Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 11, No. 05, May, 2025

eXplainable Machine Learning for Real Estate: XGBoost and Shapley Values in Price Prediction

Tris Kee [1], Winky K. O. Ho [1*]

[1] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong.

Received 13 February 2025; Revised 19 April 2025; Accepted 26 April 2025; Published 01 May 2025

Abstract

This study examines the application of eXplainable Artificial Intelligence (XAI) in property market research, utilizing housing transaction data from Quarry Bay, Hong Kong. The research employs the XGBoost algorithm to predict property prices and subsequently computes Shapley Additive Explanations (SHAP) values to quantify feature importance. A beeswarm plot is used to visualize the distribution of SHAP values, uncovering complex relationships between prices and property characteristics. The findings demonstrate how features such as square footage and property age contribute to average price predictions, offering valuable insights for urban planning and real estate decision-making. In contrast to the traditional black-box models, this study integrates XAI methodologies to enhance model interpretability, thereby fostering trust in AI-driven market analyses. The novelty of this research lies in its combination of machine learning and explainable techniques, bridging the gap between predictive accuracy and interpretability in property valuation. By advancing data-driven decision-making, this study underscores the potential of XAI in promoting transparency and facilitating informed policymaking in the property market.

Keywords: Property Prices; XGBoost; Shapley Value.

1. Introduction

The residential property market in Hong Kong constitutes a dynamic and complex ecosystem influenced by a diverse range of factors, including economic trends, demographic shifts, government policies, and local market conditions. As one of the most expensive and volatile real estate markets globally, a comprehensive understanding of the determinants of property prices is essential for various stakeholders, including investors, developers, policymakers, and urban planners. Traditional valuation methods, such as hedonic pricing models, have long been utilized to estimate property prices. However, these conventional approaches often struggle to capture the intricate and nonlinear relationships between property attributes and market dynamics. The increasing availability of large-scale property transaction data has necessitated the adoption of more advanced analytical techniques that not only enhance the predictive accuracy but also offer deeper insights into key drivers of property prices.

Recent advancements in machine learning (ML) have introduced highly effective predictive models capable of capturing the complex relationships between property prices and property attributes. Particularly, Extreme Gradient Boosting (XGBoost) has emerged as one of the most robust algorithms for property price prediction. It can model the nonlinear relationships, handle missing data, and mitigate overfitting through regularization techniques [1]. While XGBoost has demonstrated superior predictive performance, the "black-box" nature of ML models limits their interpretability. This lack of transparency raises significant concerns in high-stakes applications such as real estate valuation, where trust and explainability are crucial for informed decision-making.

* Corresponding author: winky.ho@polyu.edu.hk

http://dx.doi.org/10.28991/CEJ-2025-011-05-022

To address this challenge, the field of eXplainable Artificial Intelligence (XAI) has gained prominence, aiming to enhance model interpretability without compromising predictive performance. Among the most effective XAI techniques is the use of Shapley values [2, 3], which provide a rigorous, game-theoretic approach to quantifying the contribution of individual features to model predictions. By integrating Shapley values with XGBoost, this study seeks to identify the primary determinants of residential property prices in Hong Kong and to explain the complex relationships between prices and property characteristics. The visualization of these contributions through SHAP summary and beeswarm plots facilitates better understanding of how features contribute to the average price predictions.

Despite the increasing application of ML in real estate research, limited attention has been given to integrating predictive accuracy with interpretability, particularly in the context of the Hong Kong property market. Existing studies largely focus on enhancing predictive performance while often overlooking the necessity of transparency and explainability. This study addresses these gaps by implementing an XAI-driven framework that not only improves the accuracy of property price predictions but also provides meaningful insights into the underlying market dynamics.

The objectives of this research are fourfold: (1) to apply XGBoost for predicting residential property prices in Hong Kong, (2) to utilize Shapley values for feature importance analysis to enhance model interpretability, (3) to visualize the influence of property characteristics on price predictions through SHAP summary and beeswarm plots, and (4) to generate data-driven insights that can inform urban planning, investment strategies, and housing policy development. By bridging the gap between predictive performance and interpretability, this study contributes to a more transparent and accountable application of machine learning in real estate valuation.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on the application of XGBoost and Shapley values in real estate research. Section 3 outlines the methodological framework adopted in this study. Section 4 describes the data sources and key features used in the analysis. Section 5 presents the exploratory data analysis, highlighting initial observations and patterns within the dataset. Section 6 details the findings, emphasizing the significance of key property characteristics and their impact on market trends. Finally, Section 7 concludes the paper by summarizing key contributions and discussing the broader implications of the findings for real estate research and policy development.

2. Literature Review

The adoption of machine learning, particularly XGBoost, has significantly transformed property price prediction by offering superior efficiency, scalability, and predictive accuracy. Kumkar et al. [4] conducted a comparative analysis of ensemble methods, including Random Forest (RF), Gradient Boosting Machine (GBM), and XGBoost, for real estate price prediction in Mumbai. Their findings indicate that XGBoost exhibits the highest predictive accuracy, as measured by the mean absolute percentage error (MAPE), due to its capability to incorporate a wide range of features such as location, property attributes, and available amenities. This study not only highlights the superiority of XGBoost but also the potential of ensemble learning methodologies in real estate analytics.

Stang et al. [5] further emphasize XGBoost's accuracy in developing automated valuation models (AVMs) for the German real estate market. Their study also acknowledges the advantages of traditional econometric models, such as Ordinary Least Squares (OLS) and Generalized Additive Models (GAM), in regional real estate markets. Their results highlight the importance of integrating multiple algorithmic approaches to ensure robust and reliable property valuation models. Similarly, Calainho et al. [6] explore the application of non-linear models, including XGBoost, within a model-agnostic framework for constructing property price indices. Utilizing a dataset of 29,998 commercial real estate transactions in New York, their study demonstrates the superior predictive performance of machine learning algorithms over linear regression models. However, their study also identifies some challenges related to data dependency and model instability, particularly when applied to smaller datasets, thereby underscoring the necessity for robust strategies to mitigate the issues of bias and variance.

While XGBoost and other machine learning models have demonstrated remarkable predictive performance, their lack of interpretability is a major barrier to adoption. The opacity of these models complicates efforts to understand the decision-making processes underlying the price predictions, raising concerns regarding accountability and trust. To address this challenge, Shapley value-based techniques, such as SHapley Additive Explanations (SHAP), have emerged as essential tools for enhancing model transparency.

Lenaers & De Moor [7] compare various XAI techniques using the CatBoost algorithm for rental price prediction in Belgium. Their study concludes that no single interpretability method offers a complete explanation and proposes using a hybrid approach with multiple XAI techniques for a better understanding of property rental prices. Trindade Neves et al. [8] and Jin et al. [9] similarly apply SHAP to assess feature contributions in real estate price predictions, with the latter identifying demographic and socioeconomic variables as critical features to explain property price appreciations during the period between 2012 and 2018 in Los Angeles County.

This study aims to address the research gap in balancing predictive accuracy and interpretability in property price modeling by integrating machine learning techniques with XAI methodologies. While prior research has demonstrated the effectiveness of XGBoost in real estate analytics (Calainho et al. [6]; Kok et al. [10]; Yang & Cao [11]; Hendrayati et al. [12]; Burnwal & Jaiswal [13]), issues related to model transparency persist, limiting its practical adoption among policymakers, investors, and urban planners. To bridge this gap, the present study proposes an integrated framework that enhances both predictive accuracy and interpretability, thereby improving trust and usability in property valuation models.

The dataset employed in this research comprises real estate transaction records, ensuring a diverse representation of property types, locations, and relevant physical and accessibility characteristics. The data preprocessing phase will include procedures for handling missing values and outliers to optimize model performance. XGBoost will be the primary predictive model, given its well-documented advantages over traditional statistical methods and other ensemble approaches. To ensure a rigorous evaluation, the performance of XGBoost will be benchmarked against alternative models, including RF and GBM. Model accuracy will be assessed using key evaluation metrics such as MAPE, ensuring consistency with previous empirical studies.

Beyond predictive modeling, this study prioritizes model explainability, addressing the concerns related to interpretability that have been raised in prior research. Unlike studies that focus predominantly on improving predictive accuracy, this research establishes a structured framework that balances predictive power with transparency. By systematically evaluating the trade-off between model performance and interpretability, this study contributes to the development of more robust, transparent, and policy-relevant real estate valuation models. To achieve this objective, SHAP techniques will be implemented, including summary plots, beeswarm plots, and dependency plots, to evaluate their effectiveness in uncovering complex relationships between property prices and key features.

By advancing the integration of machine learning and XAI in real estate analytics, this study seeks to enhance transparency and accountability in property valuation methodologies. The findings are expected to provide valuable insights for stakeholders, including urban planners, policymakers, and real estate professionals, facilitating more informed and data-driven decision-making processes.

3. Research Methodology

XGBoost was selected over alternative machine learning algorithms, including RF (Breiman [14, 15], Banerjee and Dutta [16], Choy & Ho [17]) and GBM (Friedman [18], Ho et al. [19]), due to its superior predictive accuracy, computational efficiency, and scalability. Recognized as the "winning algorithm" in numerous high-profile machine learning competitions, including Kaggle (DataScientest [20]), XGBoost has gained popularity for its optimized gradient boosting framework, which incorporates bagging and regularization techniques to mitigate overfitting. Compared to RF, XGBoost achieves higher accuracy by sequentially refining weak learners, while its parallelized implementation and tree-pruning strategies enhance its computational efficiency relative to standard GBM. These advantages render XGBoost particularly well-optimized for analyzing complex real estate markets characterized by large datasets and intricate feature interactions.

The XGBoost algorithm, developed by Chen & Guestrin [1], is a highly efficient and robust tree-based machine learning technique. Building upon Friedman's Gradient Tree Boosting algorithm, XGBoost integrates both bagging and boosting strategies within an ensemble learning framework. Bagging enhances model stability and predictive accuracy by training multiple models in parallel on independently sampled subsets, while gradient boosting constructs decision trees sequentially to correct the residual errors from previous iterations. This iterative approach continues until the predefined number of trees is reached or further performance improvements become negligible.

XGBoost initializes with an initial prediction, typically the mean of the target variable, and iteratively refines it by training decision trees to predict residual errors. These predictions are scaled and incrementally added to the cumulative total to mitigate overfitting. This methodology has demonstrated high effectiveness across both regression (Kumkar et al. [4]; Lenaers & de Moor [7]; Trindade Neves et al. [8]) and classification tasks (Mahesh et al. [21]; Zhang et al. [22]), contributing to its widespread adoption in both academic research and industry applications.

XGBoost's objective function combines two key components: a differentiable convex loss function (l) that measures prediction error, and a regularization term (Ω) that prevents overfitting by penalizing model complexity. For classification, the loss function may use log loss, while regression tasks typically employ the mean squared error. Regularization involves both L1 (Lasso) and L2 (Ridge) penalties, with Ω expressed as $\gamma T + \frac{1}{2}\lambda||w||^2$, where γ and λ are the hyperparameters, T is the number of leaves, and w represents the leaf weights. This composite objective function balances model's fit and complexity, improving generalization and predictive performance.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{1}$$

Several features of XGBoost contribute to mitigating the risk of overfitting. Regularization parameters, such as lambda and alpha, regulate model complexity, while gamma prunes trees to prevent excessive depth. The learning rate (eta) ensures that the model learns gradually, and early stopping curtails training to reduce the extent of overfitting. Additionally, optimizing hyperparameters such as max_depth, min_child_weight, subsample, and colsample_bytree can further improve model performance.

Hyperparameter optimization is essential for maximizing the effectiveness of XGBoost. In this study, Optuna, a Bayesian optimization framework (Akiba et al. [23], Kee & Ho [24, 25]), is employed to efficiently explore the hyperparameter space, identifying configurations that improve model performance. The model's effectiveness is assessed using a comprehensive evaluation framework. This includes metrics like the coefficient of determination (R^2) on both training and test sets. Other metrics are also considered to provide a thorough assessment of predictive accuracy. This multi-metric evaluation offers insights into the model's strengths and limitations in addressing specific challenges.

The error metrics are formally defined as follows:

$$MAE = \frac{|(h(x^{(i)}) - y^{(i)})|}{m} \tag{2}$$

$$MSE = \frac{1}{m}\sum_{\mathrm{i=1}}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{3}$$

$$MAPE = \frac{100\%}{m}\sum_{\mathrm{i=1}}^{m}\left|\frac{(h(x^{(i)}) - y^{(i)})}{y^{(i)}}\right| \tag{4}$$

$$MSE = \sqrt{\frac{1}{m}\sum_{\mathrm{i=1}}^{m}(h(x^{(i)}) - y^{(i)})^2} \tag{5}$$

where $h\left(x^{(i)}\right)$ represents the predicted value; $y^{(i)}$ represents the actual value of the target variable Y; and m represents the number of rows in the test data.

This research applies the SHapley Additive exPlanations (SHAP) method (SHAP [26]), an advanced approach in XAI. SHAP utilizes Shapley values (Shapley [2]) from cooperative game theory to provide reliable and precise interpretations of model outputs (Lundberg & Lee [27]; Lundberg et al. [28]). In this framework, features are treated as players in a cooperative game, and model's prediction is the payoff. Shapley values fairly allocate overall prediction among features, ensuring each feature's contribution is equitably and interpretable quantified.

To calculate Shapley values, this study employs TreeExplainer algorithm, which adheres to axioms of additive feature attribution methods. This can be mathematically expressed as:

$$f(x) = \phi_0 + \sum_{i=1}^{M}\phi_i \tag{6}$$

where $f(x)$ represents the model's prediction for a specific instance, ϕ_0 represents the reference value, which is the mean prediction calculated across all samples in the dataset. ϕ_i represents the Shapley value, reflecting the impact of feature i on the model's prediction. The Shapley value ϕ_i for a given feature i is computed as follows:

$$\phi_i = \sum_{S\subseteq N\backslash\{i\}}\frac{|S|!(|N|-|S|-1)!}{|N|!}[v(S\cup\{i\}) - v(S)] \tag{7}$$

where N is the set of all features, S is any subset of features excluding feature i, and $v(S)$ represents the value function, which corresponds to the model's prediction based on the subset S of features.

First, we train an XGBoost regression model using features such as square footage, property age, floor level, rooftop, number of carparks, availability of an MTR station, and directional variables (East, South, West, North, Northeast, Southeast and Southwest). The model learns how these features influence property prices, capturing the complex, non-linear relationships. Second, after training the model, we apply TreeExplainer, which is specifically designed for tree-based models and efficiently computes exact Shapley values. The SHAP algorithm analyzes how each feature contributes to predicted price by evaluating all possible combinations of feature subsets. For a given property, it measures how the inclusion or exclusion of each feature alters model's output, averaging these marginal contributions across all possible feature subsets. This process follows Shapley formula, which ensures a fair distribution of contribution among features.

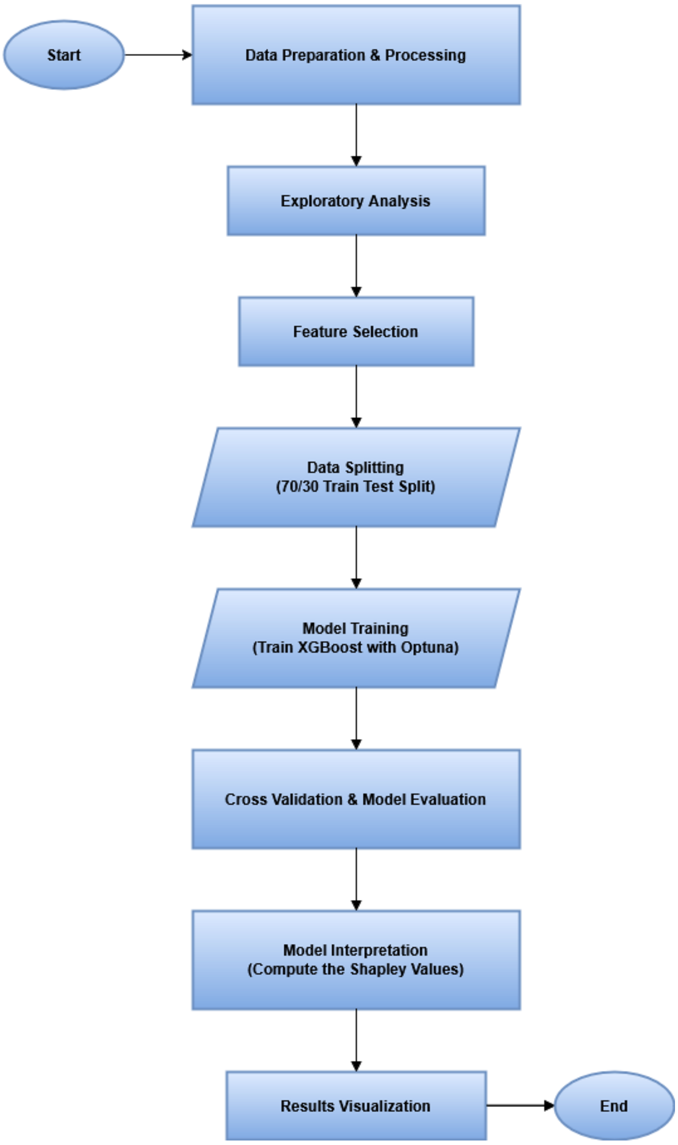

Figure 1. Research methodology workflow

4. Data Definitions and Sources

This study utilizes a dataset of 24,936 transaction records from four popular housing estates in the Quarry Bay District of Hong Kong Island, covering the period from May 1997 to May 2021. Records with missing information such as GFA and others are excluded from the analysis. To mitigate the influence of outliers and potential errors, we remove 5% of observations with the lowest and highest transaction prices. Details on private residential properties such as geographic location, transaction and occupation permit issuance dates, transaction prices, gross floor area, floor levels, and additional attributes like parking spaces and rooftops are included in our dataset. Sourced from official government records aggregated by EPRC, property prices are inflation-adjusted using the Residential Property Price Index published by the Hong Kong SAR Government (Rating and Valuation Department [29]). This adjustment ensures the standardized prices across time periods, enabling a more accurate analysis of real changes in property values. Definitions of variables used in the analysis are provided below.

RP denotes the transaction price of a residential property, expressed in HK$ million, inflation adjusted.

GFA denotes the size of a residential property, measured in square feet.

AGE denotes the age of a property at the time of the transaction. It is obtained by subtracting the issuance date of the occupation permit from the transaction date.

FL denotes the floor level of the residential property.

$ROOF$ is a dummy variable that takes the value of 1 if the property includes a rooftop and 0 otherwise.

CP denotes the number of car parks sold together with the residential property.

MTR is a dummy variable that equals 1 if the walking distance from the property to the nearest Mass Transit Railway (MTR) station is within 10 minutes, 0 otherwise.

E, S, W, N, NE, SE, SW & NW represent the orientation of the property. These are dummy variables where a value of 1 indicates that the property faces a particular direction. For comparison, NW is omitted from the analysis, and the coefficients of the other orientations are interpreted relative to this reference category.

5. Exploratory Data Analysis

Figure 2 presents histograms for each feature, providing an overview of their approximate probability distributions by showing observation frequencies across specified intervals. Figure 3 features a correlation matrix, emphasizing linear associations among all dataset features, with the analysis identifying a strong positive correlation (r = 0.8) between gross floor area and property prices. Moderate correlations are also observed between property prices and the following features: floor level (r = 0.4), property age and proximity to the MTR (r = –0.3 or 0.3), northeast orientation (r = 0.2), and features such as rooftop, car parks, and orientations facing south, west, north, or northwest (r = 0.1 or –0.1). For correlations among features, its value for floor level and age is –0.5, indicating a moderate inverse relationship where higher floors are generally associated with newer buildings. The other features in our dataset show correlations ranging between –0.001 and 0.2, which are considered weak or negligible.

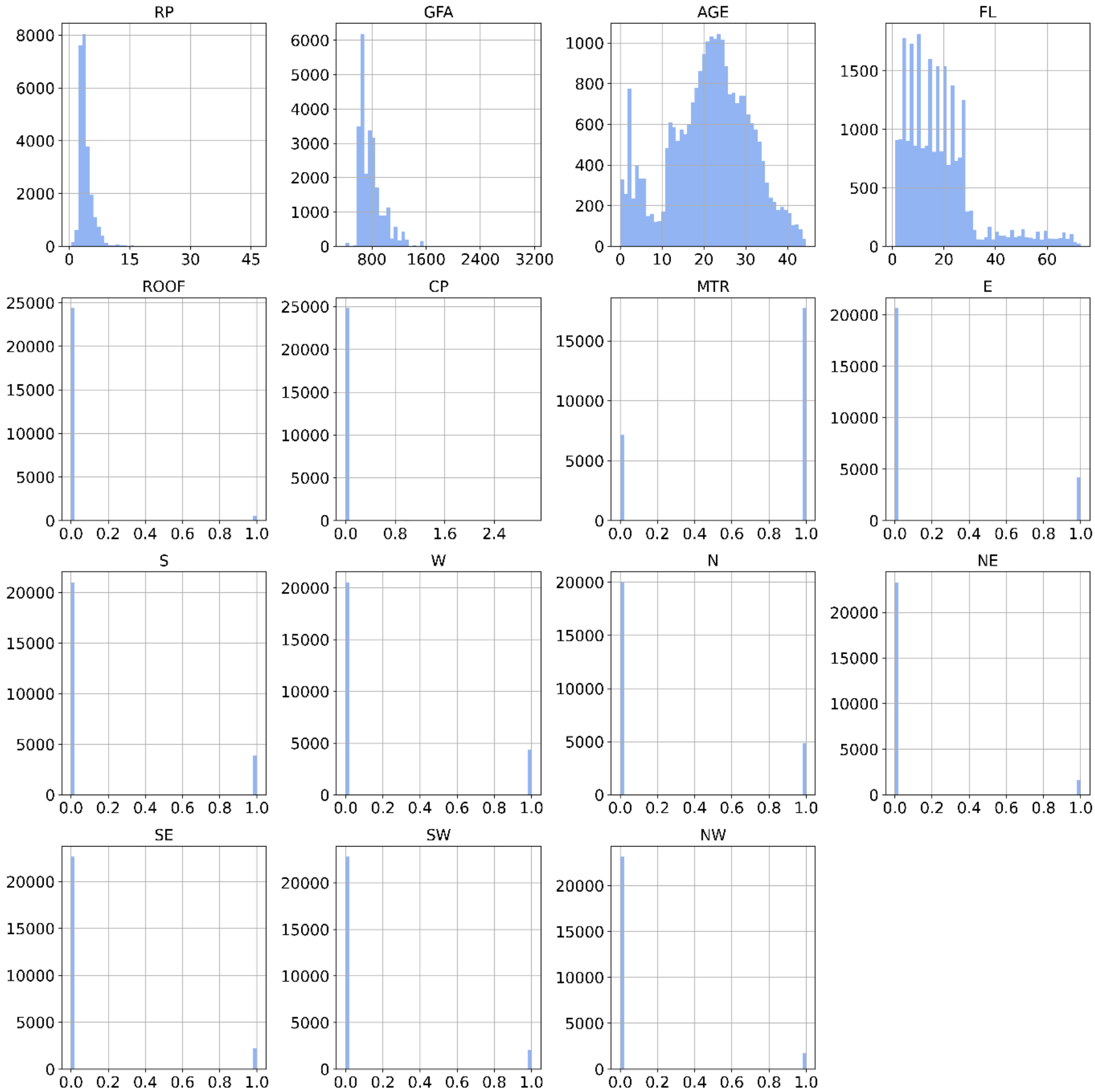

Figure 2. Distribution of property prices and key features

Figure 4 offers a visual representation of correlations based on Figure 3, highlighting relationships between property prices and features. This visualization enhances interpretability of the correlation matrix by presenting data in an intuitive format. Furthermore, Table 1 presents a concise summary of descriptive statistics for our dataset, offering an overview of central tendencies, dispersions, and distributions. The target variable is transacted property prices, inflation adjusted. The mean transaction price is approximately HK$4.05 million, with a standard deviation of HK$1.81 million (44.58%

of the mean). The mean gross floor area is 784.28 square feet, with a standard deviation of 178.15 square feet (22.71% of the mean). The average age of transacted units is 21.34 years, with a standard deviation of 9.72 years (45.56% of the mean). Furthermore, the mean floor level is 18.29 floors, with a standard deviation of 13.76 floors (75.25% of the mean). Furthermore, the mean number of carparks is 0.002, with a standard deviation of 0.06 (2,343.06% of the mean).

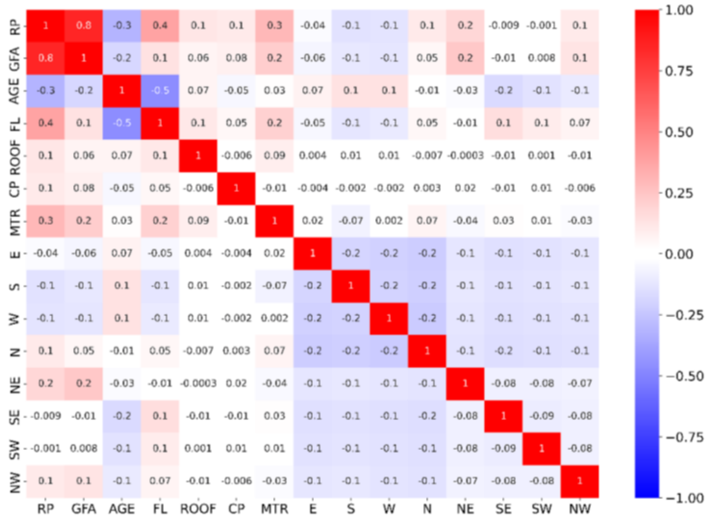

Figure 3. Correlation matrix of property prices and key features

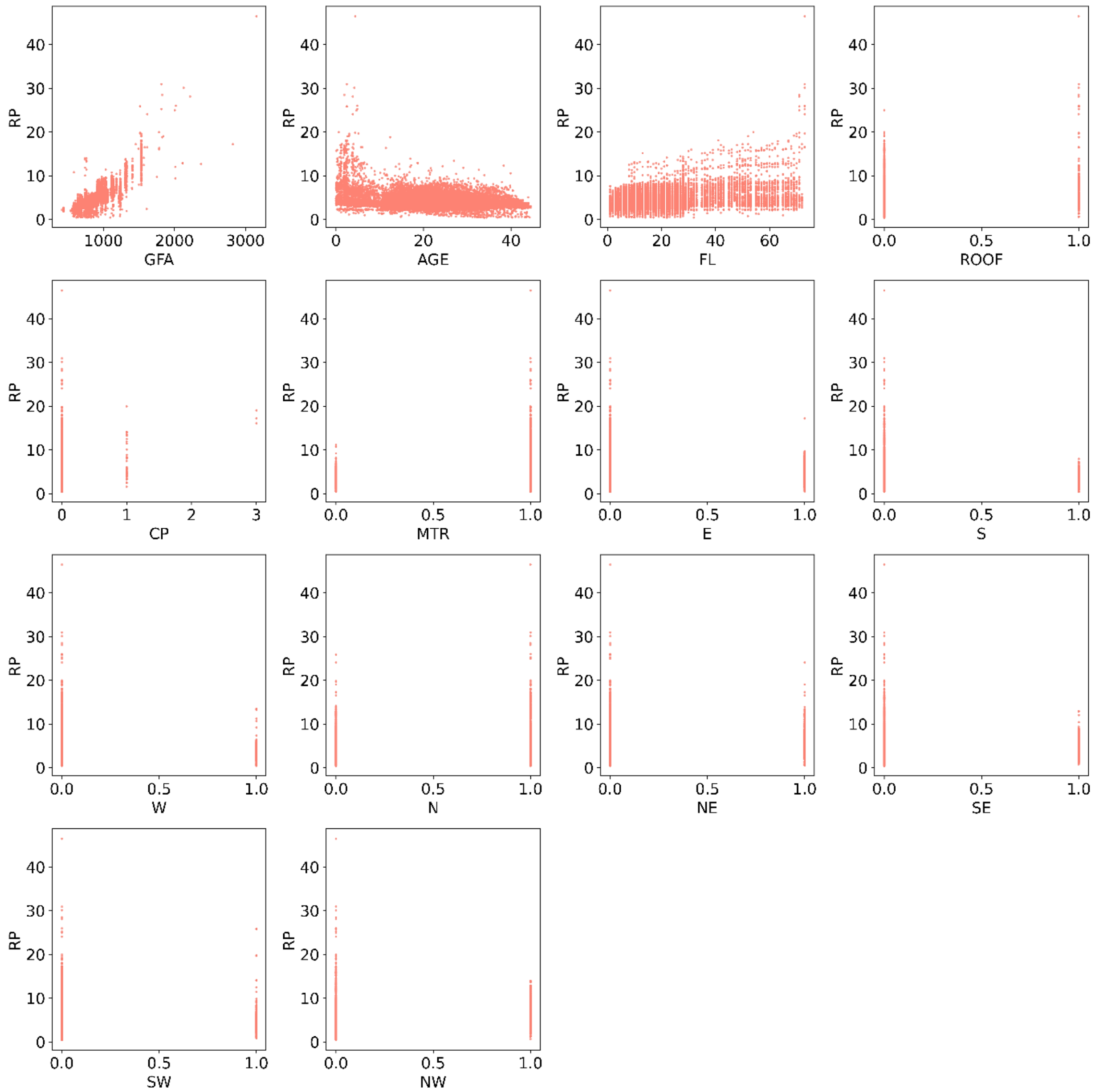

Figure 4. Data visualization of property prices and key features

Table 1. Descriptive statistics

	RP	GFA	AGE	FL	FR	ROOF	CP	MTR	E	S	W	N	NE	SE	SW	NW
Count	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863	24863
Mean	4.053	784.281	21.341	18.288	0.020	0.002	0.713	0.169	0.156	0.175	0.196	0.064	0.089	0.083	0.069	4.053
Std	1.807	178.154	9.724	13.762	0.139	0.056	0.452	0.375	0.363	0.380	0.397	0.244	0.285	0.275	0.253	1.807
Min	0.440	413	0.003	1	0	0	0	0	0	0	0	0	0	0	0	0.440
25%	2.996	675	15.189	8	0	0	0	0	0	0	0	0	0	0	0	2.996
50%	3.594	752	22.063	16	0	0	1	0	0	0	0	0	0	0	0	3.594
75%	4.570	858	28.234	24	0	0	1	0	0	0	0	0	0	0	0	4.570
Max	46.463	3155	44.449	73	1	3	1	1	1	1	1	1	1	1	1	46.463
Skew	3.644	1.654	–0.258	1.584	6.896	30.528	–0.943	1.764	1.898	1.714	1.529	3.580	2.888	3.032	3.407	3.644

6. Results and Discussion

Table 2 presents the hyperparameter space and optimal values identified by Optuna. The dataset is split into 70% for training and 30% for testing. Hyperparameter optimization is conducted iteratively using five-fold cross-validation to determine the best parameter values. Model performance and generalizability are evaluated with five-fold cross-validation. In each iteration, four folds, comprising 80% of the training set (equivalent to 56% of the entire dataset), are utilized for model training, while the remaining fold, representing 20% of the training set (or 14% of the entire dataset), is designated as the validation set. Each fold is used as the validation set once during the process.

Table 2. Hyperparameter spaces

	GBM	RF	XGBoost
booster			gbtree
bootstrap		True	
colsample_bylevel			0.7, 0.8, …, 1.0 (0.7)
colsample_bynode			0.7, 0.8, …, 1.0 (0.7)
colsample_bytree			0.7, 0.8, …, 1.0 (1.0)
criterion	friedman_mse	friedman_mse	
eval_metric			rmse
gamma			30.0, 30.1, …, 31.0 (30.8)
learning_rate	0.08, 0.09, 0.10 (0.10)		0.05, 0.06, …, 0.10 (0.09)
loss	squared_error		
max_depth	3	5, 6, …, 10 (10)	10, 11, …, 15 (14)
max_features	2, 3, …, 13 (12)	2, 3, …, 13 (9)	
max_samples		0.6, 0.7, 0.8 (0.8)	
min_child_weight			2.0, 2.1, …, 3.0 (2.4)
min_samples_leaf	13, 14, …, 17 (14)	36, 37, 38 (36)	
min_samples_split	18, 19, …, 22 (18)	36, 37, 38 (38)	
n_estimators	50, 60, …, 100 (100)	50, 60, …, 100 (60)	1100, 1110, …, 1200 (1180)
objective			reg:squarederror
reg_alpha			8
reg_lamda			7
subsample	0.8, 0.9, 1.0 (1.0)		0.8, 0.9, 1.0 (1.0)
tree_method			exact

Note: Figures in parentheses are the optimal hyperparameters' values tuned by Optuna.

Table 3 summarizes results for GBM, RF, and XGBoost for both training and test sets. The highest R^2 is achieved by XGBoost at 0.908, followed by GBM (0.902) and RF (0.876). Performance metrics show less than a 5% difference between training and test sets for all models, indicating strong generalization and an effective balance between bias and variance, reducing the risk of overfitting or underfitting. The difference between training and test evaluation metrics for XGBoost falls within a relatively small range (0.501% to 3.728%). This indicates that the model's performance is consistent across different data splits, which is a good sign of generalization. Overfitting typically occurs when a model performs significantly better on the training set than on the test set, meaning it has learned noise or specific patterns rather than generalizable relationships. In our paper, small differences in evaluation metrics, such as MAE, MSE, MAPE and RMSE, suggest that the model is not memorizing the training data but rather learning patterns that apply to unseen data as well. If the model were overfitting, we would expect to see a much larger discrepancy between training and test set performance. A highly overfitted model would exhibit minimal error on the training set but demonstrate poor performance on the test set due to its limited capacity for generalization.

Table 3. Evaluation metrics

	GBM	RF	XGBoost
R^2	0.90280 (0.90225) {–0.06122%}	0.87620 (0.87613) {–0.00835%}	0.90774 (0.90764) {–0.01056%}
MAE	0.33418 (0.34452) {3.09509%}	0.31396 (0.32928) {4.88213%}	0.32616 (0.33515) {2.75663%}
MSE	0.31388 (0.32710) {4.20955%}	0.39978 (0.41450) {3.68149%}	0.29794 (0.30904) {3.72798%}
MAPE%	10.19426% (10.66991%) {0.47565%}	9.42972% (10.04153%) {0.61181%}	10.09720% (10.59824%) {0.50103%}
RMSE	0.56025 (0.57192) {2.08308%}	0.63228 (0.64382) {1.82411%}	0.54584 (0.55592) {1.84693%}
Average CV R^2	0.87644 (0.87852) {0.23717%}	0.86737 (0.84295) {–2.81547%}	0.88731 (0.87907) {–0.92957%}

Notes: Figures in parentheses are the values for test set; figures in curly parentheses are the difference in values between training and test sets.

The acceptable threshold for overfitting varies depending on the specific application and dataset. A difference of less than 5% between training and test performance metrics is typically considered minimal, whereas differences exceeding 10% may indicate overfitting. In this study, the observed differences across all three algorithms range from 0.501% to 4.883%, suggesting minimal overfitting and well-optimized models for generalization. Low error metrics for both the training and test sets, combined with high R^2 values, confirm the absence of underfitting, wherein a model fails to capture essential data patterns. Overall, integration of hyperparameter tuning, cross-validation, and performance evaluation has resulted in well-optimized GBM, RF, and XGBoost models, demonstrating exceptional predictive accuracy and robust generalization across both datasets.

Figure 5 presents scatterplots illustrating actual residential property prices alongside residuals for the test set. The results indicate that XGBoost generally achieves a strong fit with the data. However, the model demonstrates high bias when dealing with extreme property price values, as highlighted by the green dots representing prices exceeding HK$20 million. Moreover, Figures 6 illustrates scatterplots comparing actual and predicted residential property prices for the test set, based on XGBoost model's results. The plot demonstrates a strong positive correlation between actual and predicted values, with the majority of data points aligning closely along the red reference line, signifying a high degree of agreement between model's predictions and observed values. Although a few outliers are present, the strong alignment between actual and predicted values suggests that the model effectively captures underlying patterns, achieving high predictive accuracy. The observed deviations in some instances may be attributable to data noise or unmodeled features not accounted for by our XGBoost model.

The final stage of analytical process applies SHAP value analysis to enhance interpretability of XGBoost model's predictions for residential property prices. Using TreeExplainer module within SHAP framework, the analysis quantifies relative importance of individual features and their contributions, either positive or negative, to predicted outcomes. This approach provides a nuanced understanding of features influencing property prices by illustrating impact of each feature on the model's decision-making process.

Based on our XGBoost model's results, Figures 7 and 8 present a comprehensive SHAP value analysis, offering insights into most influential features contributing to property price predictions. Figure 7, a SHAP summary plot, ranks features by their mean absolute SHAP values, which quantify average magnitude of each feature's contribution to predictions. A larger mean absolute SHAP value signifies a greater overall impact of feature. Unlike conventional feature importance metrics, SHAP values are more intuitive since they are expressed in the same units as the target variable, which in this case is millions of HK dollars.

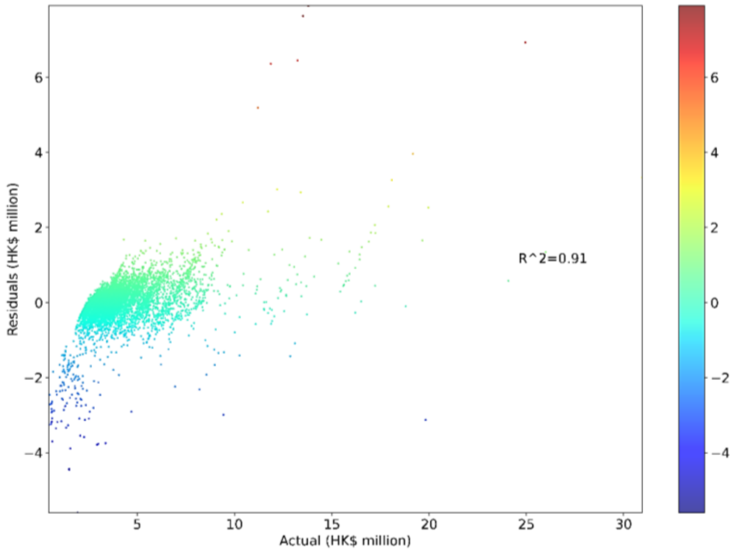

Figure 5. Actual residential property prices and residuals (Test set)

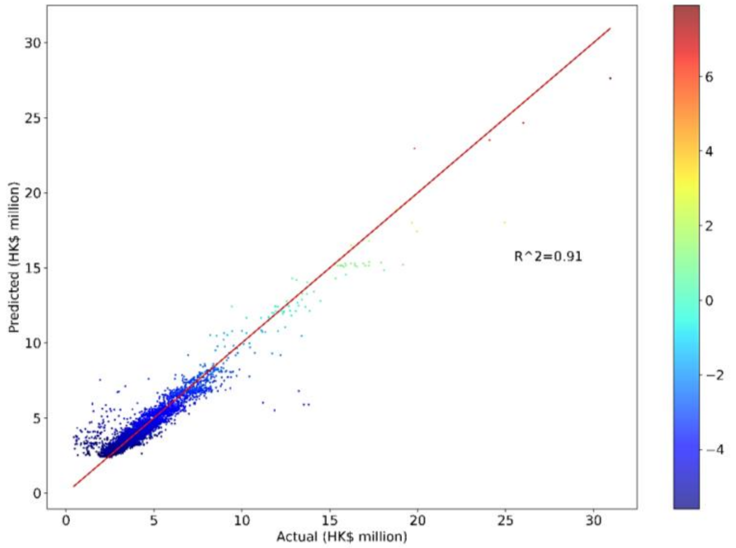

Figure 6. Actual and predicted residential property prices (Test set)

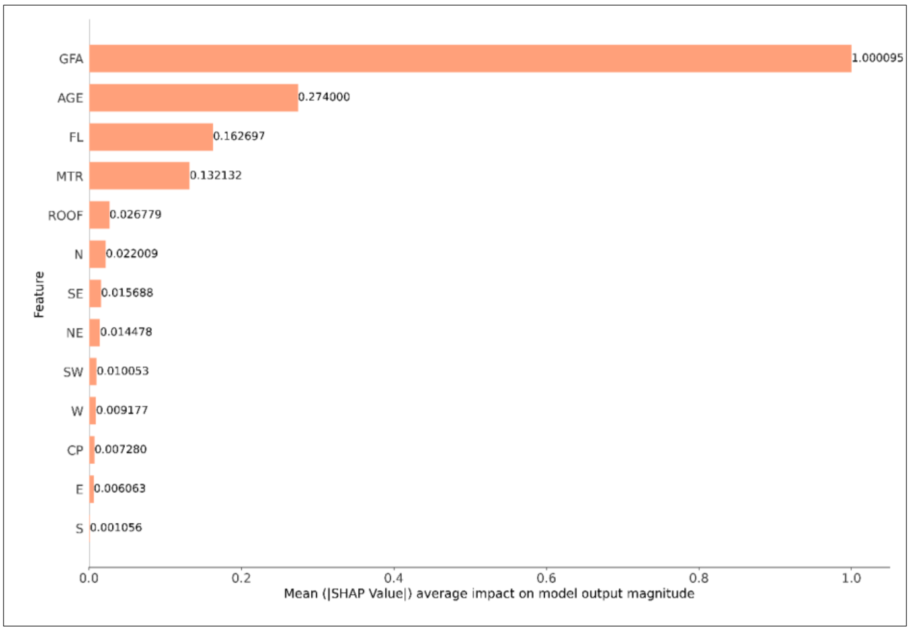

Figure 7. SHAP summary plot, in HK$ million (Test set)

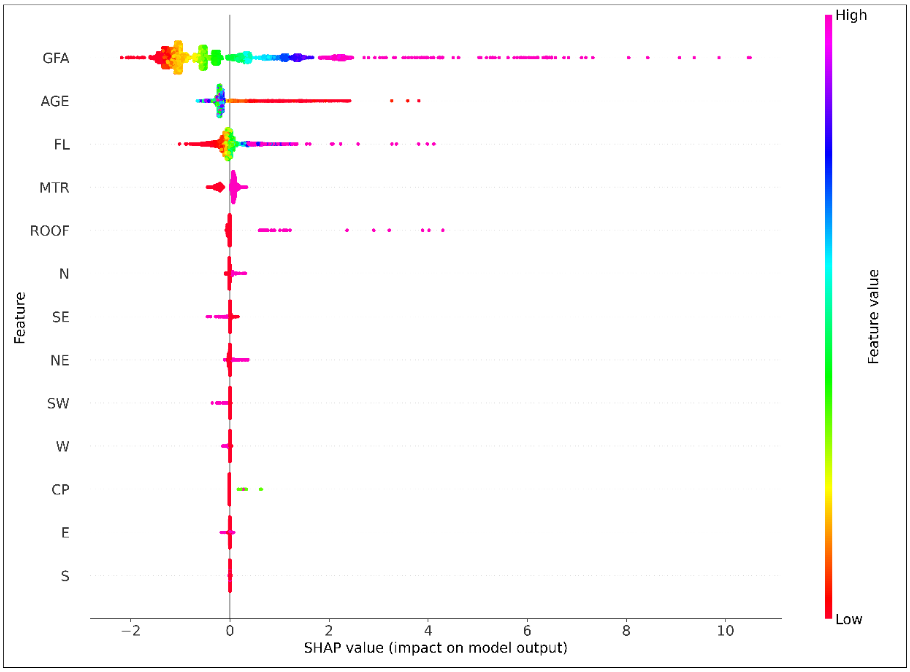

Figure 8. SHAP beeswarm plot, in HK$ million (Test set)

The analysis identifies gross floor area as most impactful feature, contributing approximately HK$1,000,095 (inflation-adjusted) to property price predictions per observation. Features such as property age (HK$274,000), proximity to MTR stations (HK$162,697), and floor level (HK$132,132) are also very important. In contrast, feature with least influence, denoted as "S," contributes only HK$1,056 per observation. These findings underscore importance of physical attributes and accessibility in shaping property prices, while less significant features have minimal influence on predictions.

6.1. SHAP Dependency Plots

This subsection provides a set of SHAP dependency plots derived from the XGBoost model's results for the test set, designed to shed light on how specific features relate to the model's predictions. Each plot shows how the SHAP values, which represent the impact of individual features on the overall prediction, change as the feature values vary. These visualizations are crucial for gaining insights into the model's functioning and understanding how different features affect its predictions (see Figures 9 to 21).

While SHAP serves as a powerful tool for model interpretability, its application to high-dimensional real estate data presents several challenges. One primary issue is computational complexity. Although TreeExplainer is optimized for tree-based models, the computation of Shapley values still necessitates evaluating multiple feature subsets, which becomes increasingly computationally demanding as the number of features grows. This challenge is particularly pertinent in real estate datasets, which often encompass a wide range of location-based, structural, and financial attributes.

Feature collinearity poses an additional challenge. In real estate, numerous features exhibit interdependencies, such as square footage and property age. SHAP assumes that features contribute independently to model predictions; however, collinear features often share importance, complicating the interpretation of individual SHAP values. This interdependence may lead to misleading attributions, as a highly correlated feature may absorb the contribution of another. Although XGBoost is generally robust to multicollinearity, Shapley value analysis remains susceptible to its effects.

To assess the impact of multicollinearity on SHAP values, the Variance Inflation Factor (VIF) is computed for all features. The results indicate that GFA and AGE exhibit relatively high VIF values of 11.90 and 6.41, respectively, suggesting a moderate to strong correlation between them. Conversely, the VIF values for the remaining features range from 1.04 to 3.99, indicating low collinearity. In the presence of multicollinearity, the marginal contribution of a feature becomes highly dependent on whether its correlated counterparts are already included in the subset. If a feature is introduced into a subset that already contains a highly correlated feature, its marginal contribution may appear negligible. Conversely, if the feature is introduced first, its contribution may seem significant. This results in high variance in assigned Shapley values, reducing their stability across different sampling strategies or feature orderings. Standard SHAP feature importance techniques assume feature independence; however, this assumption is frequently violated in the presence of multicollinearity. Traditional feature permutation methods, which involve randomly shuffling individual features, disrupt the natural relationships between correlated features. This process introduces unrealistic feature combinations that do not exist in the training data, potentially leading to misleading SHAP values and erroneous interpretations of feature importance.

To mitigate the influence of multicollinearity on Shapley values, the feature permutation setting is specified as “interventional”. This approach enhances the stability of SHAP values by ensuring robust attributions even in the presence of multicollinearity. The interventional permutation method maintains inherent relationships among features by conditioning on the observed data distribution. Instead of randomly shuffling values, it preserves the natural structure of the data. This methodology prevents unrealistic perturbations in feature values. Such distortions could otherwise cause drastic fluctuations in model predictions. Consequently, SHAP values derived through this approach are more stable and interpretable.

By maintaining natural dependencies among correlated features, the interventional permutation method reduces extreme variations in SHAP values. These variations often arise when redundant features arbitrarily alternate in importance due to random permutations. As a result, this technique provides more reliable interpretations. It prevents the overestimation of one feature’s contribution at the expense of its correlated counterparts.

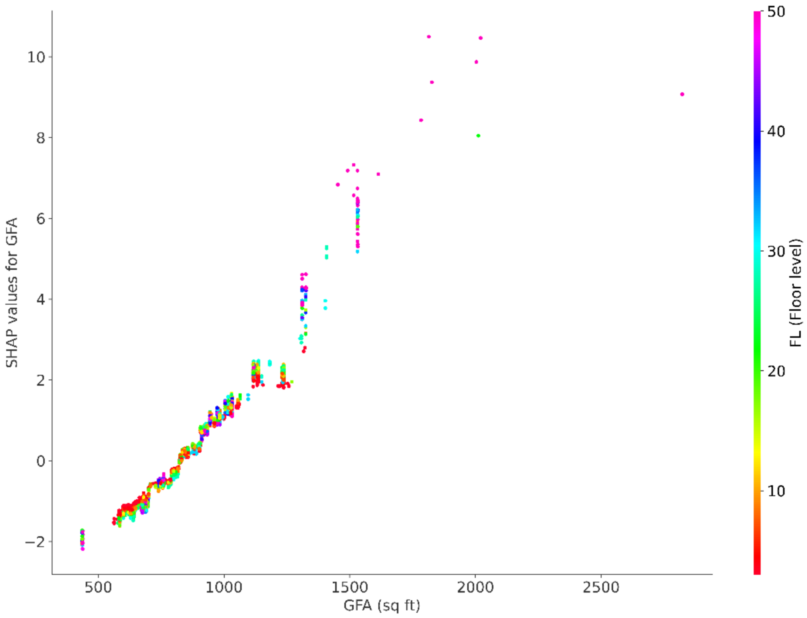

Figure 9. SHAP values for GFA (Test set)

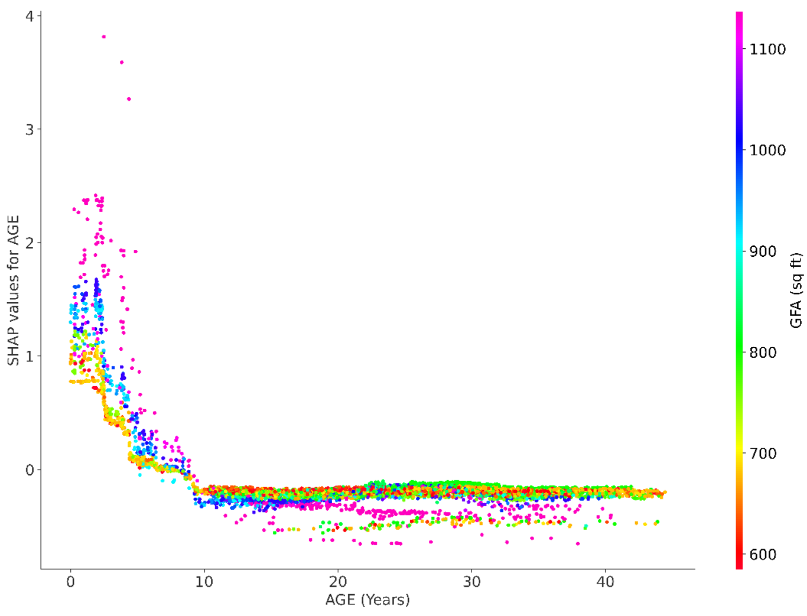

Figure 10. SHAP values for AGE (Test set)

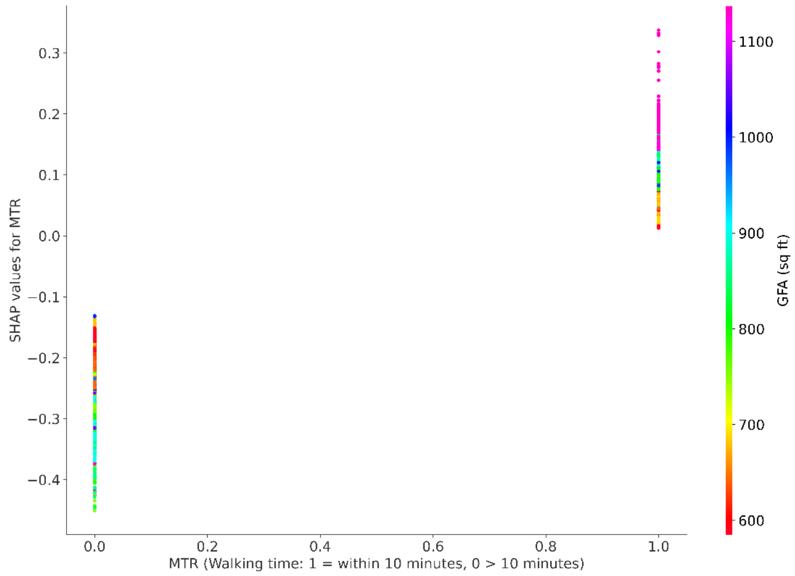

Figure 11. SHAP values for MTR (Test set)

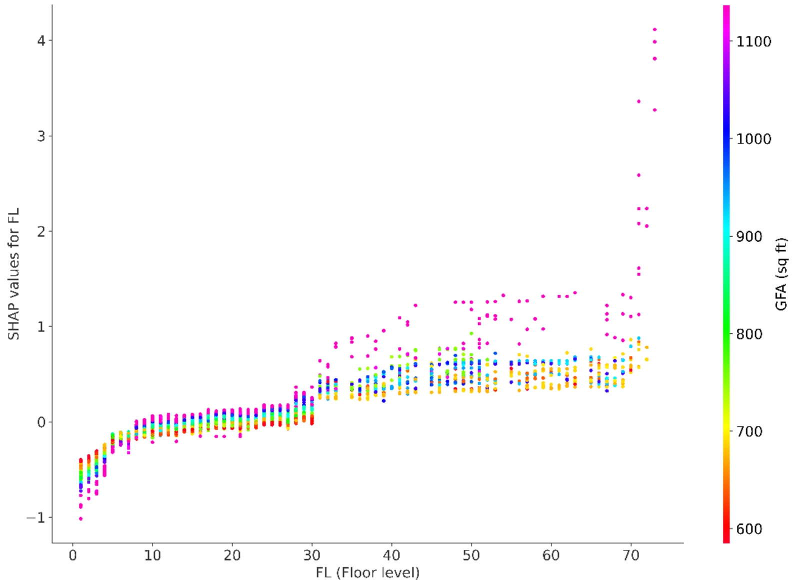

Figure 12. SHAP values for FL (Test set)

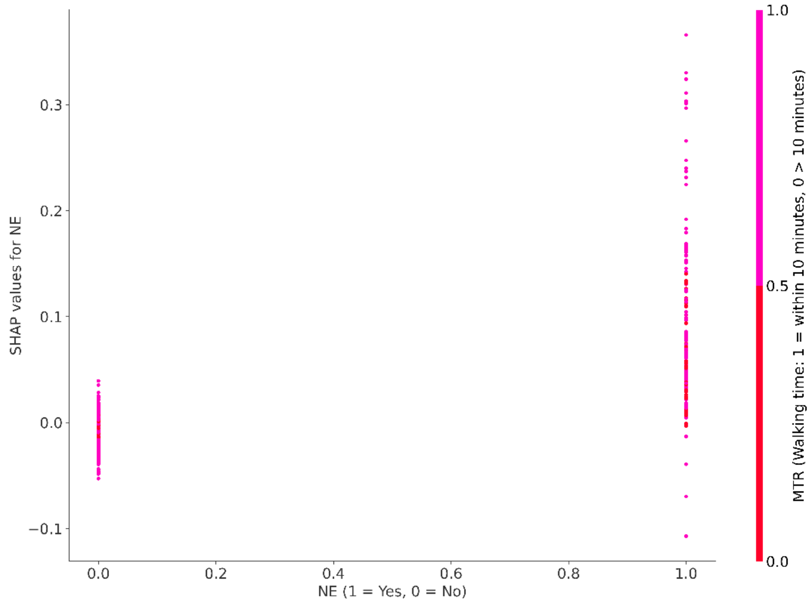

Figure 13. SHAP values for NE (Test set)

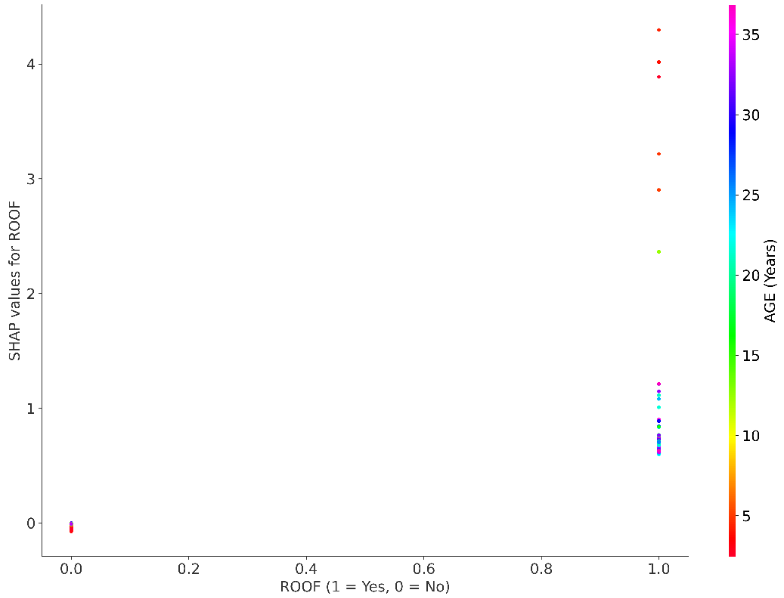

Figure 14. SHAP values for ROOF (Test set)

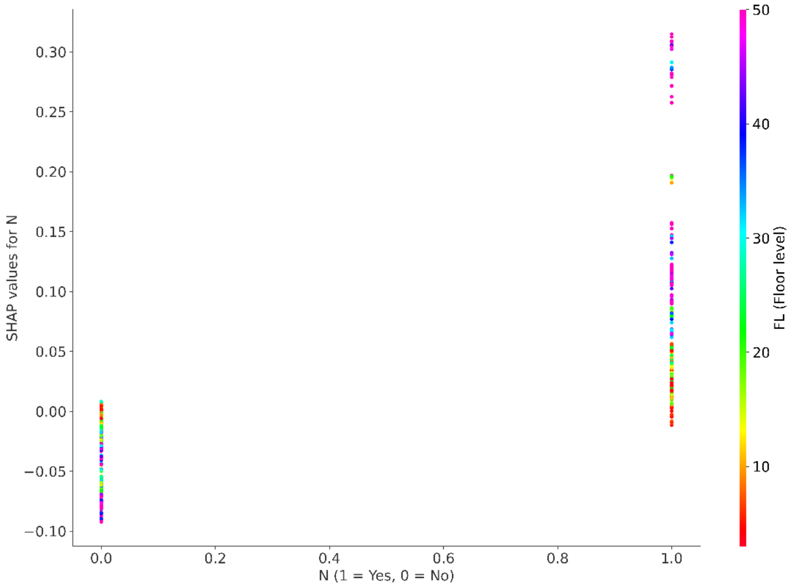

Figure 15. SHAP values for N (Test set)

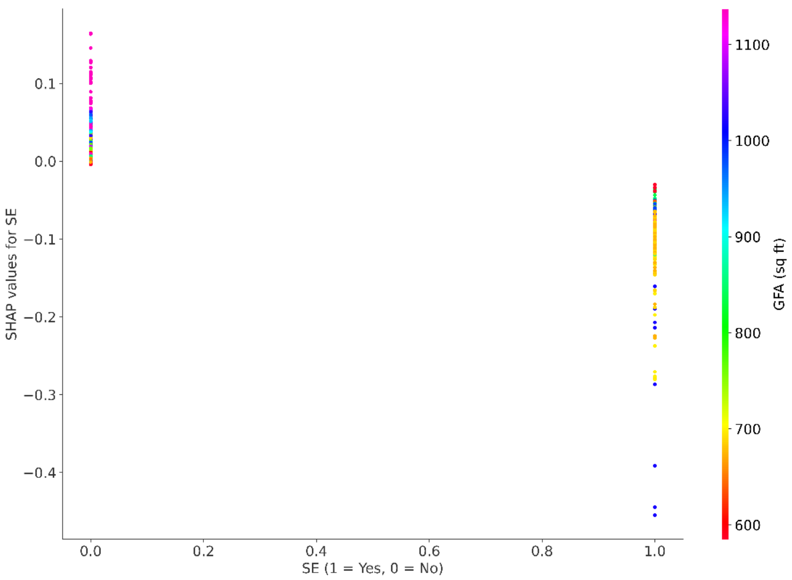

Figure 16. SHAP values for SE (Test set)

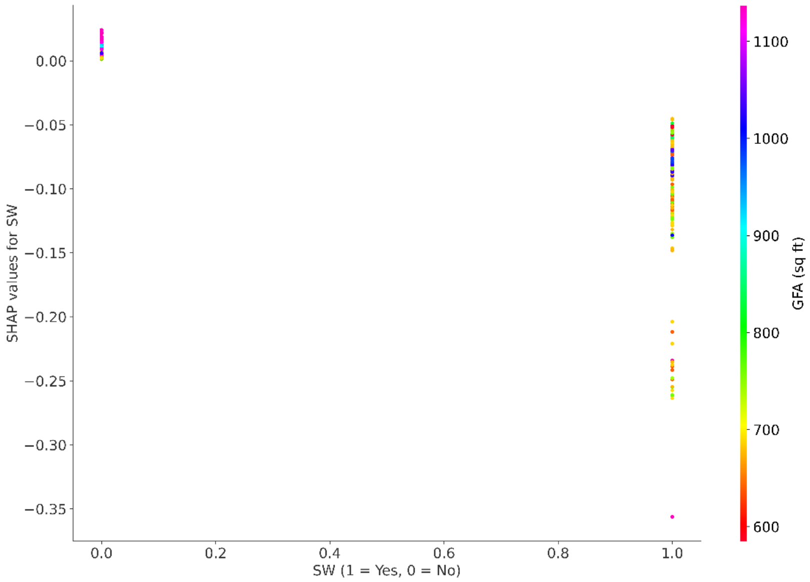

Figure 17. SHAP values for SW (Test set)

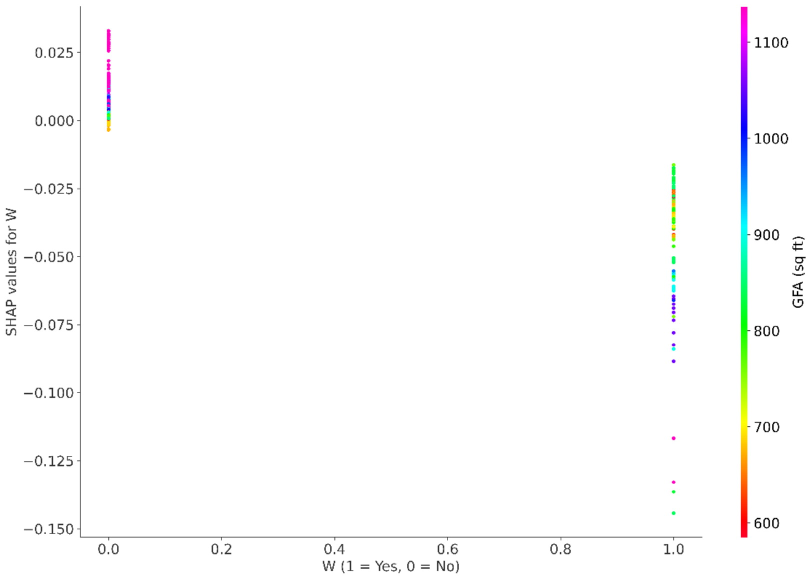

Figure 18. SHAP values for W (Test set)

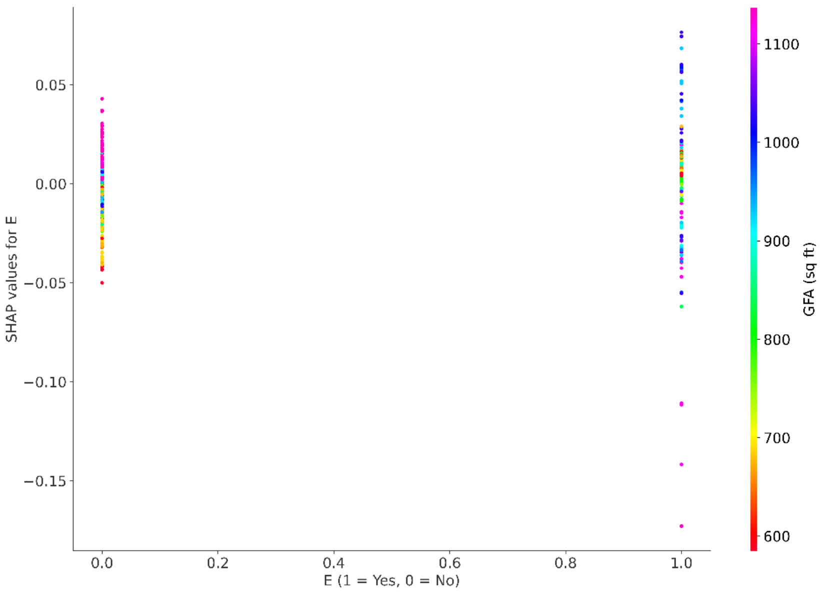

Figure 19. SHAP values for E (Test set)

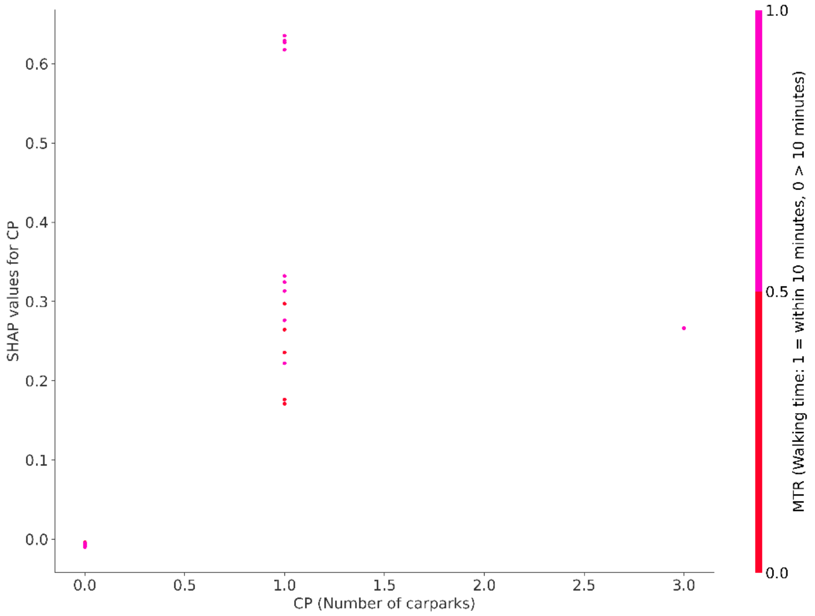

Figure 20. SHAP values for CP (Test set)

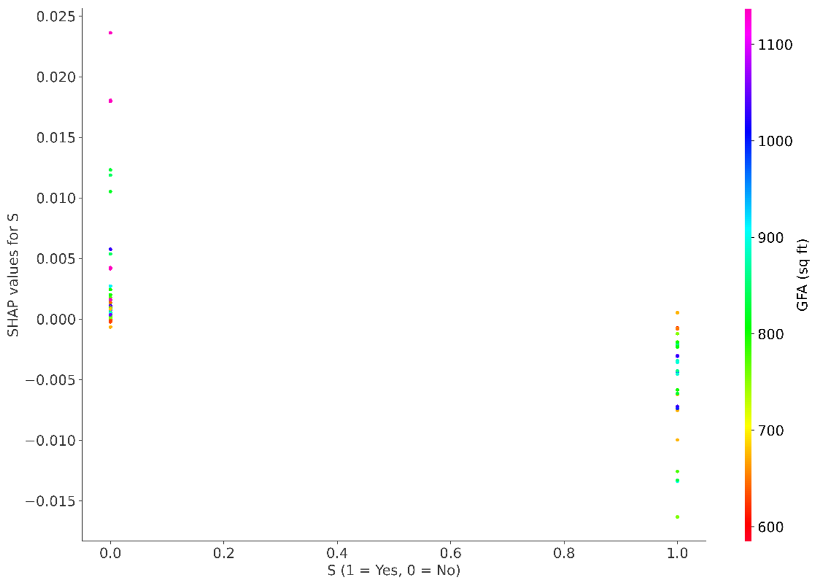

Figure 21. SHAP values for S (Test set)

7. Conclusions

This study employs XGBoost and SHAP to analyze the dynamics of Hong Kong's residential property market, enhancing transparency in identifying key determinants of property prices. The findings indicate that square footage is the most influential feature, while proximity to mass transit stations highlights contextual nuances. The interpretability provided by SHAP enables stakeholders to quantitatively assess the impact of individual features, facilitating urban planners and policymakers in detecting trends, prioritizing variables, and addressing market drivers. For instance, insights into transit accessibility can inform transportation planning and zoning policies, thereby improving connectivity and promoting equitable urban development.

Unlike existing literature (Lenaers & De Moor [7], Trindade Neves et al. [8], Jin et al. [9], Khan et al. [30]), this study employs traditional evaluation metrics such as R^2, MAE, MSE, MAPE, and RMSE for both the training and test sets, offering a comprehensive assessment of the model's performance and its susceptibility to overfitting or underfitting. In our study, these metrics demonstrate the absence of overfitting or underfitting by ensuring consistency between training and test set performance. When comparable values for these metrics are observed across both sets, it suggests that the model generalizes well rather than memorizing training data.

By presenting both traditional evaluation metrics for training and test sets and the SHAP values for the test set, this research provides a thorough assessment of the model's predictive performance and the contribution of each feature to its predictions. This dual approach enhances the robustness of the XGBoost algorithm, and the reliability of the insights derived from the SHAP analysis. Compared to existing literature, this study contributes significantly by ensuring a thorough evaluation of the model in terms of both predictive accuracy and interpretability on unseen data.

Visualization tools such as SHAP summary plots and beeswarm plots further aid in translating complex analytical outputs into accessible insights for non-technical stakeholders. By clearly illustrating the influence of features such as gross floor area, property age, and transit accessibility on property values, these tools facilitate data-driven decision-making. Beyond real estate, the integration of SHAP into urban analytics has the potential to identify disparities in housing affordability and inform interventions targeting underserved communities.

While this study focuses on a limited number of housing estates, its findings hold relevance for broader urban contexts, both locally and internationally. By leveraging XAI, this research highlights its transformative potential in bridging data science and policymaking. XAI enhances the transparency and interpretability of complex models, democratizing access to advanced analytical methodologies. This empowers policymakers, stakeholders, and communities to make informed, evidence-based decisions. Ultimately, the study underscores the pivotal role of XAI in fostering collaborative, inclusive, and sustainable urban governance.

8. Declarations

8.1. Author Contributions

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

8.2. Data Availability Statement

The data presented in this study are available in the article.

8.3. Funding

This research was supported by the Hong Kong Polytechnic University through the Start-up Funding (P0049523).

8.4. Conflicts of Interest

The authors declare no conflict of interest.

9. References

[1] Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 785–794. doi:10.1145/2939672.2939785.

[2] Shapley, L. S. (2020). A Value for n-PERSON Games. Classics in Game Theory, 69–79, Princeton University Press, Princeton, United States. doi:10.2307/j.ctv173f1fh.12.

[3] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K. W., Newman, S. F., Kim, J., & Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760. doi:10.1038/s41551-018-0304-0.

[4] Kumkar, P., Madan, I., Kale, A., Khanvilkar, O., & Khan, A. (2018). Comparison of Ensemble Methods for Real Estate Appraisal. 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 297–300. doi:10.1109/icict43934.2018.9034449.

[5] Stang, M., Krämer, B., Nagl, C., & Schäfers, W. (2023). From human business to machine learning-methods for automating real estate appraisals and their practical implications. Journal of Real Estate Economics, 9(2), 81–108. doi:10.1365/s41056-022-00063-1.

[6] Calainho, F. D., van de Minne, A. M., & Francke, M. K. (2024). A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate. Journal of Real Estate Finance and Economics, 68(4), 624–653. doi:10.1007/s11146-022-09893-1.

[7] Lenaers, I., & De Moor, L. (2023). Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study. Finance Research Letters, 58, 104306. doi:10.1016/j.frl.2023.104306.

[8] Trindade Neves, F., Aparicio, M., & de Castro Neto, M. (2024). The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities. Applied Sciences (Switzerland), 14(5), 2209. doi:10.3390/app14052209.

[9] Jin, S., Zheng, H., Marantz, N., & Roy, A. (2024). Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI. Applied Geography, 169, 103339. doi:10.1016/j.apgeog.2024.103339.

[10] Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. Journal of Portfolio Management, 43(6), 202–211. doi:10.3905/jpm.2017.43.6.202.

[11] Yang, B., & Cao, B. (2018). Research on Ensemble Learning-based Housing Price Prediction Model. Big Geospatial Data and Data Science, 1(1), 1–8. doi:10.23977/bgdds.2018.11001.

[12] Hendrayati, H., Achyarsyah, M., Marimon, F., Hartono, U., & Putit, L. (2024). The Impact of Artificial Intelligence on Digital Marketing: Leveraging Potential in a Competitive Business Landscape. Emerging Science Journal, 8(6), 2343–2359. doi:10.28991/ESJ-2024-08-06-012.

[13] Burnwal, Y., & Jaiswal, D. R. C. (2023). A Comprehensive Survey on Prediction Models and the Impact of XGBoost. International Journal for Research in Applied Science and Engineering Technology, 11(12), 1552–1556. doi:10.22214/ijraset.2023.57625.

[14] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. doi:10.1007/bf00058655.

[15] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324.

[16] Banerjee, D., & Dutta, S. (2018). Predicting the housing price direction using machine learning techniques. IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017, 2998–3000. doi:10.1109/ICPCSI.2017.8392275.

[17] Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. Land, 12(4), 740. doi:10.3390/land12040740.

[18] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. doi:10.1214/aos/1013203451.

[19] Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70. doi:10.1080/09599916.2020.1832558.

[20] DataScientest (2024). XGBoost: The Champion of Competitive Machine Learning. DataScientest, Paris, France. Available online: https://datascientest.com/en/xgboost-the-champion-of-competitive-machine-learning (accessed on April 2025).

[21] Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. Journal of Sensors, 2022. doi:10.1155/2022/4649510.

[22] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks, 18(6), 1–10. doi:10.1177/15501329221106935.

[23] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. doi:10.1145/3292500.3330701.

[24] Kee, T., & Ho, W. K. O. (2024). Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods. Preprints, 1–17. doi:10.20944/preprints202406.0264.v2.

[25] Kee, T., & Ho, W. (2024). Predicting Industrial Property Prices with Explainable Artificial Intelligence. Preprints, 1–22. doi:10.20944/preprints202409.0875.v1.

[26] SHAP (2018). Welcome to the SHAP documentation. SHAP, Osaka, Japan. Available online: https://shap.readthedocs.io/en/latest/ (accessed on April 2025).

[27] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December, 2017, Long Beach, United States.

[28] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. doi:10.1038/s42256-019-0138-9.

[29] RVD (2024). Hong Kong Property Review. Hong Kong Special Administrative Region Government, Central Government Complex, Hong Kong. Available online: https://www.rvd.gov.hk/en/publications/hkpr_previous.html (accessed on April 2025).

[30] Khan, A., Debnath, P., Sayeed, A. A., Sumon, F. I., Rahman, A., Khan, T., & Pant, L. (2024). Explainable AI and Machine Learning Model for California House Price Predictions: Intelligent Model for Homebuyers and Policymakers. Journal of Business and Management Studies, 6(5), 73–84. doi:10.32996/jbms.2024.6.5.9.