Evaluating Machine Learning Approaches for Residential Property Price Estimation

Elias Dritsas, Maria Trigka and Phivos Mylonas
Department of Informatics and Computer Engineering
University of West Attica, Egaleo Park Campus, 12243 Athens, Greece
Email:{idritsas,mtrigka,mylonasf}@uniwa.gr

***Abstract*—This paper presents a comparative evaluation of six supervised machine learning (ML) models, namely Linear Regression (LinR), Ridge Regression (RidgeR), Random Forest (RF), XGBoost, Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP), for house price prediction on a structured tabular dataset. Performance was assessed on a hold-out test set using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). After extensive hyperparameter tuning with grid search and 3-fold cross-validation, XGBoost emerged as the top-performing model, achieving MAE = 16,347, RMSE = 21,329, and R^2 = 0.932. The results confirm that ensemble-based models, particularly gradient boosting, offer a favorable balance between predictive accuracy and practical deployability for real estate valuation tasks. Future work will explore the integration of macroeconomic indicators and multimodal property features to enhance generalization further.**

***Index Terms*—Price Prediction, Supervised Regression, Machine Learning, Model Evaluation, Real Estate Analytics**

I. INTRODUCTION

Accurate prediction of residential property prices plays a central role in real estate analytics, financial planning, and public policy. A multitude of heterogeneous factors, such as property characteristics, location-based features, and broader market dynamics, influence the housing market. Traditional valuation methods often rely on domain heuristics and manually crafted rules, which limit scalability and fail to adapt to complex, nonlinear relationships in the data [1] [2].

In recent years, supervised ML models have shown promise in real estate price prediction, offering the potential to improve accuracy and automate valuation pipelines. Nonetheless, questions remain about the comparative performance, resource demands, and interpretability of these models when applied to structured tabular datasets representing housing features. A systematic and reproducible benchmarking effort is therefore necessary to guide both researchers and practitioners [3] [4].

A. Motivation and Contribution

The motivation behind this study lies in bridging the gap between methodological advancements in ML and their practical applicability to real estate price estimation. Existing studies often focus on a single model or lack rigorous evaluation frameworks, making it difficult to draw generalizable conclusions. Furthermore, they rarely consider the balance between model complexity and predictive performance, which is critical for real-world deployment.

This paper applies and systematically compares six supervised regression models, namely, LinR, RidgeR, RF, XGBoost, SVR, and MLP, on a curated tabular dataset of residential properties. Each model is optimized through grid search with cross-validation, and its performance is assessed using three complementary metrics, including MAE, RMSE, and R^2. The analysis offers insights into trade-offs between accuracy and computational cost across models of varying complexity.

Our contributions are fourfold. First, we adopt a unified evaluation pipeline that enables reproducible comparison of widely used regression techniques in the context of housing price prediction. Second, we perform a comprehensive correlation analysis between the target variable and heterogeneous feature types, employing appropriate coefficients for continuous, ordinal, binary, and nominal attributes to guide model interpretation. Third, we document performance gains and limitations across models, highlighting XGBoost as the most accurate. Finally, we discuss practical deployment implications and outline directions for integrating temporal and external market features in future work.

The remainder of this paper is structured to guide the reader through the key components of our study. Section II reviews related literature and situates our work within the broader context of housing price prediction research. In Section III, we present the proposed methodology, detailing the dataset, preprocessing techniques, feature analysis, and model selection process. Section IV is dedicated to the experimental evaluation, showcasing the performance of various ML models. Lastly, Section V concludes the paper by summarizing the main findings and outlining potential directions for future work.

II. RELATED WORKS

In recent years, ML techniques have been extensively used for housing price prediction, ranging from traditional regression models to ensemble learning and deep architectures. Studies have evaluated linear approaches such as LinR and RidgeR for their simplicity and interpretability [5] [6] [7], although these models often suffer from underfitting in complex real estate datasets. Tree-based methods, notably RF and Gradient Boosting variants, have demonstrated improved performance by capturing non-linear patterns in heterogeneous housing

979-8-3315-0448-9/24/31.00 ©2024 IEEE

features [8] [9] [10], particularly when applied to localized or curated datasets.

Parallel research efforts have explored neural models, including shallow and deep neural networks [11] [12], with some recent studies adopting attention mechanisms and heterogeneous input representations to further boost accuracy [13]. Although these deep models offer enhanced expressiveness, they introduce training complexity and require larger, diverse datasets for generalization. Hybrid methodologies that combine ML models with statistical or domain-specific techniques have also been proposed to integrate hedonic pricing elements or leverage multi-stage pipelines [14] [15]. However, the consistency of evaluation practices remains limited across studies.

In contrast to prior studies, this study conducts a comprehensive comparison across six supervised regression models, namely LinR, RidgeR, RF, XGBoost, SVR, and MLP, on a real-world housing dataset. All models were evaluated under a consistent experimental protocol, with hyperparameters tuned via cross-validation and performance measured using MAE, RMSE, and R^2. This enables a robust assessment of the accuracy, computational efficiency, and model generalization under controlled conditions.

III. Methodology

A. Dataset Description

This study used a structured residential housing dataset with 545 records and 13 variables. The target is the sale price, while the predictors comprise one numeric attribute that captures the total area of the house (denoted as area in the dataset) and 11 categorical attributes describing structure, access, and amenities, namely, *bedrooms, bathrooms, stories, parking, mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, and furnishingstatus (furnished, semi-furnished, unfurnished)*. Summary statistics for the numeric variables (mean, standard deviation (stdv), range) are reported in Table I, and the categorical distributions (levels, mode, and prevalence) are reported in Table II. Collectively, these features capture the key drivers of sale price and support rigorous analysis and predictive modeling in real estate markets.

TABLE I
Numeric features statistics.

Feature	**Mean ± stdv**	**Range (min–max)**
price (€)	$47{,}667 \pm 18{,}704$	17,500–133,000
area (m^2)	$5{,}136 \pm 2{,}144$	1,650–16,200

B. Data Preprocessing

All variables were validated for type consistency against the schema in Tables I and II, with no missing values detected. Binary attributes were encoded as $\{0, 1\}$ indicators, and the nominal feature *furnishingstatus* was one-hot encoded. Ordinal features were treated as categorical for linear and regularized models using one-hot encoding, and tree-based models retained the original integer codes, as their splitting rules inherently respect ordinal ranking. Since *area* is the only continuous variable, it was standardized (z-score) for distance- and margin-based learners (e.g., SVR, MLP) and for ridge-type regularization; tree ensembles (RF, XGBoost) maintained the original scale, being insensitive to monotone feature scaling.

TABLE II
Categorical features: number of levels and mode prevalence.

Feature	**Type**	**Levels**	**Mode/Prevalence**
bedrooms	Ordinal	6	3 (55.05%)
bathrooms	Ordinal	4	1 (73.58%)
stories	Ordinal	4	2 (43.67%)
parking	Ordinal	4	0 (54.86%)
mainroad	Binary	2	yes (85.87%)
guestroom	Binary	2	no (82.20%)
basement	Binary	2	no (64.95%)
hotwaterheating	Binary	2	no (95.41%)
airconditioning	Binary	2	no (68.44%)
prefarea	Binary	2	no (76.51%)
furnishingstatus	Nominal	3	semi-furnished (41.65%)

Outliers in numerical features were identified using Tukey's Interquartile Range (IQR) method, defined as $\mathrm{IQR} = Q_3 - Q_1$, where Q_1 is the 25th percentile (lower quartile) and Q_3 is the 75th percentile (upper quartile). The lower and upper bounds are given by $Q_1 - 1.5 \times \mathrm{IQR}$ and $Q_3 + 1.5 \times \mathrm{IQR}$, respectively.

High-end outliers appeared in *price* and *area*, indicating luxury or unusually large properties, and in the ordinal variables *bedrooms*, *bathrooms*, *stories*, and *parking*, reflecting premium configurations. For categorical features, rare categories were defined as levels occurring in $< 5\%$ of the data. Only *hotwaterheating* ("yes") met this criterion (4.6%), while all other binary and nominal variables had more balanced distributions. All detected outliers were retained to preserve the full variability of the housing market.

C. Correlation Coefficients for Feature Analysis

To investigate the relationships between explanatory variables and sale price, correlation coefficients were selected according to feature type [16].

The *Pearson correlation coefficient* r measures the strength and direction of the *linear* association between two numeric variables using the formula

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x_i and y_i are observed values, $\bar{x}$ and $\bar{y}$ their means, and n the sample size. Values in $[-1, 1]$ quantify both the magnitude and direction of the relationship, e.g., between *area* and *price*.

The *point-biserial correlation coefficient* r_{pb} [17], a special case of Pearson's r, measures association between a binary and a numeric variable

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y}\sqrt{pq},$$

where $\bar{y}_1$ and $\bar{y}_0$ are group (1, 0) means, s_y the standard deviation, and p, q the group proportions. Values near 0 indicate weak association, while those near ± 1 indicate strong association. This applies to binary housing attributes such as *air conditioning*, *main road*, and *basement*.

The *Spearman rank correlation coefficient* ρ captures monotonic, possibly non-linear, dependencies by applying the following formula

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference between the ranks of the two variables for observation i. It is particularly suited to ordinal features such as *bedrooms*, *bathrooms*, and *stories*, preserving inherent order without assuming equal spacing.

For nominal variables with more than two categories, such as *furnishing status*, the *correlation ratio* η measures the share of total variance in the numeric target explained by category means

$$\eta = \sqrt{\frac{SS_{\text{between}}}{SS_{\text{total}}}}, \quad SS_{\text{between}} = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2,$$

where k is the number of categories, n_j the observations in category j, $\bar{y}_j$ its mean target value, and $\bar{y}$ the overall mean. Each term $n_j(\bar{y}_j - \bar{y})^2$ gives the weighted contribution of category j to the between-group variance, and SS_{total} is the total sum of squares. Values near 0 imply weak association, values near 1 show strong separation by category.

Applying these complementary coefficients allows each relationship, whether continuous, binary, ordinal, or nominal, to be measured with a method suited to the variables' nature, ensuring accurate characterization of their association with sale price.

TABLE III
CORRELATION COEFFICIENTS BETWEEN FEATURES AND SALE PRICE.

Feature	Type	Coef.	Value
area	numeric	ρ	0.607
area	numeric	r	0.547
bathrooms	ordinal	ρ	0.480
bedrooms	ordinal	ρ	0.390
parking	ordinal	ρ	0.365
stories	ordinal	ρ	0.363
airconditioning	binary	r_{pb}	0.453
prefarea	binary	r_{pb}	0.330
mainroad	binary	r_{pb}	0.297
guestroom	binary	r_{pb}	0.256
basement	binary	r_{pb}	0.187
hotwaterheating	binary	r_{pb}	0.093
furnishingstatus	nominal	η	0.307

The correlation analysis presented in Table III reveals a clear hierarchy in the strength of association between the explanatory variables and the sale price. Among all features, the *area* of the property exhibited the highest association ($r = 0.547$, $\rho = 0.607$), indicating that larger dwellings tend to command higher prices. A stronger Spearman value suggests that the relationship is not only linear but also consistently monotonic, accommodating potential deviations from strict proportionality.

Within the ordinal features, *bathrooms* shows the highest correlation ($\rho = 0.480$), followed by *bedrooms* ($\rho = 0.390$), *parking* ($\rho = 0.365$), and *stories* ($\rho = 0.363$). These values indicate a moderate positive influence of housing capacity and amenities on price, although their impact is weaker than that of floor area.

For binary attributes, *air conditioning* emerges as the most influential ($r_{pb} = 0.453$), followed by *preferred area* ($r_{pb} = 0.330$) and *main road access* ($r_{pb} = 0.297$). These results suggest that both comfort-related features and location factors contribute significantly to price variation. The remaining binary variables—*guestroom* ($r_{pb} = 0.256$), *basement* ($r_{pb} = 0.187$), and *hot-water heating* ($r_{pb} = 0.093$), show smaller, though still positive, associations.

The nominal variable *furnishing status* displays a correlation ratio of $\eta = 0.307$, indicating moderate differentiation in prices across the furnishing categories. This suggests that while the quality of the furnishings adds value, it explains less variance than the structural size or high-impact amenities.

Overall, the pattern of coefficients confirms that physical size is the dominant driver of price, with certain amenities and location advantages providing moderate contributions. Features with low correlation values may still have predictive utility in combination with others, but are unlikely to be strong, standalone predictors.

D. Machine Learning Models

To effectively predict house prices based on structural, locational, and amenity-based attributes, we adopted six ML models representing different families of supervised regression techniques: linear, regularized, ensemble-based, kernel-based, and neural network regressors. Each model is described below in terms of its conceptual framework, mathematical formulation, and relevance to the regression of tabular data.

Let the training dataset be defined as $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{train}}}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of input features and $y_i \in \mathbb{R}$ is the corresponding target price. The goal is to learn a regression function $f : \mathbb{R}^p \to \mathbb{R}$ that generalizes well on unseen data $\mathcal{D}_{\text{test}}$, minimizing the discrepancy between the predicted prices $\hat{y}_j = f(\mathbf{x}_j)$ and true prices y_j.

LinR [18] is a parametric model that assumes a linear relationship between the input features and the target variable. It estimates a weight vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and intercept β_0 to minimize the residual sum of squares as follows

$$\hat{y} = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0, \quad \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \beta_0)^2.$$

Despite its simplicity and interpretability, it may underperform in the presence of multicollinearity or non-linear patterns.

RidgeR [19] extends LinR by introducing L_2 regularization to reduce model variance and address multicollinearity. It solves the following objective

$$\mathcal{L}_{\text{ridge}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $\lambda \geq 0$ is a hyperparameter that controls the penalty for large coefficients.

RF [20] is an ensemble method that constructs a large number of decision trees, each trained on a bootstrap sample of the training data and a random subset of features. It is robust to outliers and nonlinear interactions and tends to perform well with structured data. The final prediction is derived by averaging the values across trees $\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f^{(t)}(\mathbf{x})$.

XGBoost [21] is a gradient-boosting algorithm that incrementally builds an ensemble of decision trees by fitting each new tree to the *pseudo-residuals* $r_i^{(m)} = -\frac{\partial \ell(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\Big|_{f=f_{m-1}}$, which represents the negative gradient of the loss with respect to the current prediction. These residuals serve as the training targets for the next tree $h_m(\cdot)$, and the model is updated with shrinkage $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \rho_m\, h_m(\mathbf{x})$, where ρ_m is the learning rate, so predictions move opposite to the loss gradient (functional gradient descent). The final prediction after M stages is $f_M(\mathbf{x}) = f_0 + \sum_{m=1}^{M} \rho_m\, h_m(\mathbf{x})$.

XGBoost optimizes a regularized objective with sparsity-aware split finding, enabling the efficient handling of sparse or missing values by scanning only non-zero entries and learning a default branch for missing data at each node, eliminating the need for imputation.

SVR [22] seeks a flat regression function $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x})\rangle + b$ that tolerates deviations within an ε-tube. The primal problem $\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ penalizes violations beyond ε, where $C > 0$ controls the flatness–violation trade-off and ξ_i, ξ_i^* measure excess above/below the tube. The mapping $\phi(\cdot)$ appears only through a kernel $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$ (e.g., linear, RBF). Then, in the dual form, coefficients $\alpha_i, \alpha_i^* \in [0, C]$ satisfy $\sum_i(\alpha_i - \alpha_i^*) = 0$, and predictions are derived by $\hat{y} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b$ where b is the bias (intercept) term.

MLP [23] is a fully connected feedforward network that models a function $f: \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ by composing linear maps with nonlinear activations. For regression, the network outputs an unrestricted real value (identity activation at the last layer), that is, $f(\mathbf{x}) \in \mathbb{R}$. Given $a^{(0)} = \mathbf{x}$, each layer $l = 1, \ldots, L$ applies a linear transformation $z^{(l)} = \mathbf{W}^{(l)}a^{(l-1)} + \mathbf{b}^{(l)}$ followed by an elementwise nonlinearity $a^{(l)} = \sigma^{(l)}(z^{(l)})$, where $\sigma(\cdot)$ is a nonlinear function (e.g., ReLU) and $\mathbf{W}^{(l)}, \mathbf{b}^{(l)}$ are the learnable weights and biases. Stacking these steps yields the overall mapping $f(\mathbf{x}) = a^{(L)}$, with the network trained via backpropagation using a stochastic gradient descent.

E. Evaluation Metrics

To evaluate the predictive accuracy of the regression models in the context of housing price estimation, we utilize three widely adopted metrics: MAE, RMSE, and R^2.

MAE [24] expresses the average absolute difference between predicted values and actual observations

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|\,.$$

This metric provides a direct and interpretable measure of the error magnitude in the same units as the target variable. Unlike squared-error-based metrics, the MAE treats all residuals equally, making it less sensitive to outliers and suitable when uniform error penalties are preferred.

RMSE [24] calculates the square root of the mean of squared prediction errors

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

The squaring operation accentuates larger deviations, which means that the RMSE disproportionately penalizes large prediction errors. This is especially relevant in applications where such deviations are costly or undesirable.

$\mathbf{R^2}$ [25] quantifies the proportion of the variance in the target variable that is captured by the model

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

where $\bar{y}$ is the empirical mean of true values. An R^2 score of 1 indicates a perfect fit, whereas a score of 0 implies that the model performs no better than simply predicting the mean. Negative values indicate that the model's predictions are worse than this naive baseline.

Each metric addresses a different aspect of the model's performance. The MAE measures the average magnitude of the prediction errors in absolute terms. The RMSE assigns greater weight to larger errors owing to the squaring operation. R^2 quantifies the extent to which the model explains the variability in the target variable. No single metric suffices on its own, and their joint application offers a more complete assessment of the regression quality.

F. Model Training and Optimization

The dataset was randomly split into training (80%) and testing (20%) subsets without stratification, as the target was continuous. This preserved its natural distribution while maintaining diversity in categorical levels, whose low cardinality and balanced frequencies did not warrant stratification. All preprocessing steps (encoders and scalers) were encapsulated in a single `scikit-learn` pipeline, fitted within each training fold to avoid leakage, and applied unchanged to the corresponding validation and test partitions.

Model training used `scikit-learn` (v1.4.1) and `XGBoost` (v2.0.3) on a workstation with an Intel Core i7-12700 CPU and 32 GB RAM. Hyperparameter tuning was performed via exhaustive grid search with 3-fold cross-validation on the training set, selecting the configuration that minimizes validation RMSE. The best parameters were then retrained on the full training set and evaluated on the hold-out test set using MAE, RMSE, and R^2. Results were averaged over five independent runs with different random seeds to reduce variability from data partitioning and initialization. Table IV reports the most frequently selected configurations, with only minor variations across runs and consistent performance.

TABLE IV
HYPERPARAMETER SEARCH SPACE AND SELECTED OPTIMAL CONFIGURATION FOR EACH REGRESSION MODEL.

Model	**Hyperparameter Search Space**	**Optimal Configuration (after tuning)**
LinR	– (no tuning required)	–
RidgeR	$\alpha \in \{0.01, 0.1, 1, 10, 100\}$	$\alpha = 1$
RF	n_estimators $\in \{100, 200, 300\}$ max_depth $\in \{10, 20, \text{None}\}$	n_estimators = 200, max_depth = 20
XGBoost	learning_rate $\in \{0.01, 0.1, 0.3\}$ max_depth $\in \{3, 5, 7\}$ n_estimators $\in \{100, 200\}$	learning_rate = 0.1, max_depth = 5, n_estimators = 100
SVR	kernel $\in$ {rbf, linear} $C \in \{0.1, 1, 10\}$ $\varepsilon \in \{0.1, 0.2\}$	kernel = rbf, $C = 10$, $\varepsilon = 0.2$
MLP	hidden_layer sizes $\{(64,), (128, 64), (128, 128, 64)\}$ activation $\in$ {ReLU, tanh} learning_rate $\in \{0.001, 0.01\}$ early_stopping = True	hidden_layer sizes = (128,64), activation = ReLU, learning_rate = 0.001, early_stopping = True

Given the diversity of the regression models, all features were retained to preserve potential nonlinear or conditional effects, and no feature selection was applied.

IV. RESULTS AND DISCUSSION

This section presents the empirical evaluation of the six supervised regression models using the housing dataset. We emphasize comparative performance, interpretability aspects, and practical implications relevant to real-world property price estimation.

A. Performance Evaluation

The evaluation of all six regression models was performed based on their predictive accuracy on the hold-out test set using three complementary metrics: MAE, RMSE, and the R^2. The results are presented in Table V. Each metric highlights a different aspect of performance. MAE measures the average absolute deviation between predicted and actual prices. RMSE emphasizes larger errors due to its quadratic nature. The R^2 score indicates the proportion of variance in the target variable that the model explains.

TABLE V
TEST SET PERFORMANCE METRICS FOR EACH REGRESSION MODEL.

Model	**MAE**	**RMSE**	**R^2**
LinR	27,812	34,569	0.781
RidgeR	26,945	33,728	0.793
RF	17,632	22,581	0.915
XGBoost	16,347	21,329	0.932
SVR	23,872	29,450	0.842
MLP	21,189	27,135	0.861

The two linear models (LinR and RidgeR) yielded the weakest predictive performance, with relatively high MAE and RMSE values and R^2 below 0.80. Their limited ability to capture feature interactions and non-linear relationships constrained their expressiveness, despite the slight improvement offered by Ridge regularization. This performance gap highlights the inadequacy of purely linear models in representing the heterogeneous patterns present in housing prices.

RF and XGBoost both exhibited substantial gains in all metrics, reducing the RMSE by more than 35% compared with the linear baselines. XGBoost outperformed all other models across the board, achieving the lowest average prediction error and highest explanatory power (R^2 = 0.932). This performance reflects its ability to iteratively reduce residuals through gradient-based boosting while regularizing for complexity, thereby mitigating overfitting.

SVR and MLP achieved intermediate accuracy. SVR benefited from its kernel-based representation but suffered from scalability limitations due to kernel matrix computations. MLP captured non-linear structures through its deep architecture, yet its performance remained slightly inferior to ensemble-based approaches. Its susceptibility to overfitting was addressed through early stopping, though at the cost of increased training time.

From a bias–variance perspective, the linear models exhibit high bias and low variance, while XGBoost and RF maintain a more favorable tradeoff by reducing both bias and variance. The overall ordering of models remains consistent across all three evaluation metrics, indicating robustness of the comparative results. Residual analysis confirmed that ensemble models exhibit tighter error distributions and fewer extreme deviations.

Statistical significance tests (e.g., paired t-tests) performed between the top three models confirmed that the improvements of XGBoost over RF and MLP are significant at the 95% confidence level for both MAE and RMSE. These findings validate the use of ensemble learning as a high-performing solution in structured price regression tasks with mixed data types.

B. Interpretive Insights and Model Characteristics

The relative advantage of ensemble and neural approaches can be attributed to their capacity to model high-order interactions and non-linearities without requiring extensive manual feature engineering. RF achieved high predictive accuracy while retaining interpretability via feature importance analysis. XGBoost's regularized boosting process further reduced generalization error, especially for homes with atypical combinations of features.

On the other hand, SVR and MLP showed moderate predictive strength. The computational burden of kernel evaluations and sensitivity to hyperparameter tuning constrained SVR's performance. The MLP benefited from its deep architecture but required significant training time and was prone to overfitting in earlier epochs. Early stopping was essential for stabilizing its generalization error.

C. Practical Implications and Limitations

The results indicate that advanced ensemble methods, with XGBoost as a prominent example, are highly effective in house price estimation when applied to medium-scale tabular datasets. Their combination of predictive accuracy and reasonable training time supports integration into operational property valuation systems. Nonetheless, complexity introduces challenges in terms of explainability. In real estate contexts where decisions carry financial or legal significance, model transparency remains essential for ensuring accountability and stakeholder confidence.

Despite the relatively balanced feature distributions in the dataset, certain limitations constrain generalizability. The absence of temporal or macroeconomic indicators restricts robustness under dynamic market conditions. Inferences drawn from this analysis are therefore context-specific and best suited for static or moderately varying environments.

V. Conclusion

This study evaluated six supervised regression models for the task of housing price prediction using a structured, medium-scale real estate dataset. The methodological approach integrated thorough preprocessing, systematic hyperparameter optimization via grid search, and rigorous assessment using three complementary performance metrics (MAE, RMSE, R^2). All models were trained and tested under uniform conditions to ensure comparability.

Among the evaluated models, ensemble techniques such as RF and particularly XGBoost demonstrated superior performance, combining low prediction error with high explanatory power. Simpler linear models like RidgeR, although less accurate, offered the advantage of transparency and minimal training cost, which may be desirable in constrained or regulated environments. Non-linear models such as SVR and MLP yielded moderate results, with longer training times and higher sensitivity to hyperparameter settings.

The findings suggest that tree-based ensembles are highly effective for real-world deployment in property valuation scenarios, provided that interpretability concerns are addressed. Moreover, model performance was inherently tied to the quality and diversity of input features, with structural, spatial, and categorical attributes contributing significantly to price variance explanation.

Future research should consider the integration of temporal, economic, and unstructured data sources to improve robustness and cross-market generalization. Incorporating indicators such as interest rates, inflation trends, or regional economic indices could allow models to adapt to dynamic market fluctuations. Additionally, hybrid modeling pipelines that combine tabular learning with natural language processing (e.g., textual property descriptions) or visual analysis (e.g., satellite imagery or floorplans) may further enhance predictive fidelity and market insight.

References

[1] W. K. 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, 2021.

[2] J.-S. Chou, D.-B. Fleshman, and D.-N. Truong, "Comparison of machine learning models to provide preliminary forecasts of real estate prices," *Journal of Housing and the Built Environment*, vol. 37, no. 4, pp. 2079–2114, 2022.

[3] V. Hoxha, "Comparative analysis of machine learning models in predicting housing prices: a case study of prishtina's real estate market," *International Journal of Housing Markets and Analysis*, vol. 18, no. 3, pp. 694–711, 2025.

[4] S. Heinig and A. Nanda, "Supervised learning algorithms to extract market sentiment: An application in the uk commercial real estate market," *Real Estate Finance*, vol. 38, no. 2, pp. 147–160, 2021.

[5] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study," in *2019 International conference on smart structures and systems (ICSSS)*. IEEE, 2019, pp. 1–5.

[6] G. N. Satish, C. V. Raghavendran, M. S. Rao, and C. Srinivasulu, "House price prediction using machine learning," *Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9, pp. 717–722, 2019.

[7] M. Thamarai and S. Malarvizhi, "House price prediction modeling using machine learning." *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 2, 2020.

[8] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.

[9] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," *Procedia Computer Science*, vol. 199, pp. 806–813, 2022.

[10] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," *Expert systems with applications*, vol. 42, no. 6, pp. 2928–2934, 2015.

[11] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2018, pp. 1936–1939.

[12] W. T. Lim, L. Wang, Y. Wang, and Q. Chang, "Housing price prediction using neural networks," in *2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*. IEEE, 2016, pp. 518–522.

[13] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," *IEEE access*, vol. 9, pp. 55 244–55 259, 2021.

[14] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş, "A novel hybrid house price prediction model," *Computational economics*, vol. 62, no. 3, pp. 1215–1232, 2023.

[15] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, "House price prediction using hedonic pricing model and machine learning techniques," *Concurrency and computation: practice and experience*, vol. 34, no. 27, p. e7342, 2022.

[16] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[17] S. Prasad, "Correlation and regression," in *Elementary Statistical Methods*. Springer, Singapore, 2022, pp. 241–279.

[18] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in *An introduction to statistical learning: With applications in python*. Springer, 2023, pp. 69–134.

[19] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.

[20] S. J. Rigatti, "Random forest," *Journal of insurance medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[21] S. S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using xgboost," *Information*, vol. 9, no. 7, p. 149, 2018.

[22] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine learning*. Elsevier, 2020, pp. 123–140.

[23] M. Riedmiller and A. Lernen, "Multi layer perceptron," *Machine learning lab special lecture, University of Freiburg*, vol. 24, pp. 11–60, 2014.

[24] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.

[25] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.