informatics, vol. 9, 2021.

[18] R. T. Mora-Garcia, M. F. Cespedes-Lopez, and V. R. Perez-Sanchez, “Housing price prediction using machine learning algorithms in COVID-19 times,” Land, vol. 11, p. 2100, 2022, doi: 10.3390/land11112100.

[19] M. I. Hussain, A. Munir, M. Mamun, S. H. Chowdhury, N. Uddin, and M. M. Hossain, "A Transparent House Price Prediction Framework Using Ensemble Learning, Genetic Algorithm-Based Tuning, and ANOVA-Based Feature Analysis," FinTech, vol. 4, no. 3, p. 33, Jul. 2025, doi: 10.3390/fintech4030033.

[20] M. A. Shbool et al., “Real estate decision-making: Precision in price prediction through advanced machine learning algorithms,” International Journal of Housing Markets and Analysis, 2025, doi: 10.1108/ijhma-01-2025-0004.

[21] R. Obeid, “Modeling real estate price volatility in Jordan: An analysis using the EGARCH model,” SSRN, 2025, doi: 10.2139/ssrn.5261230.

[22] W. T. Al-Sit and R. Al-Hamadin, "Real estate market data analysis and prediction based on minor advertisements data and locations’ geo-codes," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 3, pp. 4077–4089, May–Jun. 2020. doi: 10.30534/ijatcse/2020/235932020.

[23] M. A. Alwadi, “Fuel sales price forecasting using time series, machine learning, and deep learning models,” Engineering, Technology & Applied Science Research, vol. 15, pp. 22360–22366, 2025, doi: 10.48084/etasr.10348.

[24] N. Thalji, "Irbid Housing Dataset (2024–2025) for Apartment Price Prediction in Jordan," Zenodo Data Repository, 2026. doi: 10.5281/zenodo.18765634.

[25] M. Mamun, S. H. Chowdhury, M. M. Hossain, M. R. Khatun, and M. S. Iqbal, "Explainability enhanced liver disease diagnosis technique using tree selection and stacking ensemble-based random forest model," Informatics and Health, vol. 2, no. 1, pp. 17–40, 2025, doi: 10.1016/j.infoh.2025.01.001.

[26] X. Zhang, C. Yan, C. Gao et al., "Predicting Missing Values in Medical Data Via XGBoost Regression," J. Healthc. Inform. Res., vol. 4, pp. 383–394, 2020. doi: 10.1007/s41666-020-00077-1.

[27] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," Journal of Big Data, vol. 7, p. 94, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00366-0