ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 446-455
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Predicting Residential Housing Rent Prices Using Machine Learning and Statistical Modelling Approaches

Nur Suhaila Bdrul Hisham[1], Azme Khamis[1]*

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 86400 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: azme@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.051

Article Info

Received: 4 January 2026
Accepted: 19 January 2026
Available online: 9 July 2026

Keywords

Residential Rental Price, Housing Market Analysis, Stepwise Regression, Multiple Linear Regression, Random Forest, XGBoost, Machine Learning

Abstract

Accurate estimation of residential rental prices is critical for property valuation, investment decision-making, and housing market analysis in urban areas. The rapid urban development in Selangor and Kuala Lumpur has heightened the complexity of determining rental prices, thereby necessitating data-driven analytical approaches. This study identifies key factors influencing residential rental prices and evaluates the predictive performance of statistical and machine learning models using housing data from Selangor and Kuala Lumpur. Stepwise regression was initially applied to identify significant explanatory variables affecting monthly rental prices. Multiple Linear Regression served as the baseline statistical model, while Random Forest and Extreme Gradient Boosting (XGBoost) were implemented to enhance predictive accuracy through nonlinear learning. The dataset was partitioned into training (80%) and testing (20%) subsets, and model performance was assessed using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results show that property size, furnishing status, location, and completion year are among the most influential factors affecting rental prices. In terms of predictive performance, machine learning models outperformed the traditional regression approach, with XGBoost achieving the lowest test-set prediction error, followed by the Random Forest. These findings demonstrate the effectiveness of machine learning techniques for predicting residential rental prices and offer valuable insights for property analysts, policymakers, and real estate practitioners in Malaysia.

1. Introduction

Housing is a fundamental necessity that provides individuals with security, stability, and shelter, while also serving as a key driver of urban development and national economic growth. In developing countries such as Malaysia, housing demand has shown a consistent upward trend, particularly in urban and metropolitan regions [1]. The residential housing market plays an important role in shaping socioeconomic development, as it is closely linked to population growth, employment opportunities, and living standards.

Rapid urbanisation and increasing population density have significantly intensified housing demand in Malaysia, especially within major urban centres such as Selangor and Kuala Lumpur. According to the Department of Statistics Malaysia (2020), Malaysia's population increased from 27.5 million in 2010 to 32.4 million in 2020, with an average annual growth rate of 1.7%. This population growth is expected to continue as Malaysia approaches 2030, leading to further urban concentration and increased pressure on the residential housing market [2]. As urbanisation accelerates, rental housing has become a critical alternative for households that are unable or unwilling to purchase property due to affordability constraints.

In Selangor, the situation is particularly critical due to the rapid development of high-end residential properties, including luxury condominiums, which have contributed to rising property prices. This trend has encouraged many urban households to rent rather than own homes. As a result, the residential rental market has expanded substantially, making the determination of rental prices more complex and dynamic. Similar patterns have been observed in Kuala Lumpur, where proximity to employment centres and urban amenities further drives rental demand [3].

In addition to domestic population growth, increasing globalisation and cross-border mobility have contributed to a growing number of foreign residents seeking employment, education, and business opportunities in Malaysia's urban areas. These individuals tend to concentrate in economically developed regions, further intensifying demand for rental housing. The concentration of foreign immigrants in low- and middle-income housing areas has created additional challenges, particularly for local residents' housing accessibility and affordability. This situation has reduced the availability of affordable rental units and increased competition within the rental market [4].

Recent residential property price updates reported by the National Property Information Centre (2025) indicate a clear hierarchy in housing prices across different property categories in Selangor, with semi-detached houses commanding the highest prices and low-cost flats remaining the most affordable. These price differentials suggest a strong relationship between property type, purchase prices, and rental market dynamics [5]. Furthermore, data from PropertyGuru Malaysia (2024) indicate that, although the rental demand index for Malaysian residential properties declined by 11% quarter-over-quarter, the supply index increased, reflecting an imbalance between supply and demand in the rental market [6]. Factors such as property size, location, and available amenities have been identified as key contributors to these rental trends [7].

Given the increasing complexity and variability of residential rental markets, traditional modelling approaches may be insufficient in capturing nonlinear relationships and interactions among housing attributes. Conventional statistical methods have been widely applied in rental price analysis; however, they often struggle to adapt to changing market dynamics, economic fluctuations, and evolving tenant preferences. While improvements to traditional approaches have been suggested to encourage the development of affordable housing without compromising profitability, their limitations remain evident in highly heterogeneous markets [8].

Consequently, residential rental prices have become a significant issue in Malaysia, particularly in Selangor and Kuala Lumpur, where rapid development and diverse housing attributes contribute to substantial variation in rental prices [9]. Accurate modelling of rental prices is therefore essential to support transparency in rental valuation and to assist stakeholders in making informed decisions. In response to these challenges, this study explores the application of statistical and machine learning techniques to analyse and predict residential rental prices in Selangor and Kuala Lumpur. Stepwise Regression is employed to identify the key factors influencing rental prices, while Random Forest and Extreme Gradient Boosting (XGBoost) models are applied to improve prediction accuracy and capture complex data patterns. The specific objectives of this study are: (i) to identify the factors influencing residential housing rent prices in Selangor and Kuala Lumpur.

2. Materials and Methods

The methodology for this study was designed to analyse residential rental price behaviour and to develop predictive models using both statistical and machine-learning techniques. Stepwise Regression was employed to identify significant factors influencing rental prices, followed by multiple linear regression as the baseline model. Additionally, Random Forest and Extreme Gradient Boosting (XGBoost) models were implemented to enhance predictive accuracy by capturing nonlinear relationships among housing attributes. The analysis was performed using RStudio, which facilitated data preprocessing, feature selection, model development, and performance evaluation through robust support for statistical modelling and machine learning.

2.1 Data Preprocessing

Secondary data were obtained from a publicly available Kaggle dataset containing residential rental property records [10]. The dataset is limited to urban residential properties in Selangor and Kuala Lumpur, Malaysia, the most developed regions in the country, and with substantial housing demand. These regions were selected due to their rapid urbanisation, high population density, and dynamic rental markets.

The dataset contains monthly rental prices and a range of property attributes, including property type, size, number of rooms, number of bathrooms, parking availability, furnishing status, completion year, and location. After data cleaning and preprocessing, 4,900 observations of residential rentals remained for analysis. The monthly rental price was designated as the response variable, and the remaining attributes were used as explanatory variables for modelling and prediction.

Prior to model development, categorical variables were encoded as numeric variables to ensure compatibility with statistical and machine learning algorithms. This encoding facilitated the incorporation of categorical attributes, such as property type, furnishing status, and location, into the modelling framework. Table 1 summarises the encoded variables and presents an overview of the final variable representations following preprocessing.

Table 1 Summary of Encoded Variables After Preprocessing

Variable Group	Encoding Method	Description
Monthly_price	Numerical (Target Variable)	Monthly rental price in Malaysian Ringgit (RM), used as the dependent variable.
Completion_year	Numerical	Year the residential property was completed.
Property_type	One-Hot Encoding	Property types encoded into binary dummy variables.
Rooms	Numerical	Number of bedrooms in the property.
Parking	Numerical	Number of parking spaces available.
Bathroom	Numerical	Number of bathrooms in the property.
Furnished	Binary Encoding	Furnishing status is encoded as 1 (furnished) and 0 (unfurnished).
Location	One-Hot Encoding	Location variables were transformed into multiple binary dummy variables.
Size	Numerical	Built-up area of the property in square feet (sq ft).

After preprocessing, the dataset was randomly split into training and test sets at an 80:20 ratio. 80% of the data was used for model development and feature selection, while the remaining 20% was reserved for performance evaluation. This partitioning approach supports reliable assessment of model generalisation and enables unbiased comparison between statistical and machine learning models [11].

The dataset provides a comprehensive representation of essential structural and locational characteristics relevant to residential rental pricing. It is appropriate to apply both statistical and machine learning methods to analyse rental price behaviour in urban Malaysia.

2.2 Feature Selection

Feature selection was implemented to improve model interpretability, reduce dimensionality, and enhance predictive accuracy in residential rental price modelling. The large number of variables resulting from categorical encoding necessitated an effective feature selection strategy to maintain model efficiency while retaining essential information.

Multiple Linear Regression (MLR) was estimated using the Ordinary Least Squares (OLS) framework for statistical modelling. Feature selection was based on statistical significance, with predictors evaluated through hypothesis testing of regression coefficients. Variables with p-values below the 0.05 significance threshold were retained, whereas those deemed insignificant were excluded. This method facilitates the identification of variables demonstrating significant linear associations with residential rental prices and supports model interpretability. The linear regression model used for feature selection is expressed as in Eq. (1) [12];

$$\boldsymbol{y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i} \tag{1}$$

Where $\boldsymbol{y_i}$ represents the monthly rental price, $\boldsymbol{x_{ij}}$ denotes the *j*-th independent variable for observation *i*, $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{\beta_j}$ are the regression coefficients, and $\boldsymbol{\varepsilon_i}$ is the error term.

The estimation of regression coefficients was carried out using the OLS estimator, which minimizes the sum of squared residuals and is shown in Eq. (2) [13];

$$\boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T y} \tag{2}$$

This method ensures unbiased, efficient coefficient estimates under standard regression assumptions. The statistical significance of each coefficient was evaluated using an individual hypothesis test as shown in Eq. (3) [14];

$$\begin{aligned} \boldsymbol{H_0}&: \boldsymbol{\beta_j} = \mathbf{0} \\ \boldsymbol{H_1}&: \boldsymbol{\beta_j} = \mathbf{0} \end{aligned} \tag{3}$$

The predictors that satisfied the rejection criterion of $\boldsymbol{H_0}$ were identified as influential factors affecting residential rental prices and retained in the stepwise regression and multiple linear regression models. Explicit feature selection was not necessary for the machine learning models Random Forest Regression and Extreme Gradient Boosting (XGBoost). Tree-based models inherently manage high-dimensional feature spaces and assess variable importance during training. Therefore, all encoded features were retained during machine learning model development to enable the algorithms to capture complex nonlinear relationships and variable interactions.

Correlation analysis among numerical predictors was performed during preprocessing to identify potential redundancy. Because tree-based models are robust to multicollinearity, correlation-based feature elimination was applied conservatively. After preprocessing and encoding, the final dataset comprised 8 independent variables, which were used consistently across all predictive models. This feature selection approach is consistent with previous research indicating that regression-based selection enhances interpretability, whereas machine learning models achieve greater predictive accuracy by retaining a comprehensive feature set.

2.3 Model Development and Evaluation

Three modelling approaches were developed to identify influential rental price factors and compare predictive performance: Multiple Linear Regression (MLR), Random Forest Regression (RF), and Extreme Gradient Boosting (XGBoost). These models encompass both traditional statistical methods and advanced machine learning techniques, thereby enabling a comprehensive evaluation of linear and nonlinear relationships in the residential rental dataset.

Multiple Linear Regression was employed as the baseline statistical model to examine the linear association between monthly rental price and housing attributes. The general model formulation is expressed in Eq. (1). The regression parameters were estimated using the Ordinary Least Squares (OLS) method as described in Eq. (2). Model adequacy was evaluated through goodness-of-fit statistics and residual diagnostics.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees through bootstrap sampling and random feature selection. In regression tasks, the final prediction is computed by averaging the outputs of all trees, as shown in Eq. (4) [15];

$$\boldsymbol{\hat{y}} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{T}} \boldsymbol{\hat{y}}^{(\boldsymbol{t})} \tag{4}$$

Where $\boldsymbol{T}$ denotes the number of trees and $\boldsymbol{\hat{y}}^{(\boldsymbol{t})}$ represents the prediction from the t-th tree. This ensemble mechanism reduces variance and improves predictive stability. Feature importance was assessed by measuring the increase in Mean Squared Error (MSE), thereby identifying dominant rental price determinants. Next, Extreme Gradient Boosting (XGBoost) is a boosting-based ensemble model that sequentially builds decision trees to minimize prediction error. The model optimizes a regularized objective function as shown in Eq. (5) [16];

$$\boldsymbol{L} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{l}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) + \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \boldsymbol{\Omega}\,(\boldsymbol{f_k}) \tag{5}$$

Where $\boldsymbol{l}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i})$ represent the loss function and $\boldsymbol{\Omega}(\boldsymbol{f_k})$ denotes the regularization term controlling model complexity. For regression tasks, the squared error loss function is applied as shown in Eq. (6) [17];

$$\boldsymbol{l}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) = (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}} \tag{6}$$

This gradient boosting framework enables modelling of nonlinear interactions and complex relationships between housing attributes and rental prices.

Model evaluation measures are expressed in Eq. (7), Eq. (8), and Eq (9) [18];

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

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

Lower RMSE and MAE values correspond to greater prediction accuracy, whereas higher $\boldsymbol{R^2}$ values demonstrate enhanced explanatory power. These metrics facilitate objective comparisons among MLR, Random Forest, and XGBoost models to identify the most effective method for predicting residential rental prices in Selangor and Kuala Lumpur.

3. Results and Discussions

This analysis aims to identify the key determinants influencing residential rental prices, evaluate the predictive performance of statistical and machine learning models, and compare the effectiveness of Multiple Linear Regression, Random Forest, and Extreme Gradient Boosting in modelling rental price behaviour in Selangor and Kuala Lumpur. The results assess the statistical significance of selected predictors, evaluate model accuracy using established metrics, and analyse feature importance to identify the dominant drivers of rental prices. The discussion clarifies the comparative strengths, predictive capabilities, and practical implications of each modelling technique, providing a comprehensive understanding of the most suitable approach for predicting residential rental prices in urban Malaysia.

3.1 Statistical Modelling Result

Stepwise regression and multiple linear regression (MLR) were employed to identify significant factors affecting residential rental prices in Selangor and Kuala Lumpur. The stepwise regression utilized the Akaike Information Criterion (AIC) to select the most suitable subset of explanatory variables. The final model included essential structural and locational attributes such as completion year, property type, number of rooms, parking availability, number of bathrooms, furnishing status, location, and property size.

Table 2 Final Stepwise Regression Model (AIC-Based Selection)

No.	Variable Remove	Degree of Freedom	Sum of Squares (SS)	Residual Sum of Squares (RSS)	AIC
1	None(Full Model)	-	-	1055421256	49043
2	Parking	1	1655397	1057076653	49047
3	Property Type	6	18016479	1073437735	49098
4	Bathroom	1	21289290	1076710546	49119
5	Rooms	1	23392422	1078814678	49127
6	Completion Year	1	37401617	1092822873	49178
7	Location	1	126051373	1181472629	49483
8	Property Size	1	230426922	1285848178	49815
9	Furnishing Status	2	279814043	1335235299	49961

The AIC-based selection procedure retained all significant predictors in the full model, indicating that each selected variable contributed meaningfully to the variation in rental prices. These findings highlight the significance of both structural housing characteristics and locational factors in determining monthly rental values.

Following feature selection, a multiple linear regression model was estimated using ordinary least squares. The final regression model is shown in Eq. 10.

$$\begin{gathered} Monthly\ Price_i = \beta_0 + \beta_1(Completion\ Year_i) + \beta_2(Property\ Type_i) + \beta_3(Rooms_i) + \\ \beta_4(Parking_i) + \beta_5(Bathrooms_i) + \beta_6(Furnishing\ Status_i) + \beta_7(Location_i) + \\ \beta_8(Property\ Size_i) + \varepsilon_i \end{gathered} \tag{10}$$

The regression coefficients indicate that property size and furnishing status are strongly positively associated with rental prices, suggesting that larger, fully furnished properties command higher rents. Location has a significant effect, as properties in Kuala Lumpur command a premium relative to those in Selangor. In contrast, structural attributes such as the number of rooms exhibit variable effects depending on property configuration and size.

Table 3 Multiple Linear Regression Coefficients

Variable	Estimate (β)	Std Error	t-value	p-value
Intercept	-40,546.481	3,514.926	-11.536	< 0.001
Completion Year	20.527	1.745	11.764	< 0.001
Property Type: Condominium	62.172	27.237	2.283	0.023
Property Type: Duplex	83.588	153.882	0.543	0.587
Property Type: Others	-130.476	166.574	-0.783	0.433
Property Type: Service Residence	192.924	28.616	6.742	< 0.001
Property Type: Studio	156.417	82.711	1.891	0.059
Property Type: Townhouse Condo	-3.300	302.110	-0.011	0.991
Number of Rooms	-144.200	15.500	-9.303	< 0.001
Parking	44.368	17.927	2.475	0.013
Number of Bathrooms	212.612	23.956	8.875	< 0.001
Not Furnished	-668.451	31.631	-21.133	< 0.001
Partially Furnished	-527.207	18.059	-29.194	< 0.001
Location: Selangor	-385.586	17.856	-21.596	< 0.001
Property Size (sqft)	1.276	0.044	29.199	< 0.001

The overall model exhibits strong explanatory power, as evidenced by the adjusted coefficient of determination, which indicates that a substantial proportion of variation in rental prices is accounted for by the selected predictors. The F-statistic further confirms statistical significance at the 5% level, indicating that the explanatory variables collectively contribute to the prediction of rental prices. Residual diagnostics indicate that the regression assumptions are adequately met, thereby supporting the validity of the statistical model as a baseline for comparison with machine learning approaches.

In summary, the statistical modelling results underscore the significance of structural property characteristics and locational factors in determining residential rental prices. These findings address the first research objective by identifying key determinants of rental prices and establishing a benchmark for evaluating the performance of Random Forest and XGBoost models in subsequent sections.

3.2 Machine Learning Model Performance

Predictive accuracy beyond the linear framework was evaluated by developing and accessing Random Forest and Extreme Gradient Boosting (XGBoost) models using the testing dataset. The performance of these machine learning models was compared to that of the baseline Multiple Linear Regression model based on Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R-squared).

Table 4 Model Performance Comparison

Model	RMSE	MAE	R-squared
Multiple Linear Regression	498.5479	348.0115	0.5177
Random Forest	411.2966	265.6406	0.6718

XGBoost	380.4181	232.4163	0.7192

Table 3 shows that both Random Forest and XGBoost outperform Multiple Linear Regression in predictive performance. The Multiple Linear Regression model produced an RMSE of 498.5479, an MAE of 348.0115, and an R-squared of 0.5177, indicating moderate explanatory power. In comparison, the Random Forest model achieved a lower RMSE of 411.2966, a reduced MAE of 265.6406, and an improved R-squared of 0.6718. This reflects a substantial reduction in prediction error and a stronger model fit.

XGBoost achieved the highest overall performance among the evaluated models, with the lowest RMSE (380.4181), the lowest MAE (232.4163), and the highest R-squared (0.7192). These results indicate superior generalisation capability and improved predictive accuracy for estimating residential rental prices. The approximately 23.7% reduction in RMSE from Multiple Linear Regression to XGBoost underscores the effectiveness of boosting-based ensemble learning in modelling complex rental price structures.

The superior performance of machine learning models demonstrates their ability to capture nonlinear relationships and interactions among housing attributes, which are often inadequately represented in linear regression frameworks. Random Forest provides robust predictive performance through bagging-based variance reduction, whereas XGBoost further improves accuracy by sequentially correcting residual errors via gradient-boosting optimisation. These findings address the second and third research objectives by confirming the effectiveness of advanced machine learning techniques and identifying XGBoost as the most suitable approach for predicting residential rental prices in Selangor and Kuala Lumpur.

3.3 Features Importance Analysis

Feature importance analysis was conducted using Random Forest and XGBoost to further interpret the predictive behaviour of the machine learning models. The analysis identifies the relative contribution of each housing attribute to residential rental prices in Selangor and Kuala Lumpur.

Random Forest assesses feature importance by quantifying the increase in Mean Squared Error (MSE) that occurs when a variable is permuted. Variables associated with a larger increase in MSE are considered more important.

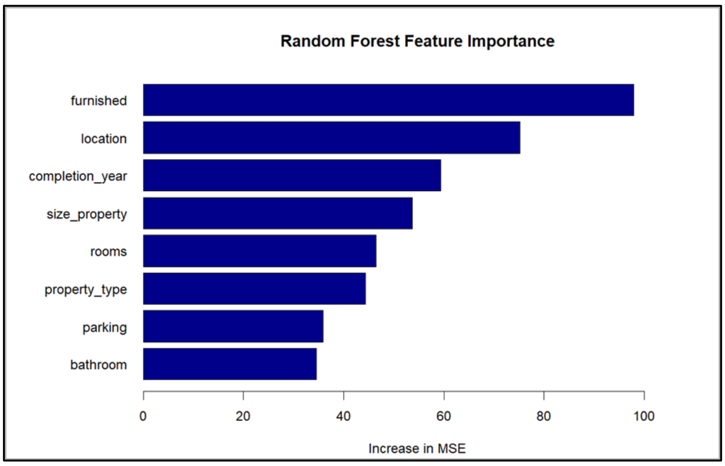

Fig. 1 Random Forest Feature Importance

Fig. 1 demonstrates that furnishing status is the most influential variable in predicting monthly rental prices. Location, completion year, and property size follow in importance. Structural attributes, including the number of rooms, property type, parking, and bathroom, also contribute to the model but have comparatively lower significance.

The prominence of furnishing status suggests that rental prices in urban areas such as Selangor and Kuala Lumpur are highly sensitive to whether a unit is furnished, partially furnished, or unfurnished. This finding reflects tenant preferences in metropolitan regions, where convenience and immediate occupancy readiness substantially influence rental value. Location also plays a significant role, consistent with the economic reality that properties in strategic urban areas, especially those near employment centers and public amenities, command higher rents. Similarly, the completion year indicates that newer properties tend to command higher rental premiums.

In contrast to Random Forest, XGBoost assesses feature importance using Gain, which quantifies each variable's contribution to enhancing the model's predictive accuracy through gradient boosting.

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 446-455

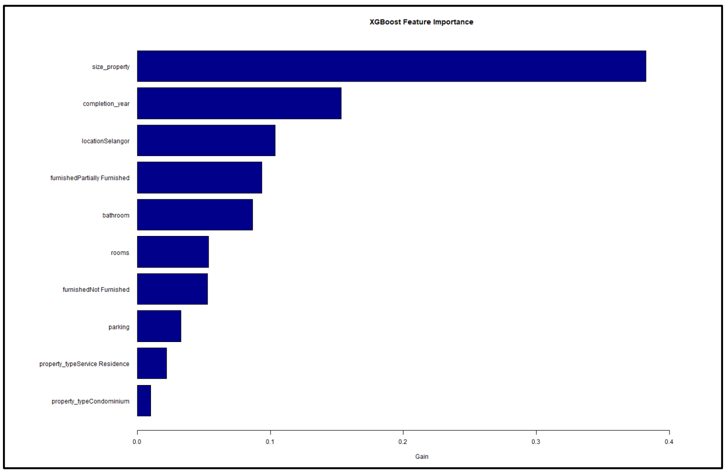

Fig. 2 XGBoost Feature Importance

Fig. 2 demonstrates that property size is the most influential predictor in the XGBoost model, followed by completion year and location. Furnishing status, bathroom count, and number of rooms also substantially contribute to predictive performance.

The prominence of property size in XGBoost suggests that the boosting mechanism effectively captures nonlinear relationships between size and rental price. Larger properties offer increased living space and flexibility, which can elevate rental demand and pricing power. The relative importance ranking contrasts with that of the Random Forest, underscoring differences in the models' learning mechanisms. Whereas Random Forest assigns greater importance to furnishing status, XGBoost prioritizes property size because its sequential boosting approach reduces residual error by focusing on the most impactful predictors at each iteration. Thus, a normalized comparison of feature-importance scores between Random Forest and XGBoost was conducted to enhance model interpretability.

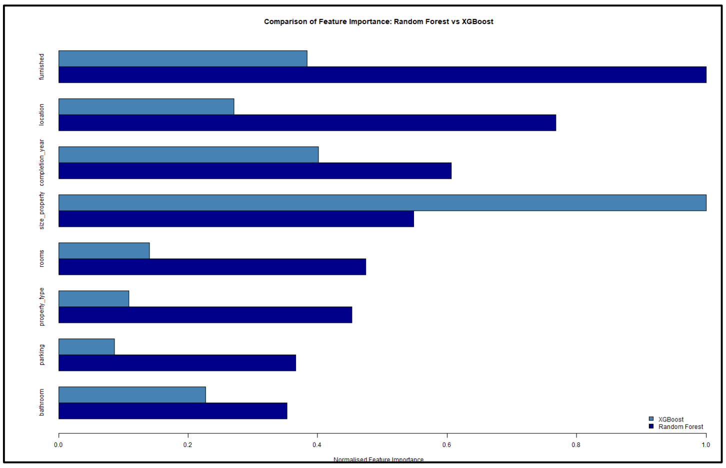

Fig. 3 Comparison of Feature Importance (RF vs XGBoost)

Fig. 3 shows that both models consistently identify structural and locational variables as the primary determinants of residential rental prices in Selangor and Kuala Lumpur. However, differences in the two algorithms' learning mechanisms lead to distinct variable rankings. The Random Forest model assigns the highest importance to furnishing status, followed by location, completion year, and property size. This result suggests that the bagging-based ensemble approach places greater emphasis on categorical distinctions, particularly the property's condition of furnishing. In urban rental markets, the level of furnishing significantly influences tenant preferences and rental premiums, which explains its strong predictive power.

In contrast, the XGBoost model identifies property size as the most influential predictor, with completion year and location also showing significant importance (see Fig. 3). As a boosting-based algorithm, XGBoost sequentially minimizes prediction errors and captures nonlinear interactions more effectively, thereby emphasizing continuous structural variables such as property size. Larger properties generally provide greater functionality and living space, which directly contributes to higher rental valuations. The consistent prominence of completion year across both models further suggests that newer developments attract stronger rental demand due to modern facilities and improved housing standards.

Although the importance rankings differ between the two approaches, both models show strong agreement regarding the significance of key predictors, including structural characteristics (property size, number of rooms, and number of bathrooms), locational attributes, property condition (completion year), and furnishing status. In contrast, parking and bathroom variables are relatively less important than primary structural and locational factors, indicating that these attributes contribute to rental valuation but with smaller marginal effects.

This comparative feature-importance analysis directly addresses the first research objective of identifying the factors that influence residential rental prices. The machine learning results confirm that rental price variability is driven by a multidimensional combination of physical property attributes, furnishing level, and geographic location. Additionally, the consistency observed between Random Forest and XGBoost enhances the analytical validity of the study and increases confidence in the robustness of the modelling framework.

Overall, the findings demonstrate that the determination of rental prices in highly urbanised regions of Malaysia is complex and multifactorial. The integration of ensemble learning techniques provides deeper insight into nonlinear patterns and interaction effects within the housing dataset. These results offer valuable, data-driven implications for tenants, landlords, investors, and policymakers seeking to understand and respond to rental price dynamics in Selangor and Kuala Lumpur.

4. Conclusion

This study developed and evaluated statistical and machine learning models to predict residential rental prices in Selangor and Kuala Lumpur, Malaysia. Stepwise regression identified property size, furnishing status, location, completion year, and key structural attributes as significant determinants of rental prices, thereby addressing the first research objective. Comparative model evaluation demonstrated that machine learning approaches substantially outperformed the traditional Multiple Linear Regression model, with XGBoost achieving the best predictive performance (RMSE = 380.4181, MAE = 232.4163, R^2 = 0.7192), followed by Random Forest (RMSE = 411.2966, R^2 = 0.6718), while the linear model showed comparatively lower explanatory power (R^2 = 0.5177). The superior performance of XGBoost indicates its strong ability to capture nonlinear relationships and complex interactions among housing attributes, making it the most suitable model for predicting residential rental prices in urban Malaysia. Feature importance analysis further confirmed that rental prices are driven by a multidimensional combination of structural characteristics, furnishing level, property condition, and geographic location. Overall, the findings demonstrate that advanced machine learning techniques, particularly boosting-based ensemble methods, yield more accurate and robust predictions than conventional regression approaches, thereby offering valuable data-driven insights for tenants, landlords, investors, and policymakers to understand rental price dynamics in highly urbanised regions.

Acknowledgement

The authors would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, for its support

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Nur Suhaila Bdrul Hisham;* ***solving the governing equation****: Nur Suhaila Bdrul Hisham, Azme Khamis;* ***data collection****: Nur Suhaila Bdrul Hisham;* ***analysis and interpretation of results****: Nur Suhaila Bdrul Hisham, Azme Khamis;* ***draft manuscript preparation****: Nur Suhaila Bdrul Hisham, Azme Khamis. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Zulkifli, A. B., Kogid, M., Pinjaman, S., & Malin, J. (2022). Macroeconomic variables and the Malaysia House Price Index. International Journal of Academic Research in Economics and Management Sciences, 11(1). https://doi.org/10.6007/ijarems/v11-i1/12214

[2] Department of Statistics Malaysia. (2020). Population and Housing Census of Malaysia, 2020. https://www.dosm.gov.my

[3] Kadir, J. M. A., Subramaniam, G., & Ghazali, I. (2020). Housing Affordability among Middle-Income Earners in Selangor, Malaysia. Environment-Behaviour Proceedings Journal, 5(15), 289–293. https://doi.org/10.21834/ebpj.v5i15.2514

[4] Leh, O. L. H., Rosli, N. N. F. M., Marhalim, F. A., & Musthafa, S. N. a. M. (2017). Social Impact Of Foreign Immigrants In Affordable Housing Area. Case Study: Mentari Court, Selangor, Malaysia. Planning Malaysia, 15. https://doi.org/10.21837/pm.v15i2.266

[5] National Property Information Centre. (2025). Residential property price update 2015–2024. Valuation and Property Services Department, Ministry of Finance, Malaysia. https://napic.jpph.gov.my

[6] PropertyGuru Malaysia. (2024). Malaysia property rental market index: Q4 2024 report. PropertyGuru Malaysia. https://www.propertyguru.com.my/property-guides

[7] Chan, H. R. (2024). Rent Price Prediction with Advanced Machine Learning Methods: A Comparison of California and Texas. Highlights in Science Engineering and Technology, 85, 501–510. https://doi.org/10.54097/84vvv580

[8] Ramli, F. (2016). Equilibrium price application modelling for affordable housing market in Malaysia. The European Proceedings of Social & Behavioural Sciences, 431–436. https://doi.org/10.15405/epsbs.2016.08.61

[9] Tan, S. Y., Maimun, N. H. A., & Daud, S. Z. B. (2025). The Power of Place: How Location Drives Residential Rental Prices in Kuala Lumpur Educational Zones. Civil and Environmental Engineering Reports, 35(2), 24–34. https://doi.org/10.59440/ceer/202030

[10] Analysis of rent pricing in KL and Selangor [Computer software]. (n.d.). Kaggle. https://www.kaggle.com/code/aysdf7/analysis-of-rent-pricing-in-kl-and-selangor/input

[11] Reddy, K. S., Athelli, T., Kulsum, S., & Stanley College of engineering and technology for women, Hyderabad, India. (2022). Analysis and Prediction Of House Price. Journal of Emerging Technologies and Innovative Research, 9(5). https://www.jetir.org

[12] Chapter 1 Simple Linear Regression. (n.d.). [Book-chapter]. In Part 2. https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%201_part2.pdf

[13] Lee, M., & Han, C. (2024). Ordinary least squares and instrumental-variables estimators for any outcome and heterogeneity. The Stata Journal Promoting Communications on Statistics and Stata, 24(1), 72–92. https://doi.org/10.1177/1536867x241233645

[14] Gorriz, J., Ramirez, J., Segovia, F., Jimenez-Mesa, C., Martinez-Murcia, F., & Suckling, J. (2025). Statistical agnostic regression: A machine learning method to validate regression models. Journal of Advanced Research, 80, 503–533. https://doi.org/10.1016/j.jare.2025.04.026

[15] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

[16] Frifra, A., Maanan, M., Maanan, M., & Rhinane, H. (2024). Harnessing LSTM and XGBoost algorithms for storm prediction. Scientific Reports, 14(1), 11381. https://doi.org/10.1038/s41598-024-62182-0

[17] Sapkota, S. C., Shrestha, A., & Tang, W. (2025b). Metaheuristic-optimized machine learning models for accurate prediction of limestone calcined clay cement (LC3) compressive strength. Case Studies in Construction Materials, 23, e05303. https://doi.org/10.1016/j.cscm.2025.e05303

[18] Shrivas, S., & Dubey, A. (2025). Prediction of specific capacitance of activated carbon electrode for energy storage device by machine learning based approach. Journal of Energy Storage, 121, 116632. https://doi.org/10.1016/j.est.2025.116632