Karshi Multidisciplinary International Scientific Journal

www.kmisj.reapress.com

K. Multidiscip. Int. Sci. J. Vol. 3, No. 1 (2026) 41–61.

Paper Type: Original Article

A Comparative Analysis of Hedonic OLS and Random Forest Models for Apartment Price Estimation

Tarek Yalouli[1], Raimonda Dervishi[2,*], Robert Rogaczewski[3]

[1] Faculty of Economics, Commerce and Management Sciences, Badji Mokhtar University-Annaba, Annaba, Algeria; tarek.yalouli@univ-annaba.dz.

[2] Department of Mathematical Engineering, Faculty of Mathematical Engineering and Physical Engineering, Polytechnic University of Tirana, Tirana, Albania; raimondadervishi@yahoo.com.

[3] Faculty of Economic and Technical Sciences, State University of Applied Sciences in Konin, Konin, Poland; r.rogaczewski@gmail.com.

Citation:

Received: 14 August 2025 *Revised: 25 December 2025* *Accepted: 27 January 2026*	Yalouli, T., Dervishi, R., & Rogaczewski, R. (2026). A comparative analysis of hedonic OLS and random forest models for apartment price estimation. *Karshi Multidisciplinary International Scientific Journal, 3*(1), 41-61.

Abstract

This study investigates apartment price formation through a comparative assessment of econometric and Machine Learning (ML) models applied to a micro-level dataset of residential properties in Tirana, Albania. The research addresses the challenge of modelling housing prices in emerging real estate markets characterized by heterogeneous property attributes and spatial variation. To ensure comparability across dwellings of different sizes, the dependent variable is defined as price per square meter. The methodological framework combines a hedonic Ordinary Least Squares (OLS) model and a nonlinear Random Forest (RF) model, allowing the evaluation of both model interpretability and predictive performance. Elastic Net and Extreme Gradient Boosting (XGBoost) models are additionally employed as robustness benchmarks. Model performance is assessed using standard prediction accuracy measures, while variable effects and importance metrics are analysed to identify the main determinants of housing prices. The results reveal that location-related factors and structural housing characteristics constitute the dominant drivers of apartment values. Apartment size is negatively associated with price per square meter, whereas the number of bathrooms is positively and statistically significantly associated. The number of rooms becomes insignificant after controlling for other explanatory variables. Strong neighbourhood effects confirm substantial spatial heterogeneity in the housing market. The comparative analysis demonstrates that RF achieves superior predictive accuracy relative to the alternative models, highlighting the ability of nonlinear methods to capture complex relationships in housing price data. The findings contribute to the application of statistical and ML techniques in real estate valuation and provide evidence on the relative strengths of interpretable and predictive modelling approaches.

Keywords: Housing price modelling, Hedonic regression, Random forest, Real estate valuation, Machine learning, Spatial heterogeneity.

1 | Introduction

The housing market represents a central pillar of urban economies, given its strong linkages to household welfare, wealth accumulation, and broader urban development dynamics [1]. Apartments serve a dual function

Corresponding Author: raimondadervishi@yahoo.com

https://doi.org/10.22105/kmisj.v3i1.117

as both consumption goods and investment assets, making apartment pricing a critical issue not only for market participants but also for policymakers concerned with housing affordability and urban planning [2]. The increasing availability of granular housing data, combined with advances in Machine Learning (ML) techniques, has significantly transformed housing price analysis, enabling more flexible and accurate modelling approaches [3]. This transformation is particularly relevant in contexts characterized by significant market heterogeneity and complex price formation processes.

Within this context, Tirana provides a particularly compelling case for empirical investigation. As Albania's primary economic, administrative, and demographic hub, the city has experienced sustained post-socialist urban expansion [4] alongside the development of a dynamic residential real estate market [5]. This expansion has been characterized by rapid, often unplanned growth, making apartment valuation an important empirical issue for understanding price formation in a transitioning urban environment [6]. Existing evidence suggests that apartment prices in Tirana are shaped not only by structural dwelling characteristics but also by location and neighborhood attributes, a finding consistent with the broader housing literature on spatial heterogeneity [7]. In Tirana, where neighborhoods differ substantially in attractiveness, accessibility, and market prestige, understanding the relative role of these factors is especially important [8].

Despite the relevance of this topic, the empirical literature on housing price determinants in Albania remains sparse and methodologically limited, particularly in terms of micro-level analysis and the application of advanced predictive techniques, relative to the broader international literature. The available studies for Tirana have mainly focused on hedonic regression approaches and are relatively few in number, while micro-level evidence based on apartment-level observations is still limited [9]. At the same time, the recent international literature has moved increasingly toward comparing conventional hedonic models with more flexible ML approaches, showing that non-linear methods often improve predictive performance while raising new questions about interpretability [10]. These findings reveal a critical gap in the Albanian context; there is still limited evidence on whether the standard hedonic framework is sufficient for the Tirana apartment market, or whether non-linear models can provide a measurable advantage in out-of-sample prediction.

This paper addresses that gap by constructing a micro-level dataset of apartment listings in Tirana from online real estate advertisements, focusing on price per square meter as the main dependent variable. The use of price per square meter allows a more meaningful comparison across apartments of different sizes—a particularly relevant consideration in heterogeneous urban housing markets. This study pursues two complementary objectives. First, it evaluates the explanatory capacity of a hedonic regression framework for modelling apartment prices per square meter. Second, it examines whether nonlinear ML algorithms provide measurable improvements in predictive performance relative to conventional econometric specifications. Accordingly, the study addresses the following research question: How effectively can linear and nonlinear modelling frameworks explain and predict apartment prices per square meter, and what is the relative contribution of structural and locational variables within these models? To answer this question, the study adopts a comparative modelling framework that integrates both econometric and ML approaches. A hedonic Ordinary Least Squares (OLS) model is employed as a benchmark linear specification, while a Random Forest (RF) model is used to capture nonlinear relationships and higher-order interactions among explanatory variables. Elastic Net and Extreme Gradient Boosting (XGBoost) are further incorporated as complementary modelling techniques to evaluate the robustness and consistency of the empirical findings across alternative statistical learning frameworks.

This modelling strategy enables a systematic comparison between interpretable parametric models and flexible data-driven algorithms, providing insights into both explanatory performance and predictive accuracy [11]. It is consistent with recent research showing that linear hedonic models remain useful for economic interpretation, whereas tree-based ML methods frequently deliver stronger predictive performance in real estate applications [12]. Importantly, in this study, "prediction" refers to out-of-sample prediction on a held-out test sample rather than time-series forecasting of future housing prices. The analysis is therefore primarily explanatory and predictive in nature, rather than causal.

The empirical analysis is guided by three testable hypotheses:

I. H_1: Apartment size, number of bathrooms, and neighborhood significantly explain apartment prices per square meter in Tirana.

II. H_2: Neighborhood effects account for a substantial share of price variation, reflecting the importance of location in the urban housing market.

III. H_3: Nonlinear ML models achieve lower prediction errors than conventional linear hedonic specifications in out-of-sample evaluation.

These hypotheses reflect the dual explanatory and predictive purpose of the study.

The main contributions of this paper, addressing both empirical and methodological gaps in the literature, are as follows:

I. It provides a comparative evaluation of linear econometric models and nonlinear ML algorithms for residential property valuation.

II. It develops and analyses a micro-level dataset of apartment listings from Tirana as an empirical application of advanced modelling techniques.

III. It examines the relative explanatory and predictive performance of hedonic OLS, RF, Elastic Net, and XGBoost models within a unified analytical framework.

IV. It contributes to the growing literature on statistical learning and housing price modelling by assessing the trade-off between model interpretability and predictive accuracy.

2|Literature Review

This section reviews key theoretical and empirical contributions to housing pricing, with a focus on hedonic pricing models and emerging ML approaches. It also highlights key research gaps that motivate this study, particularly in the context of developing housing markets such as Albania.

2.1|Hedonic Pricing Models in Housing Markets

Hedonic pricing models remain the dominant framework for analyzing housing prices because they conceptualize residential property as a bundle of attributes whose implicit values are capitalized into market prices [13]. Recent reviews confirm that the hedonic approach continues to be widely used in housing research due to its strong economic interpretability and its ability to isolate the contribution of structural, locational, and environmental characteristics to property values [14]. Importantly, the contemporary literature has moved beyond simple linear specifications by incorporating richer spatial information, accessibility indicators, and more flexible functional forms, reflecting the growing recognition that housing markets are highly heterogeneous and often characterized by non-linear relationships [15].

Recent empirical work confirms that hedonic models remain highly informative for identifying the economic determinants of house prices. Rey-Blanco et al. [15] show that improving the measurement of accessibility substantially enhances the explanatory power of hedonic models, while Marinković et al. [16] find that demographic, labour-market, and infrastructural indicators significantly shape urban housing prices in Serbia. These studies reinforce the idea that, even in developing or transition contexts such as Albania, housing values are jointly determined by dwelling-specific characteristics and broader urban conditions.

In the Albanian context, however, the empirical literature remains limited. Thanasi [5] provides one of the earliest apartment-level hedonic analyses for Tirana and shows that location and structural features both matter for apartment valuation. Related Albanian work includes hedonic modelling efforts that further support the relevance of structural housing attributes in price formation [16]. More recently, Nurja et al. [9] extended the evidence using apartment-level data and similarly found that housing attributes such as living area and internal composition are relevant for price formation. However, these studies share important

methodological limitations: They rely exclusively on linear regression models and do not explore the potential gains from non-linear approaches.

2.2 | Location and Neighborhood Effects in Real Estate Prices

A central finding in the housing literature is that prices are shaped not only by the internal characteristics of the dwelling but also by its surrounding neighborhood environment. Recent studies emphasize that neighborhood effects operate through multiple channels, including accessibility, amenities, green space, visual quality, density, and public services [17]. Consequently, location remains one of the most persistent and theoretically important dimensions for evaluating housing values, with recent work focusing on measuring location more accurately [18].

Empirical studies provide strong support for the importance of neighborhood-level variation. Zhang and Miller [19], using fine-scale housing data and explainable Artificial Intelligence (AI), find that neighborhood factors account for a substantial share of housing price variation and uncover important non-linear effects of amenities and neighborhood quality. Similarly, Rey-Blanco et al. [15] show that better accessibility measures improve both regression-based and RF models, highlighting the central role of location in price formation. Dou et al. [20] also conclude that accessibility-related variables, especially distance to the central business district and functional access to urban services, are among the most important drivers of housing prices.

The recent literature also documents the contribution of specific neighborhood amenities. Dou et al. [20] show that urban green space is positively associated with housing values, while other studies report that neighborhood greenery and service accessibility generate measurable price premiums, although the effects may be nonlinear and threshold-dependent [21]. These findings are especially relevant for urban housing markets like Tirana, where neighborhood prestige, service access, and environmental quality differ substantially across areas. Given Tirana's rapid and often uneven urban development, this line of research supports the expectation that neighborhood effects should explain an important share of apartment price variation.

2.3 | Machine Learning and Statistical Learning Approaches for Housing Price Modelling

The recent housing literature has increasingly adopted ML methods to improve predictive accuracy and to capture non-linearities and interactions that are often difficult to model in standard hedonic regressions [22]. A recent review by Anelli et al. [23] shows that ML has become an important methodological direction in real estate research, especially for valuation, prediction, and mass appraisal. Likewise, Choy and Ho [24] conclude that ML models generally offer stronger predictive performance than traditional hedonic regressions, although the latter remain valuable for economic interpretation.

Among ML methods, tree-based algorithms appear to be particularly prominent in real estate price prediction [25]. Empirical evidence suggests that these approaches can outperform traditional econometric models, especially in contexts characterized by complex and non-linear relationships between explanatory variables and property prices [26]. More recent studies reinforce this finding. For instance, AI models have been shown to outperform conventional hedonic pricing approaches in terms of appraisal accuracy, with RF demonstrating strong robustness across different variable specifications [27]. Ensemble-based valuation models have also been found to further enhance predictive performance [28]. Importantly, RF offers a balanced trade-off between predictive accuracy and interpretability, particularly when combined with explainability tools [29].

Importantly, the existing literature does not suggest that the traditional hedonic approach should be replaced entirely by ML methodologies. Rather, recent studies highlight the complementary role that both approaches offer. Specifically, hedonic OLS models are highly effective for estimating the marginal impact of individual housing attributes, providing interpretable coefficients that are essential for economic analysis [30]. ML methodologies, on the other hand, are more effective for improving forecast accuracy Zhang and Abdullah

[31] and for modelling non-linear relationships that are difficult to capture with linear specifications [32]. This complementary approach is particularly relevant for the present study, which aims not only to identify the primary drivers of apartment prices per square meter in Tirana but also to assess whether non-linear modelling improves predictive accuracy. By combining both methodologies, the study leverages the interpretability of OLS with the flexibility of ML.

The reviewed literature demonstrates that housing price modelling increasingly relies on the complementary use of econometric and ML techniques. While hedonic regression models remain the standard framework for interpreting the contribution of housing attributes, their explanatory structure may be insufficient to capture complex nonlinear relationships present in real estate data. ML algorithms offer greater flexibility and often achieve superior predictive performance, although at the cost of reduced interpretability. Consequently, recent research has shifted from treating these approaches as competing alternatives toward evaluating them within a unified modelling framework. Despite this methodological development, empirical evidence from emerging housing markets remains limited. This gap motivates the present study, which applies and compares linear and nonlinear modelling approaches using apartment-level data from Tirana.

3 | Methodology

3.1 | Data and Variables

The data consists of online apartment listings in Tirana, collected from the Çelësi real estate platform in 2025. The dataset includes information on listing price, apartment size, number of rooms, number of bathrooms, and neighborhood location. In some cases, additional functional characteristics are also reported, but the core analysis relies only on variables that are available with relatively consistent quality across the sample.

The dependent variable is the natural logarithm of apartment price per square meter, log (price), where price denotes the listing price in euros per square meter. Using the price per square meter provides a more meaningful basis for comparing apartments of different sizes. The logarithmic transformation offers several advantages. First, it reduces the influence of extreme values. Second, it helps mitigate the skewness of the price distribution. Third, it allows the coefficients of linear models to be interpreted approximately in percentage terms.

The main explanatory variables used in the analysis are:

I. Area(m2) – apartment size in square meters

II. Room – number of main rooms

III. Bathroom – number of bathrooms

IV. Neighborhood – a neighborhood in which the apartment is located

From the neighborhood, a new categorical variable is constructed by grouping neighborhoods with fewer than 10 listings into the category "other". This choice avoids categories with very few observations and improves the statistical stability of the estimates.

The variable Floor is excluded from the baseline specification because it contains a high proportion of missing values and would substantially reduce the effective sample size. For the same reason, additional characteristics such as elevator, parking, ownership certificate, and furnishing are not included in the core specification, as they are reported incompletely and inconsistently across listings.

3.2 | Hedonic Ordinary Least Squares Model

The empirical analysis begins with a benchmark hedonic model estimated using OLS, which serves as the primary linear specification for explanatory analysis. The dependent variable is the natural logarithm of apartment price per square meter, expressed in euros. This transformation is used because housing prices are

typically skewed and because it makes the estimated coefficients easier to interpret in approximate percentage terms [33].

The baseline specification is estimated using OLS and takes the following form:

$$\log(\text{price}_i)=\beta_0+\beta_1\text{area}_i(m^2)+\beta_2\text{bathroom}_i+\gamma_n\text{neighbourhood}_i+\varepsilon_i \quad (1)$$

This model estimates the association of apartment size, number of bathrooms, and location with price per square meter. Its role is to provide a simple starting point for the empirical analysis and to serve as an initial benchmark against which richer specifications can be compared. The inclusion of neighbourhood fixed effects is intended to control for spatial heterogeneity in apartment prices, given that location is expected to be one of the main determinants of housing values.

A more comprehensive hedonic specification is then estimated by adding the number of rooms as an additional structural characteristic. The extended OLS model takes the following form:

$$\log(\text{price}_i)=\beta_0+\beta_1\text{area}_i(m^2)+\beta_2\text{room}_i+\beta_3\text{bathroom}_i+\gamma_n\text{neighborhood}_i+\varepsilon_i \quad (2)$$

where γ_n denotes neighborhood fixed effects, and ε_i is the error term. A more parsimonious baseline specification excluding the number of rooms is also estimated as a preliminary benchmark. However, the extended OLS model is treated as the main empirical specification because it provides a richer representation of apartment characteristics while retaining direct coefficient interpretability. This combination makes it the preferred model for substantive economic discussion [33].

Unlike nonlinear ML algorithms, the OLS framework provides directly interpretable parameter estimates and facilitates the assessment of the conditional association between explanatory variables and apartment prices. Consequently, OLS serves as the primary explanatory model, while ML methods are employed to evaluate predictive robustness and the extent to which nonlinear structures improve model performance.

3.3 | Random Forest Model

To evaluate nonlinear modelling performance, the study additionally estimates an RF model. RF is an ensemble learning algorithm that constructs a large number of decision trees using bootstrap samples and randomly selected subsets of predictors, thereby improving predictive accuracy and reducing overfitting relative to a single decision tree [34]. Unlike linear regression models, RF does not require the specification of a predefined functional form and can accommodate complex nonlinear relationships and interaction effects among explanatory variables.

This model is estimated using the same set of explanatory variables as the extended OLS specification, namely apartment size, number of rooms, number of bathrooms, and neighbourhood group. The model is fitted on the training sample, with hyperparameters selected by minimizing Out-of-Bag (OOB) error. The tuning exercise focuses on three parameters: the number of predictors randomly evaluated at each split, the minimum terminal node size, and the sampling fraction. The preferred specification uses three predictors at each split, a minimum terminal node size of 5, and a sampling fraction of 0.7. OOB error thus serves as the standard internal validation criterion for model selection [34].

RF serves two purposes in this study. First, it functions as the main nonlinear benchmark against which the predictive performance of the linear hedonic model is evaluated. Second, it is used to derive variable importance measures. These importance scores are interpreted as indicators of each variable's relative contribution to predictive performance rather than as causal effects. Thus, RF is used to complement the OLS findings by showing which variables matter most in prediction when no linear functional form is imposed [34].

3.4 | Model Evaluation Strategy

3.4.1 | Train-test split

To evaluate out-of-sample predictive performance, the data are divided into training and test sets using an 80/20 split. The training set is used to estimate model parameters and tune hyperparameters, while the test set is reserved for final model evaluation. This separation helps avoid overly optimistic performance estimates and is standard practice in predictive modelling [35], [36].

For the ML models, hyperparameter tuning is conducted within the training sample using cross-validation. Elastic Net is tuned using 5-fold cross-validation over a grid of α values, with λ selected automatically for each specification. XGBoost is also tuned using 5-fold cross-validation, and the optimal number of boosting rounds is selected using early stopping. This tuning strategy ensures that test-set performance reflects true out-of-sample generalization rather than tuning on the evaluation sample [35], [36].

3.4.2 | Performance metrics

Model performance is assessed on the test set using several standard regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). In addition, the Mean Absolute Percentage Error (MAPE) will be reported to facilitate interpretation on the original price scale. RMSE places greater weight on large errors, while MAE provides a more direct measure of average absolute deviation. Reporting multiple metrics provides a more complete view of predictive performance [35], [36].

The purpose of this evaluation strategy is not to replace the inferential role of the hedonic OLS model, but to assess whether more flexible models can improve prediction and whether the substantive conclusions remain stable across model classes. In other words, predictive performance is used here as a complementary criterion to explanatory interpretation.

3.5 | Diagnostic Tests for the Ordinary Least Squares Model

3.5.1 | Multicollinearity

Potential multicollinearity among the regressors is assessed using Variance Inflation Factors (VIFs). VIF-based diagnostics are widely used in regression analysis to evaluate whether the variance of estimated coefficients is being inflated by linear dependence among predictors. At the same time, the interpretation of VIF thresholds requires caution, since no single cutoff is universally appropriate across all empirical contexts [37].

In this study, VIFs are used as a diagnostic tool rather than as a mechanical decision rule. Their role is to assess whether the explanatory variables in the hedonic specification are so strongly correlated that coefficient estimates become unstable or difficult to interpret. This assessment is particularly relevant because apartment size, number of rooms, and number of bathrooms may be conceptually related [37].

3.5.2 | Heteroskedasticity

The presence of heteroskedasticity is tested using the Breusch–Pagan test, a standard diagnostic procedure for examining whether the error variance depends systematically on the regressors [38].

If heteroskedasticity is detected, the statistical inference for the OLS model is based on heteroskedasticity-robust standard errors. In particular, robust covariance estimation follows the general logic of heteroskedasticity-consistent inference developed by Breusch and Pagan [39]. This approach ensures that hypothesis tests and confidence intervals remain valid even when the homoskedasticity assumption is violated.

3.6 | Robustness Checks

To assess whether the main findings are sensitive to model choice, two additional ML models are estimated as robustness checks: Elastic Net and XGBoost. These models are not treated as the main empirical

specifications for interpretation. Instead, they are used to test whether the results obtained from OLS and RF remain broadly consistent when alternative linear and nonlinear predictive structures are considered.

3.6.1 | Elastic net model

Elastic Net is a penalized linear regression method that combines the L1 penalty of the Lasso with the L2 penalty of Ridge regression. By integrating both regularization approaches, Elastic Net performs coefficient shrinkage while retaining the ability to select relevant predictors and stabilize estimation in the presence of correlated variables [40].

In this study, Elastic Net is estimated using the same set of explanatory variables as the extended OLS specification. The model serves as a linear benchmark that evaluates whether coefficient regularization improves predictive performance and model stability relative to conventional least squares estimation. Because Elastic Net preserves the linear structure of the hedonic framework, it provides a useful intermediate comparison between the fully interpretable OLS model and more flexible machine-learning algorithms.

The inclusion of Elastic Net also allows assessment of whether potential improvements in prediction originate from regularization alone or from the ability of nonlinear methods to capture complex interactions and non-proportional relationships among housing characteristics. Consequently, the model contributes to a clearer understanding of the relative importance of linear regularization versus nonlinear learning in residential property valuation.

3.6.2 | XGBoost model

XGBoost is a gradient boosting algorithm based on decision trees that constructs an ensemble of weak learners in a sequential manner, with each new tree aiming to reduce the prediction errors of the previous ensemble. Owing to its flexibility, computational efficiency, and strong predictive performance, XGBoost has become one of the most widely applied machine-learning algorithms for structured tabular data [41].

Consistent with the other model specifications, XGBoost is estimated using apartment size, number of rooms, number of bathrooms, and neighborhood group as explanatory variables. Categorical variables are converted into a numerical design matrix prior to model estimation. Hyperparameter tuning is performed using 5-fold cross-validation, considering key parameters such as learning rate, maximum tree depth, minimum child weight, row subsampling, and column subsampling. Early stopping is implemented to identify the optimal number of boosting iterations and to reduce the risk of overfitting [41].

XGBoost is included as an advanced nonlinear benchmark model capable of capturing complex interactions, nonlinear relationships, and heterogeneous effects among housing characteristics. In combination with RF, it allows evaluation of whether flexible machine-learning approaches provide meaningful improvements in predictive performance relative to linear specifications. The comparison between XGBoost, RF, Elastic Net, and OLS, therefore, offers a comprehensive assessment of the extent to which nonlinear patterns contribute to housing price formation in the Tirana apartment market.

3.7 | Explanatory and Predictive Modelling Framework

The interpretation of results follows a distinction between explanatory and predictive modelling objectives. The extended OLS specification serves as the primary explanatory model because its estimated coefficients provide a transparent assessment of the conditional association between apartment characteristics and price per square meter. This role is consistent with the traditional application of hedonic regression in housing market analysis [33].

In contrast, RF, Elastic Net, and XGBoost are employed to evaluate predictive robustness and model performance under alternative statistical learning frameworks. These models are not interpreted in terms of causal effects; instead, they are used to assess whether the patterns identified by the OLS specification remain stable when nonlinear relationships, interactions, and regularization mechanisms are introduced. For RF,

variable importance measures are interpreted as indicators of relative predictive contribution rather than economic effect sizes.

This framework enables a systematic comparison between interpretable econometric models and flexible ML algorithms. Consequently, the analysis provides both an economically interpretable baseline and an evaluation of predictive performance across alternative modelling approaches.

4 | Results

4.1 | Descriptive Statistics and Exploratory Data Analysis

The analytical sample consists of 1,726 apartment listings from Tirana. *Table 1* reports the main descriptive statistics for the variables used in the analysis.

Apartment prices average EUR 1,933 per square meter, with a median of EUR 1,826, indicating a moderately right-skewed distribution. The difference between the mean and median values suggests the presence of moderate positive skewness, motivating the logarithmic transformation applied in subsequent modelling stages. Apartment size averages 93.6 m², with a median of 93 m². The mean number of bathrooms is 1.35, while the mean number of main rooms is 1.74. The floor variable contains a substantial number of missing observations and is therefore excluded from the core empirical specification.

Table 1. Descriptive statistics of variables.

Variable	Mean	Median	SD	Min	Max	N
Price (EUR/m2)	1933.0	1826.0	637.0	725.0	4930.0	1726
log(price) (EUR/m2)	7.52	7.51	0.321	6.59	8.50	1726
Area (m2)	93.6	93.0	33.8	35.0	294.0	1726
Room	1.74	2.0	0.626	1.0	3.0	1607
Bathroom	1.35	1.0	0.523	1.0	4.0	1710
Floor	5.59	5.0	4.68	0.0	58.0	408

Fig. 1 presents the distribution of apartment prices per m², while *Fig. 2* presents the distribution of the log-transformed price variable. The distribution exhibits moderate right skewness, with a concentration of observations at lower price levels and a long upper tail. To address this skewness and improve model performance, the dependent variable is transformed using the natural logarithm in the subsequent econometric analysis. *Fig. 3* plots apartment price per square meter against apartment size, together with a fitted linear trend. The visual pattern suggests a weak negative association between apartment size and price per square meter, although the relationship does not appear perfectly linear.

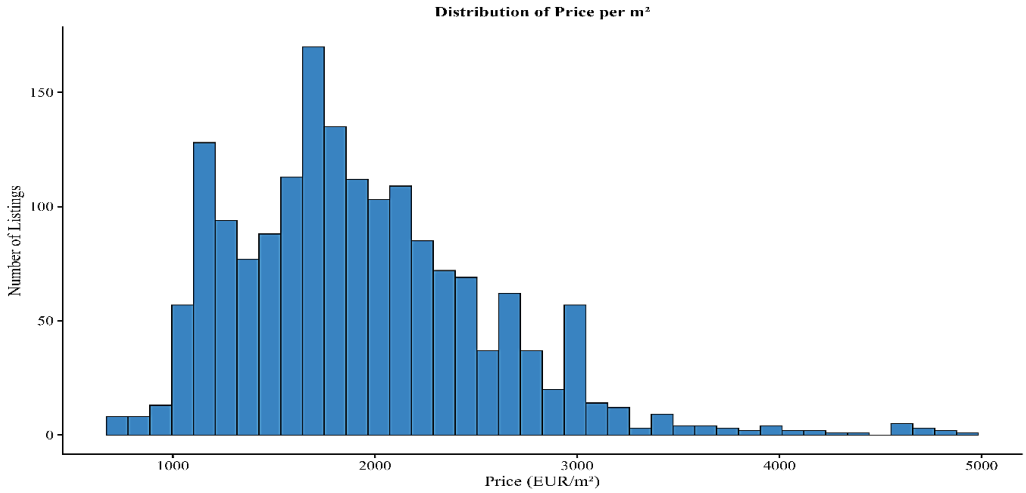

Fig. 1. Distribution of apartment prices per square meter.

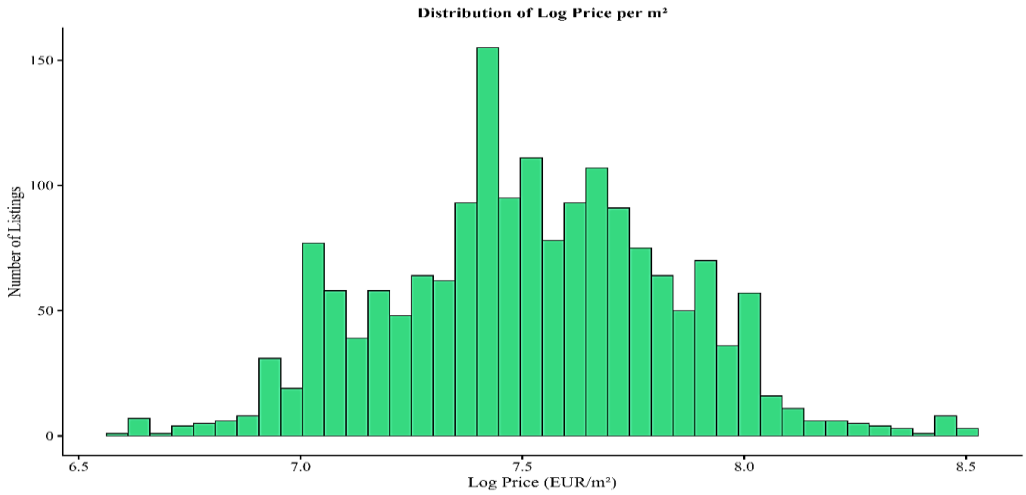

Fig. 2. Distribution of log apartment prices per square meter.

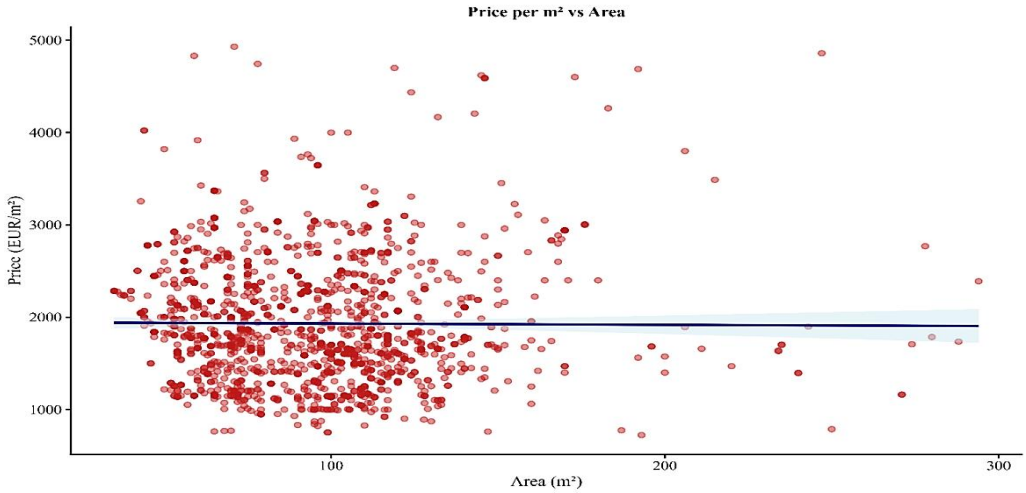

Fig. 3. Scatter plot of apartment price per square meter against apartment size, with fitted linear trend.

Fig. 4 reports median apartment prices per m² across neighbourhoods. Substantial variation is observed, with central and higher-amenity areas exhibiting higher median prices, while more peripheral neighbourhoods display lower values. The use of median prices provides a robust measure of central tendency, reducing the influence of extreme values in the data.

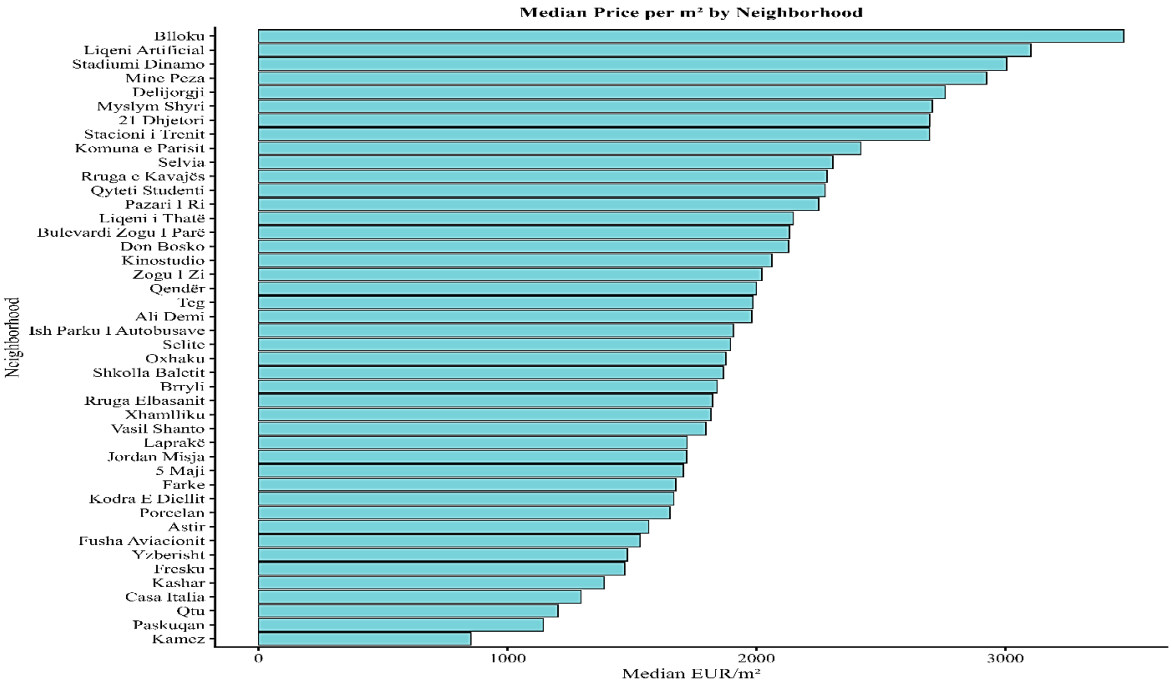

Fig. 4. Median apartment price per square meter by neighbourhood.

The observed spatial variation highlights the importance of location as a key determinant of apartment prices. At the same time, the dispersion of prices within neighbourhoods suggests that structural characteristics such as apartment size, number of rooms, and number of bathrooms also contribute to price heterogeneity.

The exploratory analysis reveals substantial variation in apartment prices across both structural and spatial dimensions. The observed heterogeneity, together with the moderate skewness of the price distribution, supports the use of a modelling framework capable of capturing both linear and nonlinear relationships. These preliminary findings provide empirical justification for the subsequent comparison of econometric and ML models.

4.2 | Hedonic Ordinary Least Squares Results

The empirical analysis begins with the estimation of two linear hedonic specifications. The baseline model includes apartment size, number of bathrooms, and neighbourhood fixed effects, while the extended model additionally incorporates the number of rooms. In the remainder of the analysis, the extended model is treated as the preferred specification, as it provides a more complete set of controls for structural apartment characteristics and allows a more direct economic interpretation of the estimated coefficients.

Table 2 compares the goodness-of-fit statistics of the two OLS specifications. The results show that the extended model performs slightly better than the baseline model. Specifically, the R^2 increases from 0.646 to 0.661, while the adjusted R^2 rises from 0.636 to 0.651. The information criteria also improve, with AIC declining from -731 to -807 and BIC from -469 to -544. These differences indicate that the extended specification offers a better fit to the data.

Table 2. Comparison between the baseline and extended hedonic OLS model.

Model	R^2	Adjusted R^2	AIC	BIC
OLS base	0.646	0.636	-731	-469
OLS extended	0.661	0.651	-807	-544

The final OLS results, reported in *Table 3* with HC1 heteroskedasticity-robust standard errors, indicate that apartment size has a negative and statistically significant coefficient at the 5% level. The estimated coefficient of -0.00084 implies that a one-square-meter increase in apartment size is associated with an approximate 0.08% decrease in price per square meter. Although statistically significant, the estimated effect size is

relatively small. By contrast, the number of bathrooms is positive and statistically significant. The coefficient of 0.05109 implies that an additional bathroom is associated with an increase of roughly 5.24% in price per square meter, holding the other observed characteristics constant. The coefficient on the number of rooms is negative but statistically insignificant, suggesting that once apartment size, bathrooms, and neighbourhood are taken into account, room count does not add meaningful explanatory power to the model.

Table 3. Selected coefficients from the extended OLS model estimated with HC1 robust standard errors.

Variable	Coefficient	Robust SE	p-value	Approx. % Effect
Area (m2)	-0.00084	0.00042	0.0467	-0.08%
Room	-0.00497	0.01526	0.7446	-0.50%
Bathroom	0.05109	0.01560	0.0011	+5.24%

Note: The percentage effects are calculated as $100\times(e^{\beta}-1)$.

Table 4 reports the main diagnostic tests for both OLS models. The multicollinearity results indicate no serious concern, as the adjusted Generalized Variance Inflation Factors (GVIF) values remain low across all core covariates in both models. In the extended model, the values are 1.788 for area, 1.644 for rooms, 1.362 for bathrooms, and 1.006 for neighbourhood, all comfortably below conventional thresholds of concern. By contrast, the Breusch–Pagan test is strongly significant in both models, with test statistics of 350.55 in the baseline model and 395.56 in the extended model, both with p-values below 0.001. The test results provide evidence of heteroskedasticity in the residuals and justify the use of HC1 heteroskedasticity-consistent standard errors for statistical inference.

Table 4. OLS diagnostic tests.

Model	Area	Rooms	Bathrooms	Neighborhood	BP Statistic	df	p-value
OLS base	1.362	—	1.326	1.004	350.55	46	<0.001
OLS extended	1.788	1.644	1.362	1.006	395.56	47	<0.001

Note: Multicollinearity is assessed using GVIF, adjusted for degrees of freedom. The Breusch–Pagan test rejects homoskedasticity in both specifications, supporting the use of heteroskedasticity-robust standard errors.

The results support the use of the extended OLS specification as the primary linear benchmark model. Relative to the baseline specification, the extended model achieves a modest improvement in explanatory performance while retaining coefficient interpretability. The diagnostic tests indicate that multicollinearity is not a concern, whereas heteroskedasticity is present and appropriately addressed through HC1 robust standard errors. Consequently, the extended OLS model serves as the principal explanatory framework and provides a reference point for comparison with the nonlinear ML models presented in the following sections.

4.3|Spatial Heterogeneity and Neighborhood Effects

To evaluate the contribution of spatial factors to apartment price variation, the extended OLS specification incorporates neighbourhood fixed effects. These parameters capture systematic differences in prices across locations after controlling for structural housing characteristics, namely apartment size, number of rooms, and number of bathrooms. The estimated coefficients provide a measure of spatial heterogeneity within the housing market and are interpreted relative to the omitted reference category (21 Dhjetori).

Table 4 reports selected statistically significant neighbourhood effects from the model using robust standard errors. The results indicate substantial spatial variation in apartment prices, even after controlling for observable apartment characteristics. Several neighbourhoods are associated with positive and statistically significant price premiums relative to the reference category. In particular, liqeni artificial and stadiumi Dinamo exhibit higher prices per square meter, suggesting that these locations command a premium in the Tirana housing market.

Table 4. Selected neighborhood fixed effects from the extended OLS model.

Neighborhood	Coefficient	Robust SE	p-value	Approx. % Effect
Liqeni artificial	0.2576	0.1066	0.0158	+29.4%
Stadiumi dinamo	0.1755	0.0370	<0.001	+19.2%
Kamëz	-1.0375	0.0437	<0.001	-64.6%

Paskuqan	-0.7924	0.0233	<0.001	-54.7%
Qtu	-0.7465	0.0358	<0.001	-52.6%
Astir	-0.4730	0.0256	<0.001	-37.7%
Yzberisht	-0.4717	0.0306	<0.001	-37.6%

Note: Coefficients are interpreted relative to the omitted reference neighborhood (21 Dhjetori). Approximate percentage effects are calculated as $e^{\beta}-1$. Estimates are reported with heteroskedasticity-robust standard errors.

In contrast, a number of peripheral or lower-priced areas show large and statistically significant discounts relative to the reference neighbourhood. The strongest negative effects are observed for Kamëz, Paskuqan, and Qtu, all of which are associated with substantially lower prices per m². Similarly, Astir and Yzberisht also display sizeable negative effects, indicating that apartments in these neighbourhoods are priced significantly below those in 21 Dhjetori, holding other characteristics constant.

The magnitude and statistical significance of the neighbourhood coefficients indicate that spatial effects represent an important component of apartment price variation. The results suggest that location-related factors contribute substantially to model explanatory performance, even after accounting for structural housing characteristics.

4.4 | Random Forest Results

To evaluate the predictive performance of a nonlinear modelling framework, the analysis incorporates an RF model estimated using the same explanatory variables as the extended OLS specification. This approach allows a direct comparison between linear and nonlinear modelling strategies under a common set of predictors. Hyperparameters were selected by minimizing OOB prediction error.

The RF model demonstrates the highest predictive accuracy on the test set. As shown in *Table 5*, it achieves an RMSE of 0.176, MAE of 0.114, and R^2 of 0.703 on the log scale, corresponding to approximately 359 EUR/m² RMSE, 220 EUR/m² MAE, and 12.2% MAPE in the original scale. This performance exceeds that of the extended OLS specification, suggesting the presence of nonlinear patterns and interaction effects that are only partially represented within the linear modelling framework. Nevertheless, RF is used here as a complementary predictive benchmark and robustness check, rather than replacing the hedonic OLS model for economic interpretation.

Table 5. Predictive performance of the RF model on the test set.

Model	Scale	RMSE	MAE	R^2	MAPE
RF	log	0.176	0.114	0.703	—
RF	EUR/m²	359	220	0.627	12.2

An additional advantage of the RF model is that it allows for the assessment of variable importance. Based on permutation importance, neighborhood emerges as the most important predictor, followed by area(m²), rooms, and bathrooms (*Fig. 5*). This result is particularly informative because it reinforces the evidence from the OLS model. While the linear specification identifies strong and statistically significant neighbourhood effects, the RF model independently confirms that location is the most influential predictor of apartment prices per square meter.

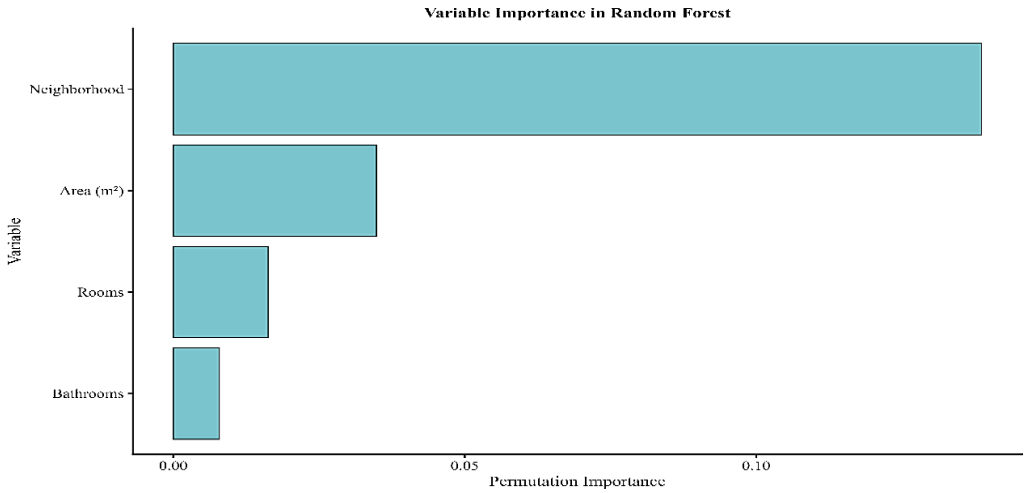

Fig. 5. Variable importance in the RF model.

This pattern is also reflected in the model's overall predictive fit. As shown in *Fig. 6*, the observed-versus-predicted plot indicates that the fitted values broadly track the observed prices. Although some dispersion remains, particularly at higher price levels, the figure suggests that the model captures the general structure of the data well and provides a reasonable approximation of the observed price distribution.

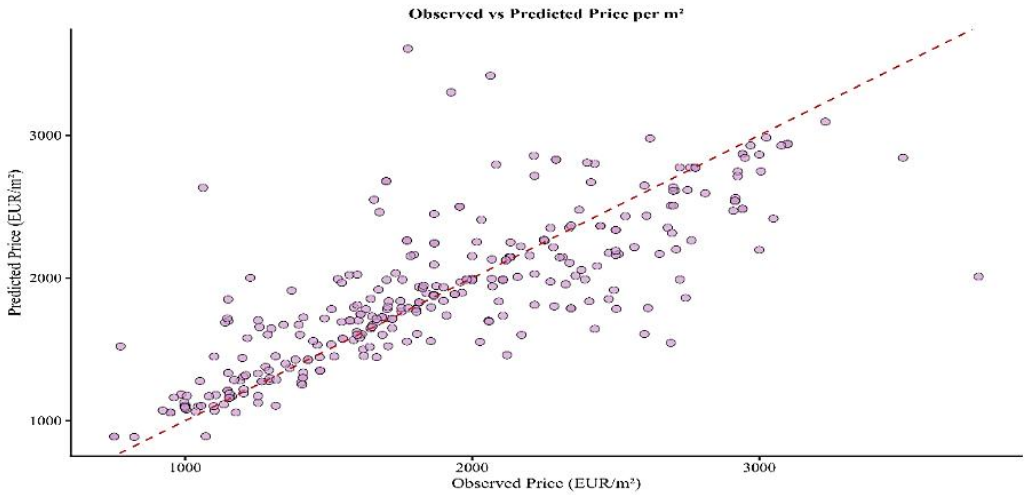

Fig. 6. Observed vs predicted price per m² in the RF model.

Fig. 7 reports partial dependence plots for apartment area, number of rooms, and number of bathrooms. The plots show that apartment area has a clearly non-linear association with predicted price per m², with higher predicted values for smaller units and a general decline as size increases. The number of rooms displays only a limited variation in predicted price once other characteristics are taken into account, suggesting a weaker marginal role. The number of bathrooms shows a more noticeable effect, although the relationship is not strictly linear.

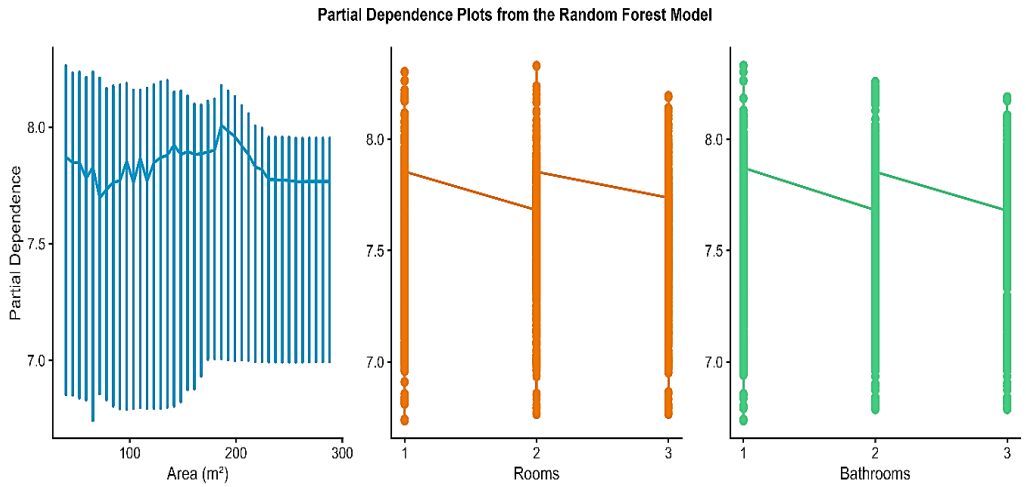

Fig. 7. Partial dependence plots from the RF model.

Taken together, the RF results demonstrate that nonlinear modelling improves predictive performance relative to the linear benchmark while preserving the importance of spatial information identified by the OLS analysis. The variable importance rankings, observed-versus-predicted comparison, and partial dependence plots consistently indicate that both location-related variables and nonlinear structural relationships contribute to model performance. As such, the findings provide empirical support for the use of flexible statistical learning algorithms in housing price modelling. Variable importance measures should not be interpreted as causal effects, but rather as indicators of relative predictive contribution within the model.

4.5 | Model Comparison and Robustness Checks

In addition to the benchmark OLS and RF specifications, the analysis incorporates Elastic Net and XGBoost models as robustness checks. These alternative modelling frameworks enable the evaluation of predictive stability across both regularized linear methods and advanced nonlinear algorithms. The objective is to determine whether the observed empirical patterns are model-specific or remain consistent across different statistical learning approaches.

Table 6 presents the out-of-sample predictive performance of all models. RF consistently achieves the highest accuracy, followed by XGBoost. In contrast, the extended OLS and Elastic Net show nearly identical performance, with RMSE around 0.196, MAE ≈ 0.144–0.145, and R^2 ≈ 0.634–0.635 on the log scale, corresponding to RMSE of 402–403 EUR/m² and MAPE of 14.9% in the original scale. These results suggest that penalization does not substantially enhance predictive performance relative to the linear benchmark.

Table 6. Out-of-sample predictive performance across all estimated models.

Model	Scale	RMSE	MAE	R^2	MAPE
OLS extended	log	0.196	0.144	0.634	—
OLS extended	EUR/m²	403	278	0.530	14.9
Elastic Net	log	0.196	0.145	0.635	—
Elastic Net	EUR/m²	402	278	0.531	14.9
RF	log	0.176	0.114	0.703	—
RF	EUR/m²	359	220	0.627	12.2
XGBoost	log	0.183	0.125	0.681	—
XGBoost	EUR/m²	364	239	0.616	13.2

The nonlinear specifications perform appreciably better. XGBoost reduces prediction error relative to both OLS and Elastic Net, achieving an RMSE of 0.183 and an MAE of 0.125 on the log scale, together with an RMSE of 364 EUR/m² and a MAPE of 13.2% on the original scale. However, its performance remains modestly below that of RF, which emerges as the best-performing model overall, with an RMSE of 0.176, an MAE of 0.114, and an R^2 of 0.703 on the log scale, and an RMSE of 359 EUR/m² and a MAPE of 12.2% on the original scale.

Two broader implications follow from these comparisons. First, regularization alone appears insufficient to generate meaningful predictive gains in this setting. The near equivalence between Elastic Net and OLS suggests that shrinkage does not materially alter the information content of the linear specification. The predictive limitations of the linear model do not appear to arise primarily from overfitting or coefficient instability. Second, the superior performance of the tree-based algorithms suggests the presence of nonlinear structures and interaction effects that are only partially captured by linear specifications. The fact that both RF and XGBoost outperform the linear alternatives lends further support to this interpretation.

The model-specific tuning results are also informative. For Elastic Net, cross-validation selects $\alpha = 0.7$ and $\lambda = 0.0002288$ as the optimal parameter combination. Yet, despite this data-driven regularization, predictive performance remains virtually unchanged relative to OLS on both evaluation scales. For XGBoost, the best-performing specification uses a learning rate of 0.05, a maximum tree depth of 5, a minimum child weight of 1, full row subsampling, a column subsampling rate of 0.8, and 500 boosting rounds (*Table 7*). This specification yields a substantial improvement over the linear models, but it does not surpass the predictive performance of RF.

Table 7. Summary of robustness check models and their main implications.

Model	Key Tuning Result	Main Implication
Elastic Net	$\alpha = 0.7$, $\lambda = 0.0002288$	Predictive performance is nearly identical to OLS
XGBoost	Learning rate = 0.05; maximum depth = 5; minimum child weight = 1; row sampling fraction = 1; column sampling fraction = 0.8; 500 boosting rounds	Improves prediction over linear models, but remains slightly below RF

The comparative evaluation demonstrates that the principal findings are robust across alternative modelling frameworks. While the explanatory conclusions derived from the OLS specification remain stable, the predictive assessment consistently favours nonlinear algorithms, particularly RF. The results indicate that the primary gains in model performance arise from the ability of flexible statistical learning methods to capture nonlinear relationships and interaction effects rather than from coefficient regularization alone. Consequently, the evidence supports the complementary use of interpretable econometric models and ML algorithms in housing price modelling.

5 | Discussion

5.1 | Spatial Effects in Housing Price Modelling

Neighborhood effects emerge as the strongest determinant of apartment prices in Tirana. High-price areas, such as Liqeni artificial and stadiumi Dinamo, exhibit substantial premia, whereas Kamëz, Paskuqan, QTU, Astir, and Yzberisht experience notable discounts. This spatial variation indicates the presence of substantial location-related heterogeneity within the housing market. The estimated neighbourhood effects capture systematic differences in apartment prices that remain after controlling for observed structural characteristics, suggesting that location contains important information not fully represented by the available dwelling attributes.

Economically, this suggests that apartment values are shaped not only by the physical characteristics of the dwelling but also by the spatial distribution of urban opportunities. Neighborhood captures a broad set of location-specific advantages and disadvantages, including accessibility, prestige, environmental quality,

proximity to services, and perceived social status, which are often unobservable in listing data. Consequently, neighborhood coefficients are not merely control variables; they provide substantive evidence of pronounced spatial stratification in the Tirana housing market.

This pattern is particularly salient in the context of Tirana's recent urban expansion and uneven development. Buyers price apartments not only for their structural attributes but also for the neighborhood's position within the city's social and economic geography. The observed premiums and discounts underscore the centrality of location: Any model that omits neighborhood heterogeneity would fail to capture a substantial component of the market's economic structure.

5.2 | Comparing Linear and Non-Linear Models: Interpretation Versus Prediction

The comparison of OLS, Elastic Net, RF, and XGBoost highlights a key methodological insight. While the linear hedonic model provides interpretable coefficients and a clear economic interpretation, it does not fully capture the data's structure. RF delivers the strongest out-of-sample performance, with XGBoost also outperforming the linear models, whereas Elastic Net closely resembles OLS.

The model comparison suggests that regularization alone adds little value, consistent with low VIFs, and that predictive gains stem from modeling nonlinearities and interactions. Tree-based models better capture complex combinations of housing attributes, particularly when effects are non-proportional or conditional on other variables.

The results highlight a common trade-off: Linear models are preferred for interpretation, whereas nonlinear models enhance prediction. OLS effectively explains apartment size, bathrooms, and neighborhood effects, while RF improves predictive accuracy. Importantly, neighborhood remains the most important predictor in both models, linking explanatory and predictive insights. Overall, these findings support combining hedonic regression with ML to balance interpretability and predictive robustness.

5.3 | Practical and Methodological Implications

The findings carry both practical and methodological implications. From a modelling perspective, the results demonstrate the importance of incorporating spatial information and nonlinear modelling techniques when analysing housing price data. The consistent importance of neighbourhood variables across alternative specifications indicates that location-related effects constitute a major source of variation that should not be ignored in residential valuation models. First, location emerges as a central organizing factor, with strong neighborhood premia and discounts highlighting spatial differences in affordability and potential housing inequality across the city.

Second, the positive effect of bathrooms and the limited independent role of room count indicate that market valuation prioritizes perceived quality and functional convenience over simple interior divisions, guiding developers toward design choices that enhance comfort and modern standards rather than maximizing room numbers.

Third, the superior predictive performance of RF further indicates that flexible statistical learning algorithms can capture complex relationships that are only partially represented within conventional linear frameworks. This finding is consistent with recent evidence supporting the complementary use of econometric and ML approaches in property valuation and predictive modelling.

While the results should be interpreted in light of the study's data and modelling constraints, the overall findings remain consistent across alternative specifications. Taken together, the evidence provides a useful empirical benchmark for housing price modelling and residential property valuation in emerging urban markets.

6 | Limitations of the Study

Several limitations should be acknowledged in this study. First, the analysis relies on listing prices rather than transaction prices; therefore, the models explain variation in advertised prices rather than finalized market transactions. Second, the available dataset does not include several potentially relevant housing attributes, such as building age, construction quality, renovation status, legal status, and detailed accessibility measures. In addition, variables such as floor level, elevator availability, parking, furnishing, and certificate-related information were excluded from the core modelling framework because of substantial missingness and inconsistent reporting. While this decision preserves sample size and model comparability, it may limit the explanatory power of the estimated models.

Third, the absence of exact geographic coordinates prevents the application of explicit spatial modelling techniques. Spatial variation is instead represented through neighbourhood fixed effects, which capture broad location-related heterogeneity but cannot account for finer-scale spatial dependence or spatial autocorrelation. Future research could incorporate georeferenced property data and spatial econometric methods to address this limitation.

Finally, although the ML algorithms improve predictive performance, they offer a lower degree of interpretability than the hedonic OLS specification. Consequently, the study adopts a complementary modelling strategy in which econometric models provide explanatory insights, while ML algorithms are primarily used for predictive evaluation and robustness assessment.

7 | Conclusion

This study examined apartment price formation in Tirana using a micro-level dataset of residential property listings and a comparative modelling framework that integrated a traditional hedonic regression model with Elastic Net, RF, and XGBoost algorithms. By evaluating both econometric and ML approaches within the same analytical setting, the study provides a comprehensive assessment of the factors influencing residential property values and the predictive capabilities of alternative modelling strategies.

The findings reveal that spatial location constitutes the most influential component of apartment valuation in Tirana, reflecting substantial heterogeneity across neighbourhoods. In addition, selected housing characteristics contribute to price variation, although their effects are less pronounced than those associated with location. These results indicate that residential property values are shaped by a combination of structural housing attributes and broader spatial-economic factors embedded within the urban environment.

A key contribution of this study is the demonstration that predictive performance varies considerably across modelling approaches. While the hedonic OLS framework remains valuable for identifying and interpreting the economic significance of individual housing attributes, nonlinear machine-learning methods provide superior forecasting accuracy. The strong performance of RF and XGBoost suggests that housing prices are influenced by complex relationships that cannot be fully captured through conventional linear specifications alone.

From a broader perspective, the study contributes to the growing literature on residential property valuation in emerging urban markets by providing empirical evidence from Tirana, a rapidly developing city that remains underrepresented in international housing research. The results highlight the importance of incorporating both spatial heterogeneity and advanced predictive techniques when modelling urban housing markets, particularly in cities experiencing rapid demographic and economic transformation.

The findings may be relevant for property valuation professionals, real-estate professionals, urban planners, and policymakers seeking to better understand housing market dynamics and residential affordability patterns. More accurate valuation models can support informed investment decisions, improve market transparency, and contribute to more effective urban development strategies.

Future research could build upon these results by incorporating transaction-level data, detailed geospatial information, neighbourhood socioeconomic indicators, and spatial econometric methods. Such extensions would enable a deeper investigation of spatial dependence, improve model interpretability, and further enhance the accuracy of residential property valuation models.

Acknowledgments

Not applicable.

Author Contributaion

Conceptualization, R. D. and T. Y.; methodology, R. D. and T. Y.; software, R. D.; validation, R. D., T. Y. and R. R.; formal analysis, R. D.; investigation, R. D.; resources, T. Y. and R. R.; data curation, R. D.; writing—original draft preparation, R. D.; writing—review and editing, T. Y., R. D. and R. R.; visualization, R. D.; supervision, T. Y. and R. R.; project administration, R. D.; funding acquisition, R. R. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data supporting the findings of this study were collected from publicly available real estate advertisements published on the Çelësi real estate platform (https://www.celesi.com). The processed dataset generated and analyzed during the current study is available from the corresponding author upon reasonable request. All data were derived from publicly accessible sources and were used exclusively for academic research purposes.

Funding

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Consent for Publication

The author has given consent for the publication of this manuscript.

Ethics Approval and Consent to Participate

This study does not involve any research conducted on human participants or animals.

References

[1] Zhao, C., & Liu, F. (2023). Impact of housing policies on the real estate market-systematic literature review. *Heliyon, 9*(10), e20704. https://www.cell.com/heliyon/fulltext/S2405-8440(23)07912-4

[2] Han, F., Lu, M., Qin, D., Zheng, G., Zeng, G., Tan, Y., ... & He, H. (2025). Exploring housing price dynamics in sustainable cities through a cooperated big data driven machine learning method: Case study on a typical city in China. *City and environment interactions*, *28*, 100223. https://doi.org/10.1016/j.cacint.2025.100223

[3] Çılgın, C., & Gökçen, H. (2023). Machine learning methods for prediction real estate sales prices in Turkey. *Revista de la construcción*, *22*(1), 163-177. http://dx.doi.org/10.7764/rdlc.22.1.163

[4] Pojani, D. (2010). Tirana. *Cities, 27*(6), 483–495. https://doi.org/10.1016/j.cities.2010.02.002

[5] Thanasi, M. (2016). Hedonic appraisal of apartments in Tirana. *International journal of housing markets and analysis, 9*(2), 239–255. https://doi.org/10.1108/IJHMA-03-2015-0016

[6] Liu, T., Wang, J., Liu, L., Peng, Z., & Wu, H. (2025). What are the pivotal factors influencing housing prices? A spatiotemporal dynamic analysis across market cycles from upturn to downturn in Wuhan. *Land*, *14*(2), 356. https://doi.org/10.3390/land14020356

[7] Chwiałkowski, C., Zydroń, A., & Kayzer, D. (2022). Assessing the impact of selected attributes on dwelling prices using ordinary least squares regression and geographically weighted regression: A case study in Poznań, Poland. *Land*, *12*(1), 125. https://doi.org/10.3390/land12010125

[8] Shtepani, E., & Yunitsyna, A. (2023). Evaluation of the spatial quality of apartments from different price categories using the visibility graph analysis: A case of Tirana, Albania. *International journal of real estate studies, 17*(1), 83–92. https://doi.org/10.11113/intrest.v17n1.268

[9] Nurja, I., Jaupi, F., & Elezaj, O. (2022). Appraisal of apartments in Albania using hedonic regression. *WSEAS transactions on business and economics, 19*, 1816–1823. https://doi.org/10.37394/23207.2022.19.163

[10] Ko, D., & Park, S. (2024). Investigating the correlation between air pollution and housing prices in Seoul, South Korea: Application of explainable artificial intelligence in random forest machine learning. *Sustainability, 16*(11), 4453. https://doi.org/10.3390/su16114453

[11] Kim, W., & Hong, J. (2024). Stacked ensemble model for the automatic valuation of residential properties in South Korea: A case study on Jeju Island. *Land, 13*(9), 1436. https://doi.org/10.3390/land13091436

[12] Wan, H., Chowdhury, P. K. R., Yoon, J., Bhaduri, P., Srikrishnan, V., Judi, D., & Daniel, B. (2025). Explaining drivers of housing prices with nonlinear hedonic regressions. *Machine learning with applications, 21*, 100707. https://doi.org/10.1016/j.mlwa.2025.100707

[13] Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land, 11*(3), 334. https://doi.org/10.3390/land11030334

[14] Khoshnoud, M., Sirmans, G. S., & Zietz, E. N. (2023). The evolution of hedonic pricing models. *Journal of real estate literature, 31*(1), 1–47. https://doi.org/10.1080/09277544.2023.2201020

[15] Rey-Blanco, D., Zofio, J. L., & Gonzalez-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. *Expert systems with applications, 235*, 121059. https://doi.org/10.1016/j.eswa.2023.121059

[16] Marinković, S., Džunić, M., & Marjanović, I. (2024). Determinants of housing prices: Serbian cities' perspective. *Journal of housing and the built environment, 39*(3), 1601–1626. https://doi.org/10.1007/s10901-024-10134-5

[17] Shehu, E., Afezolli, A., & Kondi, I. (2015). The model for determining the market value of residential properties in Tirana city. *Proceedings of international conference on innovation in civil and structural engineering (ICICSE)* (pp. 164-169). Unique Conferences Publishing. http://dx.doi.org/10.17758/UR.U0615307

[18] Wang, Z., Wang, Y., Xia, X., Chen, S., & Jiang, W. (2025). How does built environment influence housing prices in large-scale areas? An interpretable machine learning method by considering multi-dimensional accessibility. *ISPRS international journal of Geo-information*, *14*(11), 436. https://doi.org/10.3390/ijgi14110436

[19] Zhang, Y., & Miller, E. J. (2025). Location choice of residential housing supply: An application of the multiple discrete-continuous extreme value (MDCEV) model. *Journal of choice modelling*, *54*, 100535. https://doi.org/10.1016/j.jocm.2024.100535

[20] Dou, M., Gu, Y., & Fan, H. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. *Applied geography, 158*, 103032. https://doi.org/10.1016/j.habitatint.2024.103212

[21] Li, J., Ossokina, I. V, & Arentze, T. A. (2024). The impact of urban green space on housing value: A combined hedonic price analysis and land use modeling approach. *Journal of sustainable real estate, 16*(1), 2432758. https://doi.org/10.1080/19498276.2024.2432758

[22] Li, C., Zhou, Y., Wu, M., Xu, J., & Fu, X. (2025). Exploring nonlinear threshold effects and interactions between built environment and urban vitality at the block level using machine learning. *Land, 14*(6), 1232. https://doi.org/10.3390/land14061232

[23] Anelli, D., Morano, P., Tajani, F., & Guarini, M. R. (2025). The interpretative effects of normalization techniques on complex regression modeling: An application to real estate values using machine learning. *Information, 16*(6), 486. https://doi.org/10.3390/info16060486

[24] Choy, L. H. T., & Ho, W. K. O. (2023). The use of machine learning in real estate research. *Land, 12*(4), 740. https://doi.org/10.3390/land12040740

[25] Moreno-Foronda, I., Sánchez-Martínez, M. T., & Pareja-Eastaway, M. (2025). Comparative analysis of advanced models for predicting housing prices: A review. *Urban science*, *9*(2), 32. https://doi.org/10.3390/urbansci9020032

[26] Maselli, G., & Nesticò, A. (2025). Machine learning algorithms and explainable artificial intelligence for property valuation. *Real estate, 2*(3), 12. https://doi.org/10.3390/realestate2030012

[27] Soltani, A., & Lee, C. L. (2024). The non-linear dynamics of South Australian regional housing markets: A machine learning approach. *Applied geography, 166*, 103248. https://doi.org/10.1016/j.apgeog.2024.103248

[28] Jafary, P., Shojaei, D., Rajabifard, A., & Ngo, T. (2025). AI, machine learning and BIM for enhanced property valuation: Integration of cost and market approaches through a hybrid model. *Habitat international, 164*, 103515. https://doi.org/10.1016/j.habitatint.2025.103515

[29] Andrade-Girón, D. C., Marin-Rodriguez, W. J., & Zuñiga-Rojas, M. G. (2025). Intelligent feature selection ensemble model for price prediction in real estate markets. *Informatics, 12*(2), 52. https://doi.org/10.3390/informatics12020052

[30] Yang, Y., & Wang, H. (2025). Random forest-based machine failure prediction: A performance comparison. *Applied sciences, 15*(16), 8841. https://doi.org/10.3390/app15168841

[31] Zhang, Q., & Abdullah, F. (2026). Hedonic beats utilitarian: Differential effects of AI chatbots and AR/VR on consumer engagement in e-commerce. *Journal of theoretical and applied electronic commerce research, 21*(2), 60. https://doi.org/10.3390/jtaer21020060

[32] Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data science and management, 4*, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

[33] An, S., Song, Y., Jang, H., & Ahn, K. (2025). Toward transparent and accurate housing price appraisal: Hedonic price models versus machine learning algorithms. *Financial innovation, 11*(1), 141. https://doi.org/10.1186/s40854-025-00874-w

[34] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of political economy, 82*(1), 34–55. https://doi.org/10.1086/260169

[35] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[36] James, G. (2013). *An introduction to statistical learning with applications in R*. Springer. https://doi.org/10.1007/978-3-031-38747-0

[37] Kuhn, M. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

[38] O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

[39] Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society, 47*(5), 1287–1294. https://doi.org/10.2307/1911963

[40] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the econometric society, 48*(4), 817–838. https://doi.org/10.2307/1912934

[41] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society series b: Statistical methodology, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x