

## Article

# Research on the Nonlinear Relationship Between Carbon Emissions from Residential Land and the Built Environment: A Case Study of Susong County, Anhui Province Using the XGBoost-SHAP Model

Congguang Xu <sup>1,2,3</sup>, Wei Xiong <sup>1,2,\*</sup>, Simin Zhang <sup>3</sup>, Hailiang Shi <sup>2</sup> , Shichao Wu <sup>2</sup>, Shanju Bao <sup>4</sup> and Tieqiao Xiao <sup>5</sup>

<sup>1</sup> Science Island Branch, University of Science and Technology of China, Hefei 230026, China; 18056010101@163.com

<sup>2</sup> Anhui Institute of Optics and Fine Mechanics, Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; hlshi@aiofm.ac.cn (H.S.); wusc@aiofm.ac.cn (S.W.)

<sup>3</sup> Anhui Provincial Key Laboratory of Building Earthquake Disaster Mitigation and Green Operations, Anhui Institute of Building Research & Design, Hefei 230088, China; 18756528624@163.com

<sup>4</sup> School of Geography and Tourism, Anhui Normal University, Wuhu 241000, China; 2014222@ahnu.edu.cn

<sup>5</sup> School of Architecture and Planning, Anhui Jianzhu University, Hefei 230009, China; xtg020032@ahjzu.edu.cn

\* Correspondence: frank@aiofm.ac.cn; Tel.: +86-0551-65593098

**Abstract:** Residential land is the basic unit of urban-scale carbon emissions (CEs). Quantifying and predicting CEs from residential land are conducive to achieving urban carbon neutrality. This study took 84 residential communities in Susong County, Anhui Province as its research object, exploring the nonlinear relationship between the urban built environment and CEs from residential land. By identifying CEs from residential land through building electricity consumption, 14 built environment indicators, including land area (LA), floor area ratio (FAR), greening ratio (GA), building density (BD), gross floor area (GFA), land use mix rate (Phh), and permanent population density (PPD), were selected to establish an interpretable machine learning (ML) model based on the XGBoost-SHAP attribution analysis framework. The research results show that, first, the goodness of fit of the XGBoost model reached 91.9%, and its prediction accuracy was better than that of gradient boosting decision tree (GBDT), random forest (RF), the Adaboost model, and the traditional logistic model. Second, compared with other ML models, the XGBoost-SHAP model explained the influencing factors of CEs from residential land more clearly. The SHAP attribution analysis results indicate that BD, FAR, and Phh were the most important factors affecting CEs. Third, there was a significant nonlinear relationship and threshold effect between built environment characteristic variables and CEs from residential land. Fourth, there was an interaction between different dimensions of environmental factors, and BD, FAR, and Phh played a dominant role in the interaction. Reducing FAR is considered to be an effective CE reduction strategy. This research provides practical suggestions for urban planners on reducing CEs from residential land, which has important policy implications and practical significance.

**Keywords:** residential land; machine learning; XGBoost model; SHAP algorithm; carbon emissions; built environment factors; nonlinear relationship



Academic Editor: Chuanrong Zhang

Received: 10 January 2025

Revised: 12 February 2025

Accepted: 18 February 2025

Published: 20 February 2025

**Citation:** Xu, C.; Xiong, W.; Zhang, S.; Shi, H.; Wu, S.; Bao, S.; Xiao, T.

Research on the Nonlinear Relationship Between Carbon Emissions from Residential Land and the Built Environment: A Case Study of Susong County, Anhui Province Using the XGBoost-SHAP Model. *Land* **2025**, *14*, 440. <https://doi.org/10.3390/land14030440>

**Copyright:** © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the Creative Commons

Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cities, as hubs of human economic activity and resource consumption, are the primary regions responsible for global warming and climate change. In 2023, carbon emissions (CEs)

from urban areas accounted for over 80% of the world's total, making cities the primary source of CEs [1]. The household is the fundamental unit of urban life. Residential land serves as the fundamental unit of urban land use. According to relevant statistical data, residential energy consumption in China accounts for 30–40% of the energy consumption within the urban construction sector, and CEs from buildings on residential land occupy a significant proportion [2]. In addition, low-carbon residential areas serve as crucial spatial carriers and fundamental units for building low-carbon cities. Therefore, investigating the mechanisms by which the spatial structure and other built environments of residential areas influence residents' CEs holds immense importance for urban low-carbon construction and is a vital step toward achieving carbon neutrality goals in China and beyond [3].

Scholars have engaged in heated discussions surrounding the accounting and prediction methods for CEs from different types of land use. Based on different principles and methods, CE prediction for different types of land use can be divided into qualitative prediction and quantitative prediction. The former relies on the observer's ability to analyze and make predictions based on experience, while the latter involves constructing mathematical models using historical data and mathematical methods. Currently, research on CE prediction from urban construction land at the macro scale mainly focuses on quantitative analysis. Common data sources include remote sensing data, land use data, electricity consumption, and other statistical data. Analysis methods mainly include time series analysis, regression analysis, and principal component analysis. Common prediction models include the DPSIR model [4], the TOPSIS model [4], the LMDI model [5], the STIRPAT model [6–9], the LEAP model [10], and the EKC model [11,12], which have played an important role in predicting CEs from buildings, planning pathways for carbon neutrality, and evaluating environmental policies. For example, analytical methods such as grey correlation analysis [13], multiple linear regression analysis [14], Pearson analysis [15], one-way ANOVA [16], and the chi-square test [17] are often used to analyze the effects and interactions between spatial and non-spatial factors and CEs from construction land in a comprehensive way. However, these methods have certain limitations in dealing with complex nonlinear relationships and multi-dimensional feature data. On the other hand, research on CE prediction from residential land at the micro scale is relatively scarce. Some domestic scholars, such as Jiang Haiyan [18], Wang Weiqiang [19], Wu Wei [20], and Zhang Li chunyi [21], have used multiple linear regression models to explore the impact mechanism of residential land CEs from the perspective of built environment factors, such as floor area ratio (FAR), greening rate (GR), building density (BD), and gross floor area (GFA). The research findings confirm that various built environment factors significantly impact CEs from residential land. A favorable built environment can notably enhance the microclimate, energy demand, and energy intensity of residential land, thereby influencing its CEs. This provides a valuable reference and insight.

These studies reveal the impact of the built environment on CEs to a certain extent. However, due to the limited data sources and high variable complexity, traditional regression models have limitations in prediction accuracy and applicability.

In recent years, the advantages of machine learning (ML) in handling multi-dimensional data have been continuously verified, and the modeling results are often more scientific and accurate. It can compensate for the deficiencies of mathematical models and, therefore, is widely used by scholars in carbon emission prediction models in different fields. Commonly used ML models include support vector machine (SVM) [22–25], random forest (RF) [26,27], extreme gradient boosting (XGBoost) [28–35], gradient boosting decision tree (GBDT) [36], and back propagation (BP) [37,38]. Among them, in the XGBoost model, as an improved algorithm of GBDT, regularization rules have been incorporated to mitigate the risk of overfitting, which significantly reduces model prediction errors

by optimizing the algorithm sequence and integrating decision trees [39]. The XGBoost model excels in handling nonlinear relationships and complex interactive effects in spatially heterogeneous data, offering advantages such as a fast computation speed, high accuracy, and strong robustness. It has been widely applied in various fields such as transportation CE prediction [40], automobile CE prediction [41], user behavior prediction [42], and heart disease risk prediction [43,44]. However, the “black box” nature of the XGBoost model makes its solution process opaque, making it difficult to explain the importance of various influencing factors. To address this issue, Shapley Additive exPlanation (SHAP) values, as an interpretability tool for ML models, have been widely used in feature importance analysis in recent years. By fairly allocating variable contributions, SHAP values can effectively identify key features and explore relationships between variables, offering excellent visualization effects. Scholars such as Santos R M [45], Alireza F [46], Shi W [47], Li Yangtao [41], Wang Zhijiang [48], and Yang Dongfeng [49] have successfully introduced SHAP values into fields such as transportation CE prediction and user behavior analysis for quantitative analysis of the contribution of features to prediction results. For instance, Li Yangtao et al. uncovered the critical factors influencing vehicle CEs by utilizing SHAP values in conjunction with the XGBoost model [41]. Wang Zhiqiang et al. employed SHAP values to scrutinize the individual and interactive effects of multiple features on CEs during the building design phase [48]. Yang Dongfeng et al. utilized SHAP values to investigate the nonlinear correlation between the built environment and street vitality as well as the interactive effects among various built environment elements [42]. SHAP values possess advantages, such as the strong global and local interpretability of variables, the fair allocation of variable contributions, and excellent visualization effects, compensating for the poor interpretability of tree-based models such as decision trees and RF. However, there has been little research on using the XGBoost model combined with SHAP values to carry out CE prediction for residential land.

This study aimed to conduct research on the nonlinear relationship between CEs from residential land use and the built environment based on an ML multi-model comparison in an attempt to address the limitations of existing research. Firstly, given that most research is conducted macroscopically and lacks an in-depth exploration of CE prediction and impact mechanisms, especially at the residential community level, this study innovatively focuses on the prediction of CEs from residential land, combining built environment characteristics such as FAR and BD to explore their nonlinear relationship. Secondly, in response to the lack of data dimensions, this study integrates multiple sources of data, including remote sensing data, land use data, and electricity consumption data, to construct a comprehensive, high-dimensional feature set that aims to improve the model prediction accuracy. In terms of model selection, this study uses the XGBoost model as the core, optimizing the prediction performance through its advantages in handling nonlinear relationships and complex interactive effects. It also utilizes SHAP values to conduct an explanatory analysis on the model prediction results, quantifying the contribution weights of each feature to CEs, revealing key influencing factors and their mechanisms of action, and providing a CE prediction method for residential community planning and design. Furthermore, by employing feature engineering methods such as recursive feature elimination, the feature selection process was optimized, improving the prediction accuracy and reliability of the model. Establishing a CE prediction model for residential land use based on the XGBoost model demonstrates a certain level of reliability and innovation. This study transforms the prediction of CEs from residential land from a subject matter of other disciplines into a problem that urban planning can explore. It not only achieves breakthroughs in the accuracy of CE prediction for residential land but also provides a scientific basis for the construction of low-carbon cities. In the context of global climate change in particular, it

holds significant theoretical and practical implications for achieving carbon neutrality goals and low-carbon city practices.

## 2. Materials and Methods

### 2.1. Study Area

This study selected the central urban area of Susong County of Anhui Province as the research area. The county, located in the southwest of Anhui Province at the junction of the Anhui, Hubei, and Jiangxi provinces and having a history of about 1000 years, possesses a typical climate characteristic of hot summers and cold winters. By combining indicators such as the construction time of residential land, the number of households (NOH), the size of the residential land area (LA), and the FAR, 84 typical residential land samples within the central urban area of the county were selected. To make the survey samples more representative, the selected residential land was of various types, including old residential areas, affordable housing, resettlement housing, and large residential communities, which have diverse built environmental characteristics. This provides a certain level of representativeness for studying the relationship between residential land CEs and construction environmental factors as well as predicting residential land CEs. The distribution of residential land samples in the study area is shown in Figure 1.

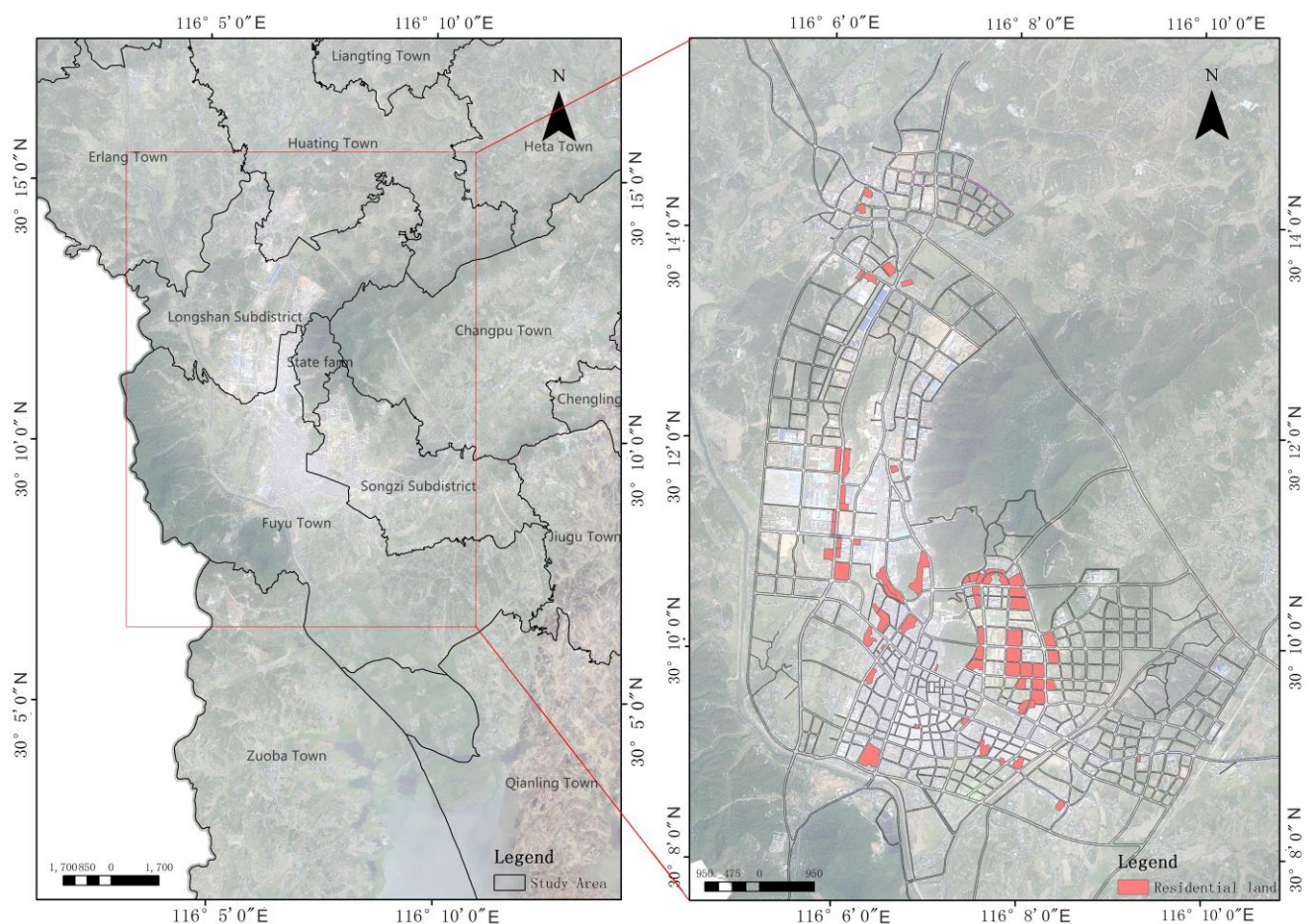


**Figure 1.** Schematic diagram of the residential land sample.

### 2.2. Data Sources

Data on electricity consumption by buildings and built environment characteristics across various residential land uses provide crucial support for the prediction of CEs from residential land. Through data collection and field research, this study gathered information

on electricity consumption by buildings and built environment characteristics from sample residential land plots in the central urban area of Susong County. Specifically, this included electricity consumption data for buildings within the sample residential land plots in the county from 2021 to 2023 as well as built environment characteristic indicators such as LA, GFA, GR, FAR, and BD. Data on the permanent population density (PPD) and workforce density (WFD) were obtained through discussions and visits to relevant departments. The statistical data on explanatory variables and target characteristics of residential land in the study area are shown in Table 1.

**Table 1.** Statistical data on explanatory variables and target characteristics.

| Variables                               | Unit                               | Mean      | Std Dev   | Min     | Max        |
|-----------------------------------------|------------------------------------|-----------|-----------|---------|------------|
| <b>Independent Variables</b>            |                                    |           |           |         |            |
| Land Area (LA)                          | m <sup>2</sup>                     | 33,490.37 | 27,366.59 | 631.67  | 176,357.74 |
| Number of Households (NOH)              | —                                  | 480.00    | 345.00    | 6.00    | 2080.00    |
| Gross Floor Area (GFA)                  | m <sup>2</sup>                     | 71,807.07 | 59,913.52 | 1084.64 | 438,641.01 |
| Floor Area Ratio (FAR)                  | —                                  | 1.93      | 0.62      | 1.10    | 4.16       |
| Greening Rate (GR)                      | %                                  | 0.32      | 0.04      | 0.20    | 0.46       |
| Building Density (BD)                   | %                                  | 27.86     | 5.04      | 16.21   | 39.45      |
| Permanent Population Density (PPD)      | —                                  | 476.40    | 188.24    | 180.00  | 860.00     |
| Workforce Density (WFD)                 | —                                  | 1.70      | 1.11      | 0.11    | 4.92       |
| Land Use Mix Rate (Phh)                 | —                                  | 0.76      | 0.12      | 0.45    | 1.00       |
| Road Network Density (RND)              | km/km <sup>2</sup>                 | 4.10      | 0.79      | 2.15    | 6.15       |
| Intersection Density (ITD)              | —                                  | 2.00      | 1.66      | 0.23    | 8.12       |
| Public Transport Network Density (PTND) | —                                  | 1.47      | 0.38      | 1.05    | 2.60       |
| Point of Interest (POI)                 | —                                  | 3.88      | 2.40      | 0.15    | 9.54       |
| Parking Area Ratio (PAD)                | —                                  | 0.99      | 0.40      | 0.15    | 2.30       |
| <b>Dependent Variables</b>              |                                    |           |           |         |            |
| Carbon Emissions (CEs)                  | 10 <sup>3</sup> kgCO <sub>2e</sub> | 3282.91   | 3016.23   | 123.15  | 15,086.93  |
| CEs per unit building area              | 10 <sup>3</sup> kgCO <sub>2e</sub> | 0.08      | 0.10      | 0.01    | 0.47       |

Due to the different residential land areas, to mitigate the impact of land area size on CEs and address the variable scale differences caused by data with different dimensions, data normalization was applied to ensure the reasonableness of the sample size in the feature space. Normalization aims to reduce the data scale, minimize data prediction errors, and accelerate algorithm convergence.

### 2.3. Research Methods

#### 2.3.1. Analytical Framework

This study analyzed the complex nonlinear relationship between the built environment of residential land and CEs, progressing from data collection and processing to the construction and comparison of various ML models, and, finally, exploring the SHAP impact mechanism (Figure 2). This included collecting data and calculating indicators for the independent variables of the model, specifically the built environment of residential land, and the dependent variable of the CEs. This also included utilizing the Python 3.13.2 programming language to establish an interpretable ML model based on the XGBoost-SHAP attribution analysis framework. This was achieved through comparative analysis between the XGBoost model and other multi-model prediction outcomes, assessing the significance of feature variables on the model's prediction results. The SHAP method was then employed to interpret and visualize the model's results, thereby exploring the nonlinear relationship between the built environment and residential land CEs as well as the interactive effects among various built environment elements.

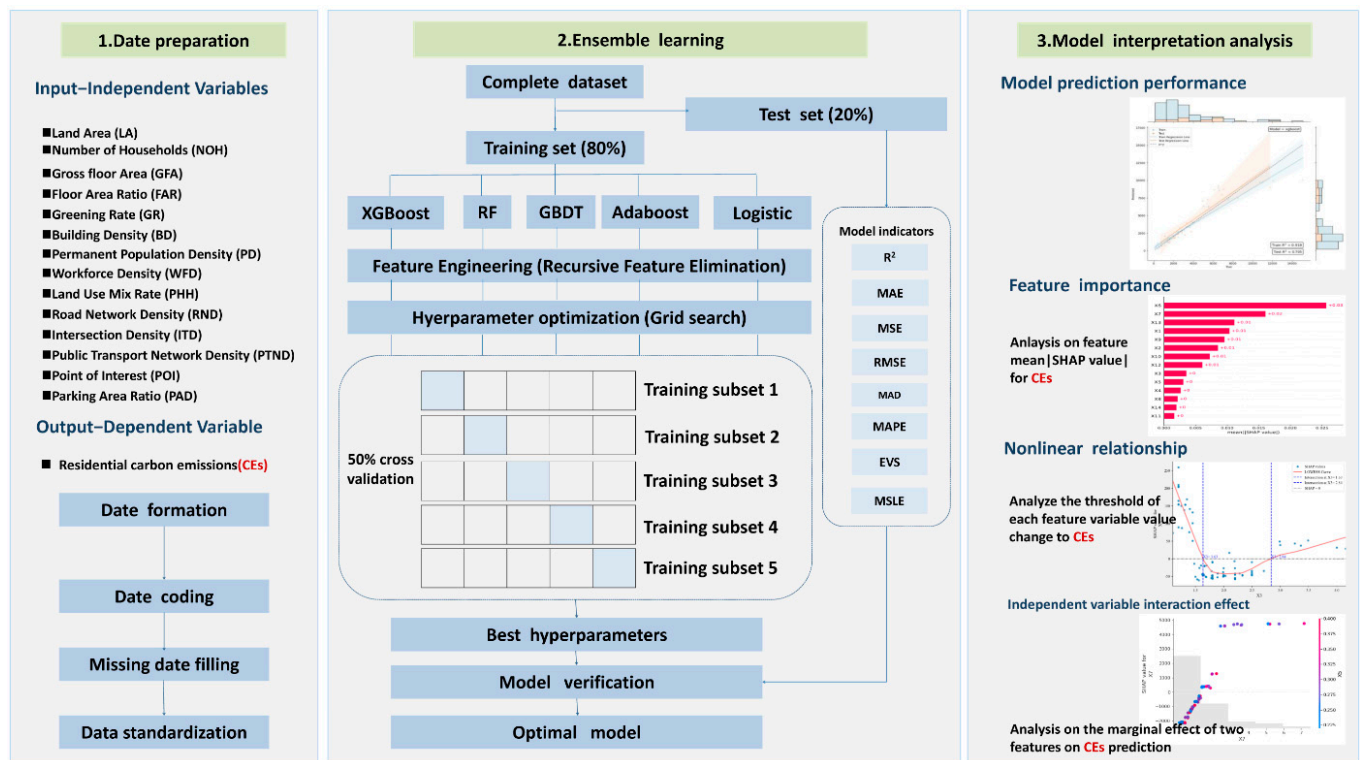


**Figure 2.** Schematic showing the workflow of this study.

### 2.3.2. Construction of Model Feature Set

#### (1) Dependent Variables

The dependent variable of the residential land carbon emission prediction model in this study was the carbon emission amount from residential land. According to the carbon emission coefficient method, the carbon emission amount from residential land is calculated by multiplying the energy consumption data by the carbon emission coefficient per unit of energy consumption. When calculating CEs from household travel and household living, the carbon emission coefficient values refer to the Carbon Emission Intensity Table for Public Transportation in China published by the domestic profit organization Green Source Group and the Citizen Energy Conservation and Emission Reduction Handbook from the Ministry of Science and Technology of China, which indicates that the carbon emission coefficient for electricity per unit of energy consumption is 0.98.

$$C = \sum S_i \times e_i \times F$$

where C is the building factor, S is the building area, e is the unit building energy consumption, and F is the unit energy consumption carbon emission coefficient.

#### (2) Independent Variables

At the level of the built environment, scholars such as Ewings have depicted the built environment using the “5Ds” model [49], which includes density, diversity, design, destination accessibility, and distance to transit. Chinese scholars generally believe that factors such as LA, GFA, FAR, PPD, and BD are closely related to land CEs. The study’s selection of independent variable indicators for residential land primarily considered the controllability, quantifiability, and accessibility of the indicators from the perspective of spatial planning. Ultimately, the most relevant factors affecting CEs from residential land were selected as independent variables, including LA, GFA, FAR, GR, and BD. The

definitions and measurement methods of the independent variables are summarized in Table 2.

**Table 2.** Summary of definitions of independent variables and measurement methods.

| Variable Grouping   | Variable                                | Measuring Method                                                                                                                           | Data Sources                                    |
|---------------------|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------|
| Basic Attributes    | Land Area (LA)                          | /                                                                                                                                          | Planning data                                   |
|                     | Number of Households (NOH)              | /                                                                                                                                          | Planning data                                   |
|                     | Gross Floor Area (GFA)                  | /                                                                                                                                          | Planning data                                   |
| Density             | Floor Area Ratio (FAR)                  | Total floor area above ground/Research scope area                                                                                          | Planning data                                   |
|                     | Greening Rate (GR)                      | Green land area/Research scope area                                                                                                        | Planning data                                   |
|                     | Building Density (BD)                   | Total area of building foundation/Total area of construction land                                                                          | Planning data                                   |
|                     | Permanent Population Density (PPD)      | Population within the research scope/Research scope area                                                                                   | Statistical yearbook, departmental research     |
|                     | Workforce Density (WFD)                 | Number of positions within the research scope/Research scope area                                                                          |                                                 |
| Diversity           | Land Use Mix Rate (Phh)                 | The logarithmic ratio of the various types of land areas to areas of regions is from 0 to 1 (the closer to 1, the lower the mixing degree) | Survey data on land use change in Susong County |
| Design              | Road Network Density (RND)              | Road length/Research scope area                                                                                                            | Internet map                                    |
|                     | Intersection Density (ITD)              | Number of intersections/Research scope area                                                                                                |                                                 |
|                     | Parking Area Ratio (PAD)                | Building parking spaces in residential areas/Total number of households in residential areas                                               | Planning data                                   |
| Distance to transit | Public Transport Network Density (PTND) | Total length of bus routes/Research scope area                                                                                             |                                                 |
| Destination         | Point of Interest (POI)                 | POI/Research scope area                                                                                                                    | Internet map                                    |

### 2.3.3. Machine Learning Model: The XGBoost Model

The XGBoost model was first proposed by Chen T Q [50] as an ensemble-tree-based model. It iteratively generates numerous decision trees and then combines the results of all the trees to create the final model. Each individual decision tree is built on the basis of minimizing a specified loss function, and, each time a new decision tree is created, cases that were misestimated by the previous tree are given higher weights. This method can be used to construct both regression and classification models, significantly improving the computational efficiency of GBDT through distributed computing. Generally speaking, the XGBoost model is more than ten times faster than other GBDT algorithm frameworks. The XGBoost model is built using multiple weak learners, with CART decision trees as the base learners, employing a gradient boosting algorithm. It introduces Lasso regularization (L1) and Ridge regularization to control the complexity of the model, preventing overfitting and

enhancing the model's generalizability. Additionally, the XGBoost model supports various types of base classifiers, calculates the negative gradient of the model, and iteratively trains multiple times, continuously adding trees during the training process to fit the residuals of the previous predictions and, thereby, gradually improve the model's accuracy.

The XGBoost model randomly selects 80% of the dataset samples as the training set for modeling and uses the remaining 20% of the dataset samples as the validation set to validate the constructed model (Figure 3). The CE prediction function expression of the XGBoost model is:

$$\hat{y}_i^{(t)} = f_p^t(x_i^{\text{set}})$$

where  $\hat{y}_i^{(t)}$  is the predicted CE value for the  $t$ -th iteration;  $x_i^{\text{set}}$  is the feature variable for CEs; and  $f_p^t(x_i^{\text{set}})$  is the prediction function regarding  $x_i^{\text{set}}$ .

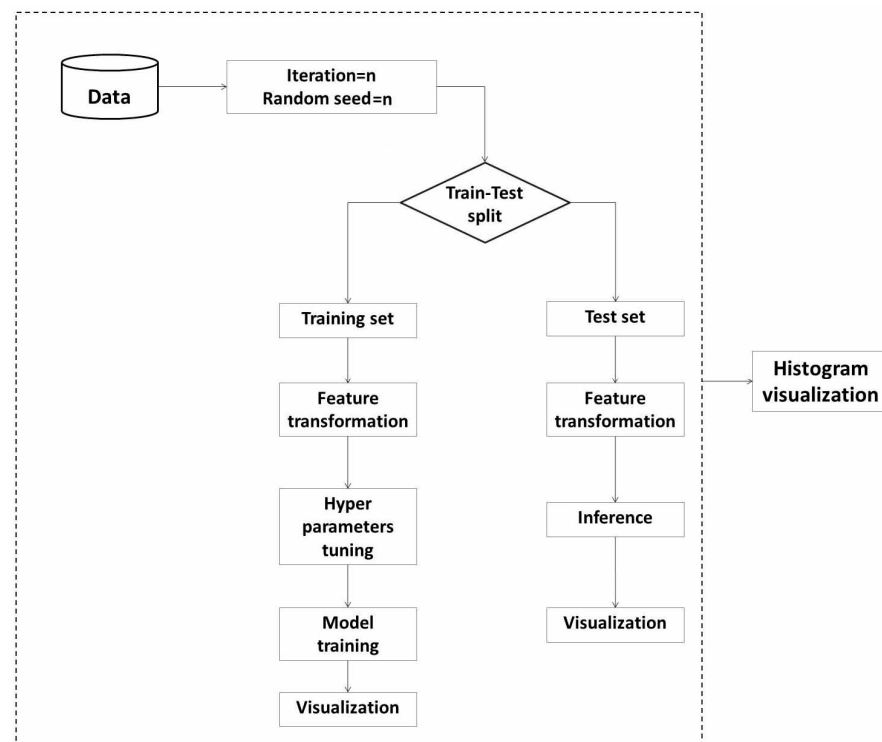


**Figure 3.** Calculation process of the XGBoost model.

This study used the “hyper-parameter grid search” method to adjust key parameters, including the learning rate (which controls the learning speed and has a significant impact on overfitting), max-depth (which determines the complexity of each tree and thus affects overfitting), and lambda and alpha (which prevent overfitting). Specifically, it involves training all listed parameter combinations in a loop until the optimal parameter combination is found. In detail, the possible values for the learning rate are 0.01, 0.1, and 0.3, the max-depth can take values of 4, 5, 6, 7, 8, 9, and 10, the min-child-weight (which determines the minimum sample weight of leaf nodes and is used to prevent overfitting) can take values of 1, 2, 3, and 4, and the lambda and alpha can take values of 0.05, 0.1, 0.2, and 1. By traversing all parameter combinations, the optimal parameters that achieve the best model performance were found to be a learning rate of 0.01, a maximum tree depth of 10, a min-child-weight of 4, a lambda of 0.05, and an alpha of 0.2. As an ensemble tree-based model, XGBoost first iteratively generates numerous decision trees and then combines the results of all the trees to create the final model. Each individual decision tree is built on the basis of minimizing a specified loss function and, each time a new decision tree is constructed, cases that were misestimated by the previous tree are given higher weights.

The XGBoost model is built by combining multiple weak learners, using decision trees as base learners, and employing the gradient boosting algorithm. It controls the complexity of the model by introducing Lasso regularization (L1) and Ridge regularization, which help to prevent overfitting and improve the model's generalizability. Additionally, the XGBoost model supports various types of base classifiers and trains iteratively by calculating the negative gradient of the model, gradually improving the model's accuracy.

#### 2.3.4. Analysis of Correlation Between CEs and Characteristic Factors

Although the XGBoost model has advantages such as high predictive accuracy and a fast prediction speed, its inherent “black box” nature in machine learning leads to a lack of transparency in the prediction process due to the complexity of the model, which contains a large number of parameters and connections. While the model excels at identifying correlations between input and output data, it cannot explain the causal relationships involved, resulting in limited interpretability [51–53]. For example, a disease prediction model may be able to predict high-risk populations but cannot explain the underlying reasons. To address this issue, a hybrid model based on an improved XGBoost model combined with SHAP is proposed. SHAP, derived from the Shapley value principle in game theory, calculates the marginal contribution of characteristic factors (influencing factors) to the model's predictions. The SHAP value (attribution value) corresponding to each feature can be understood as the importance of individual input features in contributing to the model's predicted value, revealing the thresholds and interactive synergistic effects between explanatory variables and the explained variable. The SHAP formula is:

$$\text{SHAP}(f, x, S) = \sum_{T \in S} \frac{|T|!(n - |T|)!}{n!} [f(x) - f(x_{S/T})]$$

where  $f$  is the prediction function of the model;  $x$  is the observed feature vector;  $S$  is the geometry of all features;  $n$  is the number of features; and  $x_{S/T}$  is the observed value of all features except  $T$ . The main impact of each variable was determined by subtracting the SHAP interaction value from the SHAP value.

#### 2.3.5. Precision Evaluation

The prediction accuracy of the machine model was evaluated through the coefficient of determination ( $R^2$ ), mean absolute error (MAE), mean-square error (MSE), root mean square error (RMSE), median absolute deviation (MAD), mean absolute percentage error (MAPE), expected value saved (EVS), and mean squared logarithmic error (MSLE) indicators. The calculation formulas for these indicators are as follows:

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

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

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

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

$$\text{MAD} = \sum |y_i - \bar{y}| / n$$

$$\text{MAPE} = \frac{1}{n} \sum (|y_i - \hat{y}_i| / y_i) * 100\%$$

$$\text{MAPE} = \frac{1}{n} \sum (|y_i - \hat{y}_i| / y_i) * 100\%$$

$$\text{MSLE} = \frac{1}{n} \sum_{i=0}^{n-1} (\log_e(1 + y_i) - \log_e(1 + \hat{y}_i))^2$$

where  $y_i$  is the actual value;  $\bar{y}_i$  is the actual mean;  $\hat{y}_i$  is the predicted value;  $n$  is the sample size; and  $\text{Var}$  is the variance. The RMSE, MAE, MSE, and MAPE range from 0 to  $+\infty$ , with a value of 0 indicating perfect alignment between predicted and actual values; thus, the larger the error, the greater the corresponding value.  $R^2$  and EVS range from 0 to 1, with scores closer to 1 indicating a better model fitting performance. The smaller the MSLE value, the slighter the difference between the predicted result and the true value, indicating a better fit of the model. The MAD represents the absolute value of the residual between the predicted value and the median, unaffected by outliers, and the smaller the better.

Based on these predictive accuracy metrics, a preliminary assessment and selection were made among the five algorithms. Further testing was required to assess the tolerance to missing values and understand the applicability of the optimal model in different scenarios.

### 3. Results and Analysis

#### 3.1. Model Performance and Comparison of Different Models

To verify the performance of the XGBoost model, multiple commonly used ML models were established for comparison, including the GBDT model, the RF model, the Adaboost model, and the traditional logistic model (Table 3, Figure 4). The results indicate that the accuracy of the XGBoost model surpasses that of the other ML models. The XGBoost model achieves a prediction  $R^2$  of 91.9%, demonstrating excellent performance with minimal differences between training and testing errors.

**Table 3.** Comparison table of the performance of different ML models.

|       | XGBoost | GBDT  | RF    | Adaboost | Logistic |
|-------|---------|-------|-------|----------|----------|
| $R^2$ | 0.919   | 0.732 | 0.836 | 0.538    | 0.610    |
| MAE   | 0.043   | 0.063 | 0.055 | 0.061    | 0.051    |
| MSE   | 0.012   | 0.015 | 0.011 | 0.013    | 0.035    |
| RMSE  | 0.112   | 0.122 | 0.106 | 0.114    | 0.215    |
| MAD   | 0.019   | 0.021 | 0.026 | 0.022    | 0.031    |
| MAPE  | 0.261   | 0.115 | 0.144 | 0.191    | 0.121    |
| EVS   | 0.291   | 0.160 | 0.358 | 0.257    | 0.481    |
| MSLE  | 0.008   | 0.009 | 0.007 | 0.008    | 0.439    |

The comparative analysis between the XGBoost model and the RF model found that the accuracy of the RF model is not significantly different from that of the XGBoost model. Both the RF and XGBoost models are ensemble learning algorithms based on decision trees. The difference lies in the fact that the RF model consists of multiple randomly composed decision trees, each of which is independent. Test data are evaluated separately in each decision tree, and the results from each tree are then combined to determine the majority output as the result of the RF model. Conversely, the decision trees of the XGBoost model are constructed sequentially, with each tree learning from the prediction results and residuals of previous decision trees. Therefore, the prediction accuracy of the XGBoost model is higher than that of the RF model.

A comparative analysis of the XGBoost model with the Adaboost model and the GBDT model reveals that the Adaboost model initially sets initial weights for each weak learner. If the prediction result of a weak learner after traversing all training data samples is inconsistent with the true result, the total error of that weak learner increases, and its weight changes accordingly. Through iteration, each weak learner obtains its corresponding

weight, which is then combined according to different weights to form a strong learner. Similarly, due to the iterative nature of the GBDT model, the training of each tree relies on the residuals of the previous tree. This makes it difficult to achieve full parallelization during training. When dealing with high-dimensional sparse datasets, invalid features can increase the computational load and may lead to overfitting. The performance of the GBDT model is inferior to that of other algorithms, such as the RF model and the XGBoost model.

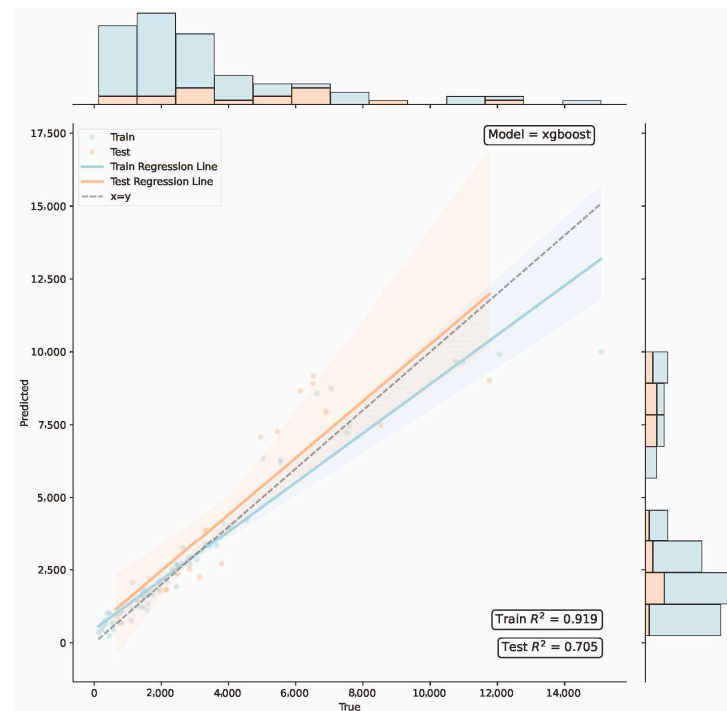


**Figure 4.** The correlation between predicted CEs and actual values.

Upon comparison and analysis of the XGBoost model and the traditional logistic model, we found the latter to be a nonlinear regression model with a logistic function form. Regression models are prone to underfitting and have a limited ability to handle nonlinear relationships, difficulty in handling data imbalance issues, and limited feature selection capabilities, resulting in significantly lower prediction accuracy compared with other ML models. In summary, the XGBoost model demonstrates certain advantages.

The XGBoost model combined with SHAP values can handle high-dimensional feature models, capture nonlinear relationships, and does not rely on the continuity of spatial data, making it suitable for complex spatial data analysis. It has certain advantages, especially when dealing with nonlinear relationships and high-dimensional feature data. The XGBoost model method boasts numerous advantages, including: (1) extremely high prediction accuracy, especially compared with traditional statistical models, with a superior predictive capability; (2) no restrictions on data types (such as continuous and categorical variables) and data distributions; (3) better handling of missing values and outliers; (4) no pre-defined relationship type (such as linear) between independent and dependent variables, allowing for better detection of potential nonlinear relationships; and (5) the XGBoost model combined with SHAP values predicts the contribution of each feature, providing a high degree of interpretability and enhancing the model's transparency. These advantages have made the XGBoost model one of the most popular ML methods in recent years. However, the XGBoost model also has some drawbacks, such as an inability to perform significance tests and a lack of coefficients and confidence intervals for independent variables, similar to other ML models. However, the XGBoost model can provide the relative importance of feature variables, which reflects the magnitude of the impact of independent variables on

the dependent variables. At the same time, commonly used partial dependence plots can visualize the relationship between changes in independent variables and the dependent variables, which can compensate for the aforementioned disadvantages.

### 3.2. Analysis of Feature Importance Based on SHAP Values

The contribution of independent variables to the prediction of CEs from residential land was analyzed through the XGBoost model combined with SHAP values, as illustrated in Figure 5. The relative importance was used to accurately indicate the contribution and ranking of a certain independent variable to the prediction of the CEs, as demonstrated in Table 4. Among the independent variables, BD had the greatest impact on CEs from residential land, with a relative importance of 33.16%. This was followed by FAR, land use mix rate (Phh), road network density (RND), PPD, and GFA. The variables with a lower contribution were WFD, POI density, GR, and intersection density (ITD). Most of the SHAP values for these points were scattered around 0, indicating that these variables have a small impact on CEs from residential land. BD, FAR, RND, PPD, and GFA had a positive impact on the model prediction, while public transport network density (PTND) and GR had a negative impact on the model.

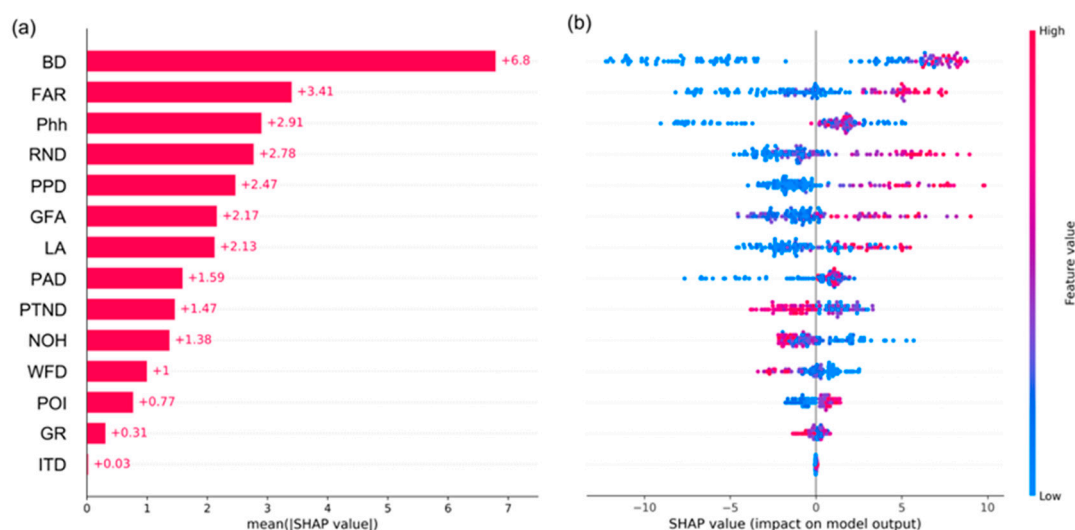


**Figure 5.** SHAP feature importance of built environment factors. (a) Importance score ranking of the model prediction characteristics; (b) Every feature's impact on the model's output.

**Table 4.** Relative importance of built environment characteristic indicators.

| Variable                                | Relative Importance% | Rank |
|-----------------------------------------|----------------------|------|
| Land area (LA)                          | 2.21                 | 7    |
| Number of households (NOH)              | 0.61                 | 10   |
| Gross floor area (GFA)                  | 4.30                 | 6    |
| Floor area ratio (FAR)                  | 27.74                | 2    |
| Greening rate (GR)                      | 0.20                 | 13   |
| Building density (BD)                   | 33.16                | 1    |
| Permanent population density (PPD)      | 6.04                 | 5    |
| Workforce density (WFD)                 | 0.56                 | 11   |
| Land use mix rate (Phh)                 | 11.93                | 3    |
| Road network density (RND)              | 10.72                | 4    |
| Intersection density (ITD)              | 0.15                 | 14   |
| Public transport network density (PTND) | 0.84                 | 9    |
| POI density                             | 0.33                 | 12   |
| Parking area ratio (PAD)                | 1.21                 | 8    |

### 3.3. The Nonlinear Impact and Threshold Effect of the Built Environment on CEs from Residential Land

In order to better reveal the complex nonlinear relationship between the dependent variables and independent variables, a fitting curve was introduced into the predicted SHAP scatter plot through locally weighted scatterplot smoothing (LOWESS) to smooth the variations between data points, visually capture the trend of the feature variables, and exhibit the nonlinear relationship and threshold effect between the built environment's independent variable characteristic indicators and residential land CEs (Figure 6).

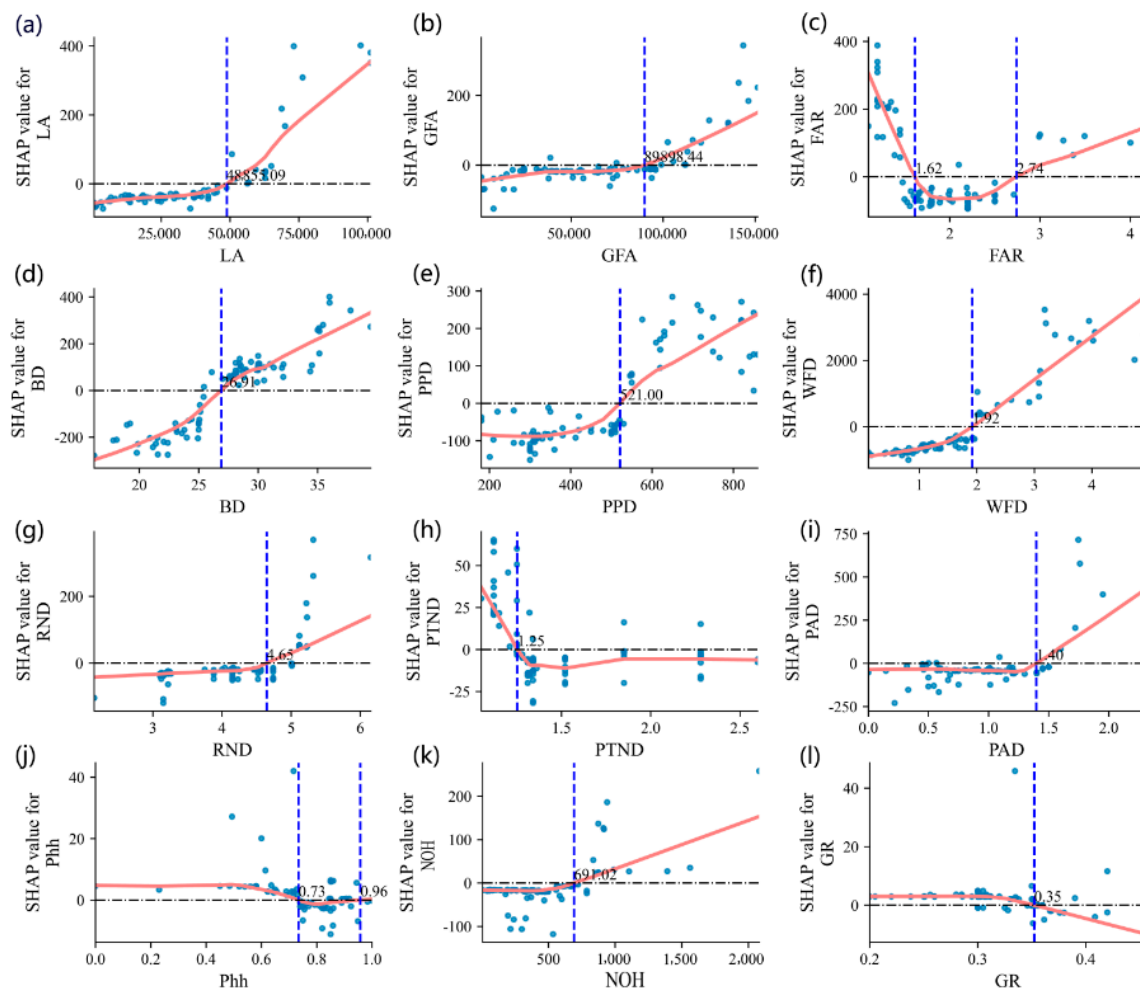


**Figure 6.** Nonlinear impact of built environment indicators on CEs from residential land. (a) LA, (b) GFA, (c) FAR, (d) BD, (e) PPD, (f) WFD, (g) RND, (h) PTND, (i) PAD, (j) Phh, (k) NOH, (l) GR.

As shown in Figure 6a,b,e,f,k, as the values of LA, GFA, BD, PPD, WFD, and NOH increased, the corresponding SHAP values also increased, indicating that higher feature values have a greater positive impact on the model's predictions, leading to an increasing trend in CEs. When the LA value is below 48,855.09 m<sup>2</sup>, the GFA value is below 89,898.44 m<sup>2</sup>, the BD value is below 26.91%, the PPD value is below 521.00 ten thousand people/km<sup>2</sup>, and the WFD value is below 1.92, the SHAP values are negative, indicating that excessively low LA, GFA, BD, PPD, WFD, and NOH values are negatively correlated with CEs from residential land, which aligns with conclusions from existing research [54–59]. A larger community land and total building area implies a greater number of permanent residents. At the same time, high LA and GFA values affect factors such as sunlight, thermal radiation, and air flow on the plot. An excessively high GFA increases the plot's absorption of solar radiation, thereby affecting the local microclimate of the residential area. Rosenfeld A

believes that, as the BD on a plot increases, the heat island effect will intensify, leading to an increase in CEs [56]. Jiang Haiyan et al. conducted field surveys in 33 communities in Guangzhou and similarly found a positive correlation between the building area of the community and CEs [18]. Wang Xiaodan's research found that an expansion in community size and an increase in total building area would result in higher CEs [57]. On the other hand, high BD values can reduce the distance between buildings to some extent, leading to poorer lighting and ventilation conditions for some buildings. To meet indoor lighting and comfort needs, residents may need to rely more on artificial lighting and mechanical ventilation systems, thereby increasing electricity consumption and resulting in higher CEs. High BD values can also cause shading between buildings, affecting the utilization of solar energy in winter and increasing energy consumption for heating. The Urban Morphology Research Laboratory in France analyzed the relationship between the average plot BD and CEs in 25 cities around the world, proving that BD is a key factor influencing CEs from built-up areas [58]. Zhou Xuefan conducted simulations based on a mesoscale meteorological model (Weather Research Forecasting, WRF) and found that neighborhoods with higher BD values also had correspondingly higher building energy consumption [59].

As shown in Figure 6c, communities with lower FAR values are more likely to generate high CEs. There is a U-shaped curve between plot ratios and residential land CEs, with critical values of 1.62 and 2.74. FAR values above the critical value have a promoting effect on CEs. Further analysis reveals that the impact of FAR on CEs from residential land is related to its land use intensity. When residential land is utilized in an extensive manner, the high base coverage leads to an increase in FAR, which, in turn, has a positive effect on CEs. Conversely, when residential land is utilized in a compact manner, the low base coverage results in a lower FAR, leading to higher CEs from residential land [60].

As the Phh within residential land increased, the smaller corresponding SHAP value indicated a more pronounced negative predictive impact on the model. When the Phh exceeded approximately 0.73, it exerted a negative influence on CEs from residential land (Figure 6j). Masnavi's research shows that, as the Phh within residential areas increases, CEs from these areas gradually decrease. The Phh not only affects the CEs from buildings on a plot but also influences residents' choice of transportation modes. Residential land with a higher Phh can reduce the probability of vehicle travel by 70% compared with land with a lower mix, thereby significantly decreasing CEs from residential land [61].

As the RND and parking area ratio (PAD) increased, the corresponding SHAP value also increased. When the RND was higher than 4.65 and the PAD was greater than 1.40, there was a clear proportional relationship between the interval and the CEs (Figure 6g,i). Multi-story residential land with a relatively high PAD and built after 2015 had the highest CEs. The proportion of residents traveling by car in such residential areas was as high as 50%, which increased the CEs of residential land to a certain extent. Residents in residential land that were built earlier had a relatively higher proportion of walking and non-vehicle travel, resulting in lower per capita CEs.

As Figure 6h,l illustrate, an increase in PTND and GR corresponds to a lower SHAP value, indicating a negative correlation with CEs from residential land. A high PTND near residential areas encourages residents to opt for low-carbon modes of transportation, thereby reducing CEs to some extent. The GR within residential areas can enhance the local microclimate, absorbing carbon dioxide and releasing oxygen, which helps lower greenhouse gas emissions, mitigate the urban heat island effect, and reduce the usage of air conditioning and cooling equipment, ultimately leading to decreased energy consumption and CEs from these areas.

### 3.4. Analysis of Interaction Effects Among Key Built Environment Variables

The feature interaction dependency graph illustrates the complex interaction relationships among multiple feature indicators (Figure 7). The analysis results indicate that multiple characteristic indicators interact with BD and there is a positive interaction between GFA and BD. Both factors interact in the same direction, indicating that a higher GFA and BD will positively interact with and exacerbate CEs. Conversely, there is a negative interaction between FAR and BD. When the FAR is below approximately 2.0 and the BD exceeds approximately 35%, the two exhibit a positive interaction, thereby increasing CEs in residential land.

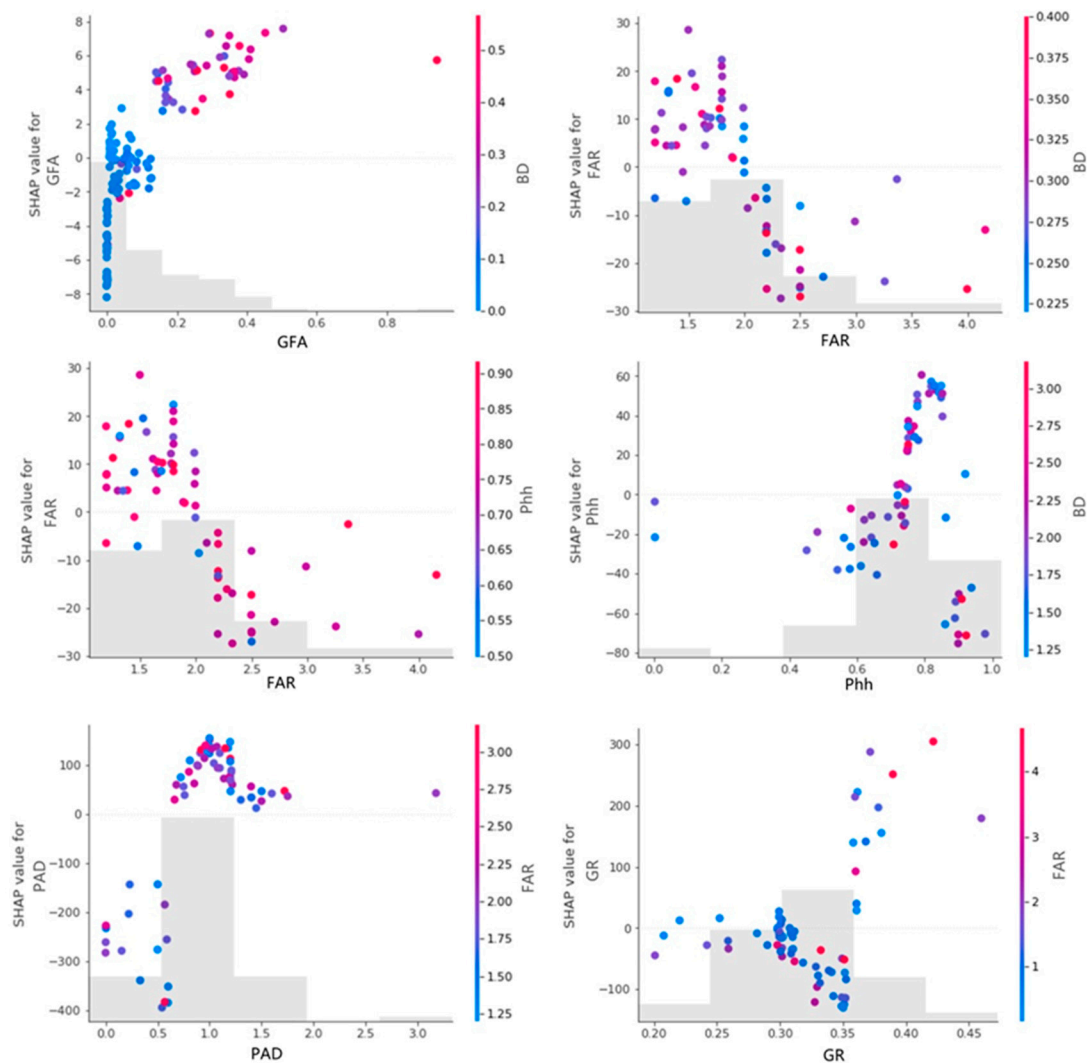


**Figure 7.** The interactive effect of independent variables on CEs from residential land.

The impact of land Phh and BD on CEs in residential land is inversely related. Only when the Phh is lower than 0.8 and combined with a high BD do the two positively interact, thereby increasing CEs from residential land. When the FAR is less than approximately 2.0, an increase in Phh generally results in stable or even slightly reduced CEs. This suggests that reducing the FAR can be an effective strategy to decrease CEs from residential land. Additionally, PAD and public PTND are mutually exclusive, and both values at a low level negatively impact CEs. Therefore, to reduce CEs from residential land, it is necessary to match the two within a reasonable range. The existence of these interactive effects indicates that changes in a single factor may produce unexpected results through interactions with other factors.

## 4. Conclusions and Limitations

### 4.1. Conclusions

This study took 84 residential plots in the Susong County of Anhui Province as its research object and used the XGBoost model combined with the SHAP value analysis method to deeply explore the nonlinear relationship between built environment factors and residential land CEs, providing data support and guidance to planning practice for achieving the goal of urban carbon neutrality. The main conclusions are as follows:

- (1) Comparative analysis revealed that the XGBoost model exhibits excellent performance. With a fitting degree of 91.9% and good prediction accuracy, the XGBoost model significantly outperforms mainstream ML models such as the GBDT model, the RF model, the Adaboost model, and the traditional logistic model, validating its outstanding advantages in quantitatively analyzing nonlinear relationships. Combined with the SHAP value analysis corresponding to each feature, the interpretability of the model was further enhanced, providing an intuitive analytical path for revealing the complex relationships between built environmental factors;
- (2) In the analysis of key influencing factors and their importance, analysis of the SHAP value showed that BD, FAR, and Phh are key factors affecting CEs from residential land. The relative importance of these features was 33.16%, 27.84%, and 18.72%, respectively. In addition, RND and PPD are also important variables that cannot be ignored. The results show that the CEs of residential land are affected by the comprehensive influence of multi-dimensional built environment characteristics, but the marginal effects of some variables are more prominent;
- (3) In the analysis of the nonlinear relationship and threshold effect, we found a significant nonlinear relationship and threshold effect between residential land CEs and major built environment factors. For example, FAR and carbon emissions show a “U”-curve relationship. When the FAR reached a critical value (1.62 to 2.74), CEs began to increase significantly. An increase in LA, GFA, BD, and PPD also showed a positive dynamic effect on CEs. In addition, Phh, PTND, and GR had a restraining effect on CEs. This indicates that, in the actual planning process, the complex interactions between variables should be rationally balanced to avoid the negative effects of high CE “inflection points”;
- (4) In terms of the interaction effects between features and optimization strategies, through an analysis of feature interaction dependency graphs, it was found that the interaction effects between built environment elements deserve attention. For example, there was a positive interaction between BD and GFA, and a joint increase in the two will aggravate CEs. There was a negative interaction between FAR and BD, and reducing the FAR can be an effective means of reducing CEs. Combining these mechanisms, urban planners should implement structural interventions based on multi-feature collaborative optimization design, targeting key variables such as FAR and Phh to achieve a win-win situation of carbon emission reductions and efficiency improvements;
- (5) In terms of insights from planning practice, the findings of this study offer specific implementation pathways for low-carbon urban planning, encompassing optimizing FAR, enhancing Phh and PTND, and reasonably controlling BD. Simultaneously, this study underscores the significance of multi-feature collaborative design in enhancing carbon reduction outcomes, thereby providing intuitive and practical guidance for policymakers and planning practitioners;
- (6) In terms of the significance of the theoretical innovation, introducing the XGBoost model into the field of CEs from residential land provides a new and efficient predictive tool for this area that has high theoretical value. The XGBoost model has

advantages such as a fast computation speed and the ability to handle large-scale data. Compared with traditional methods like linear regression models and time series analysis, it can more accurately capture the nonlinear relationships between CEs from residential land and various complex built environment factors, improving the accuracy and reliability of predictions and providing stronger technical support for subsequent research and practice. Secondly, research on land use has always been one of the core contents of spatial planning theory and practice. As a fundamental basis for guiding low-carbon urban development, spatial planning mainly achieves this by controlling land use characteristics such as land nature, development intensity, and spatial form. Through the analysis of the results from the XGBoost model, it is possible to quantitatively reveal the relationships between specific built environment factors such as BD, FAR, PPD, and the completeness of transportation facilities with CEs from residential land. This changes our previous qualitative or vague understanding of how built environment factors affect CEs, allowing for better control of land use characteristics such as land nature, development intensity, and spatial form. The potential for using spatial planning as a structural adjustment method for emission reductions in CE control is enormous. It not only provides a more accurate theoretical basis for governments and relevant departments to formulate targeted emission reduction measures but also offers scientific support for developing urban planning and environmental protection policies, enhancing the relevance and effectiveness of these policies. Finally, the research findings contribute to enriching and improving the development of low-carbon city theory. In-depth studies on CEs from residential land and their relationship with built environment factors can provide empirical support and a theoretical foundation for constructing a more comprehensive and systematic theoretical framework for low-carbon cities. This will promote the application and development of low-carbon city theory in practice and provide theoretical guidance for achieving urban sustainable development goals.

#### 4.2. Limitations

Although this study achieved certain innovative results, there are also some shortcomings that require further improvement in future research:

- (1) The data coverage and variables need further expansion. The limitations of the data sources have resulted in some important variables not being included in the analysis. For example, factors such as household income, occupational structure, and differences in user energy consumption behavior may indirectly affect residential land CEs. The absence of these variables may lead to an incomplete analysis of the impact mechanism of residential land CEs. Future research should expand the data collection scope from multiple channels and introduce more socioeconomic variables to further optimize the explanatory power of the model;
- (2) Further multi-model comparative analysis is needed to optimize model algorithms. Currently, ML models are primarily based on the XGBoost model for analysis. However, in the field of ML, new algorithms, such as the LightGBM model, have demonstrated excellent performance in multiple domains due to their high efficiency and low memory consumption. Future research could attempt to introduce emerging algorithms, such as LightGBM and CatBoost, for supplementary analysis and model comparison, thereby further enhancing the model's performance and verifying the robustness of conclusions;
- (3) More comparisons of regional differences and spatial multi-scale analysis need to be performed. This study took Susong County in Anhui Province as a case study, which has a certain degree of regional representativeness. However, there are significant

- differences in built environment characteristics, resident behavior patterns, and carbon emission structures among different regions. In the future, cross-regional comparative studies can be conducted on a larger scale while introducing multi-scale analysis methods, ranging from the block scale to the overall city scale, to explore the impact of spatial scale changes on model performance and the importance of key variables;
- (4) Policy applications and emission reduction simulation testing need to be optimized. This study provides several quantitative results and policy recommendations but fails to further simulate the dynamic impact of the implementation of specific planning policies on CEs. In the future, policy experiments and dynamic simulation tools can be combined to assess the actual emission reduction effect of specific planning schemes through scenario construction and intervention simulation, providing more reliable data support for policy feasibility studies.

In summary, this study not only verifies the superiority of the XGBoost model in predicting CEs from residential land and analyzing the nonlinear relationship between built environmental factors and CEs from residential land but also reveals the crucial significance of optimizing built environmental factors of residential land use in achieving urban carbon neutrality. Future research needs to further improve the data and algorithm systems, expand the scope of research, deepen policy applications, and provide more comprehensive and scientific empirical evidence and decision support for the construction of low-carbon cities.

**Author Contributions:** Methodology, C.X. and W.X.; software, S.Z.; validation, S.Z. and S.W.; formal analysis, S.Z.; investigation, H.S.; data curation, H.S.; writing—original draft preparation, S.Z. and C.X.; writing—review and editing, S.B. and T.X.; visualization, S.Z.; project administration, C.X.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Natural Science Foundation of China (42371249), the Anhui Social Science Innovation and Development Research Project (2022CX188), the Scientific Research Project for Colleges and Universities in Anhui Province (2022AH050128), the Project of the Open Research Fund of the Anhui Provincial Key Laboratory of Building Earthquake Disaster Mitigation and Green Operations, Anhui Institute of Building Research & Design (Grant NO. 2023-JKYL-001), and the Research Project of the Anhui Academy of Architectural Sciences and Design (Sensitivity Simulation of Urban Carbon Emission Factors and Low-Carbon Land Layout).

**Data Availability Statement:** The data that have been used are confidential.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Jiang, L.; Ding, B.-N.; Shi, X.-N.; Li, C.-H.; Chen, Y.-M. Household Energy Consumption Patterns and Carbon Emissions for the Megacities-Evidence from Guangzhou, China. *Energies* **2022**, *15*, 2731. [\[CrossRef\]](#)
2. Wei, X. Research on Reducing Carbon Consumption in Residential Community Spaces as Influenced by Microclimate Environments. *Urban Plan. Dev.* **2021**, *147*, 04021037. [\[CrossRef\]](#)
3. Wang, Y. Carbon Emissions from Food Consumption and Reduction Potential in Urban Residents: A Case Study of Provincial Capitals in the Middle and Lower Reaches of the Yellow River. *Sustainability* **2025**, *17*, 690. [\[CrossRef\]](#)
4. Ding, X.-X.; Xu, X.-Z.; Wang, Q.; An, X.-D. Research on Influencing Factors of Carbon Emission of Prefabricated. *J. Eng. Manag.* **2022**, *36*, 47–51. [\[CrossRef\]](#)
5. Wang, C.-J.; Wang, F.; Zhang, X.-L.; Wang, Y.; Su, Y.-X.; Ye, Y.-Y.; Wu, Q.-T.; Zhang, H.-O. Dynamic features and driving mechanism of coal consumption for Guangdong province in China. *J. Geogr. Sci.* **2022**, *32*, 401–420. [\[CrossRef\]](#)
6. Wang, H.-P.; Yang, J.-Z. Scenario Prediction of Carbon Emissions from the Paper Industry in Guangdong Province Based on the Extended STIRPAT Model. *Environ. Sci.* **2024**, *11*, 1–17. [\[CrossRef\]](#)
7. Wang, H.-B.; Li, B.-W.; Khan, M.-Q. Prediction of Shanghai electric power carbon emissions based on improved STIRPAT model. *Sustainability* **2022**, *14*, 13068. [\[CrossRef\]](#)

8. Yang, F.; Shi, L.; Gao, L. Probing CO<sub>2</sub> emission in Chengdu based on STRIPAT model and Tapio decoupling. *Sustain. Cities Soc.* **2023**, *89*, 104309. [\[CrossRef\]](#)
9. Liu, Z.-J.; Ju, M.-T. Scenario Analysis of Carbon Peaking in a Northern Coastal City Based on the STIRPAT Model. *Environ. Sci.* **2024**, *9*, 1–23. [\[CrossRef\]](#)
10. Hu, H.-W.; Wang, Q.-Z.; Zhu, L.; Zhang, Y.-B. Prediction and Analysis of Carbon Emissions in Building Sector Based on LEAP Model and LMDI Decomposition. *J. Beijing Univ. Civ. Eng. Archit.* **2023**, *39*, 80–87. [\[CrossRef\]](#)
11. Yuan, K.-H.; Gan, C.-L.; Yang, H.-L.; Liu, Y.; Chen, Y.-R.; Zhu, Q.-Y. Validation of the EKC and Characteristics Decomposition between Construction Land Expansion and Carbon Emission: A Case Study of Wuhan City. *China Land Sci.* **2019**, *33*, 56–64.
12. Gill, R.-A.; Viswanathan, K.-K.; Hassan, S. A test of environmental Kuznets curve (EKC) for carbon emission and potential of renewable energy to reduce green house gases (GHG) in Malaysia. *Environ. Dev. Sustain.* **2018**, *20*, 1103–1114. [\[CrossRef\]](#)
13. Feng, Y.-X.; Hong, Z.-X.; Cheng, J.; Jia, L.-K.; Tan, J.-R. Low Carbon-Oriented Optimal Reliability Design with Interval Product Failure Analysis and Grey Correlation Analysis. *Sustainability* **2017**, *9*, 369. [\[CrossRef\]](#)
14. Wang, Y.; Wu, Y. Analysis and forecast of carbon emissions in the Yangtze River Delta region. *J. Anhui Agric. Univ.* **2023**, *50*, 1051–1058. [\[CrossRef\]](#)
15. Tang, S. Research on Low-Carbon Building Development and Carbon Emission Control Based on Mathematical Models: A Case Study of Jiangsu Province. *Energies* **2024**, *17*, 4545. [\[CrossRef\]](#)
16. He, W.-J.; Liu, D.; Wang, C. Are Chinese Provincial Carbon Emissions Allowances Misallocated over 2000–2017? Evidence from an Extended Gini-Coefficient Approach. *Sustain. Prod. Consum.* **2022**, *29*, 564–573. [\[CrossRef\]](#)
17. Qi, T.; Tao, Z.; Rong, Y. The nexus of electricity infrastructure investment, household income, and carbon emissions in rural China. *Util. Policy* **2024**, *91*, 101849. [\[CrossRef\]](#)
18. Jiang, H.-Y.; Xiao, R.-B.; Wu, J. The Impact Pattern of Urban Household Carbon Emission and Implications for the Low-Carbon Resident Community Planning: A Case Study of Guangzhou. *Mod. Urban Res.* **2013**, *28*, 100–106. [\[CrossRef\]](#)
19. Wang, W.-Q.; Li, J. Correlation Between Residential Density and Carbon Emissions Caused by Household Energy Consumption: A Case Study on Caoyang Xincun, Shanghai. *Urban Plan.* **2017**, *41*, 83–91.
20. Wu, W.; Song, Y.; Hong, Z.-S.; Lu, Y.-H. Research of the Influence of Community Form on residential Energy Consumption: A Case Study on Ningbo City. *Urban Dev. Stud.* **2018**, *25*, 15–20+28. [\[CrossRef\]](#)
21. Zhangli, C.-Y.; Sun, Q.-Z.; Wu, H.-T.; Xu, H.-L.; Du, M.-G.; Yuan, D.-C. Relationship Between Carbon Emissions and Residential Environment: Taking residential buildings in Tianjin as Example. *Build. Energy Effic.* **2020**, *48*, 93–101+107. [\[CrossRef\]](#)
22. Hou, S.-N.; Dang, G.-F. Simulation of urban thermal environment based on machine learning and research on the temporal and spatial variation characteristics of urban heat island. *Sci. Surv. Mapp.* **2022**, *47*, 200–207. [\[CrossRef\]](#)
23. Yu, H.; Yang, Y.; Li, B.; Liu, B.-W.; Guo, Y.-H.; Wang, Y.-Q.; Guo, Z.-F.; Meng, R.-H. Research on the community electric carbon emission prediction considering the dynamic emission coefficient of power system. *Sci. Rep.* **2023**, *13*, 5568. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Li, Y.-Z.; Wu, Y.-T.; Yu, J.; Bai, L.-H.; Wang, L.-X. Prediction Model for Carbon Emissions of Life Cycle Assessment of High-rise Office Buildings Based on SVR: Take Tianjin as an Example. *J. BEE* **2021**, *49*, 25–30. [\[CrossRef\]](#)
25. Deng, F.-M.; Xu, L.; Fang, Y.; Gong, Q.-X. PCA-DEA-tobit regression assessment with carbon emission constraints of China's logistics industry. *J. Clean. Prod.* **2020**, *271*, 122548. [\[CrossRef\]](#)
26. Kong, F.-W.; Zhang, G.-H.; Wang, Y.-H. The Importance Research of Influencing Factors of Carbon Emissions in Construction Industry in Liaoning Province Based on Random Forest Algorithm. *J. Shenyang Jianzhu Univ.* **2023**, *25*, 58–63. [\[CrossRef\]](#)
27. Yu, W.-M.; Zhang, T.-T.; Shen, D.-J. County-level spatial pattern and influencing factors evolution of carbon emission intensity in China: A random forest model analysis. *China Environ. Sci.* **2022**, *42*, 2788–2798. [\[CrossRef\]](#)
28. Wang, W.; Su, W. Impacts of urban block form on carbon and pollutant emissions from urban life in China from the perspective of regional differences. *Sustain. Cities Soc.* **2024**, *115*, 105849. [\[CrossRef\]](#)
29. Chen, C.; Tang, Y. Nonlinear effects of built environment on travel safety of the elderly pedestrians: A case in Yuzhong District, Chongqing. *Sci. Technol. Eng.* **2023**, *23*, 7112–7119.
30. Samat, A.; Li, E.; Wang, W.; Liu, S.; Abuduwaili, J. Meta-XGBoost for Hyperspectral Image Classification Using Extended MSER-Guided Morphological Profiles. *Remote Sens.* **2020**, *12*, 1973. [\[CrossRef\]](#)
31. Piraei, R.; Afzali, S.-H.; Niazkar, M. Assessment of XGBoost to Estimate Total Sediment Loads in Rivers. *Water Resour. Manag.* **2023**, *37*, 5289–5306. [\[CrossRef\]](#)
32. Wu, Y.; Zhang, Q.; Hu, Y.; Sun-Woo, K.; Zhang, X.; Zhu, H.; Jie, L.; Li, S.-Y. Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems. *Future Gener. Comput. Syst.* **2022**, *129*, 1–12. [\[CrossRef\]](#)
33. Shao, Z.; Ahmad, M.-N.; Javed, A. Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface. *Remote Sens.* **2024**, *16*, 665. [\[CrossRef\]](#)
34. Uddin, M.-N.; Al-Amin; Hossain, S. Revolutionizing engineered cementitious composite materials (ECC): The impact of XGBoost-SHAP analysis on polyvinyl alcohol (PVA) based ECC predictions. *Low-Carbon Mater. Green Constr.* **2024**, *2*, 11. [\[CrossRef\]](#)

35. Ma, S.-H.; Chen, X.-F.; Yang, L.; Zhao, Y.-Z.; Zeng, Y. Impacts of Accessibility on Spatial Heterogeneity of Multimodal Transportation Carbon Emissions in Urban Agglomeration. *J. Transp. Syst. Eng. Inf. Technol.* **2024**, *24*, 64–74. [\[CrossRef\]](#)
36. Wang, K.; Lu, J.; Liu, A.; Song, Y.-L.; Xiong, L.; Zhang, G.-Q. Elastic gradient boosting decision tree with adaptive iterations for concept drift adaptation. *Neurocomputing* **2022**, *491*, 288–304. [\[CrossRef\]](#)
37. Hu, Z.; Gong, X.; Liu, H. Prediction of household consumption carbon emission in western cities Based on BP model: Case of Xi'an city. *J. Arid. Land Resour. Environ.* **2020**, *34*, 82–89. [\[CrossRef\]](#)
38. Yan, F.-Y.; Liu, S.-X.; Zhang, X.-P. Prediction of Carbon Emission for Land Use Based on PCA-BP Neural Network. *J. Hum. Settl. West China* **2021**, *36*, 1–7. [\[CrossRef\]](#)
39. Aly, M.-A.; Alsedais, N. Heat and mass transfer of NEPCM in a wavy porous cavity with variable hot and high-concentration zones: A study using ISPH and XGBoost models. *Int. Commun. Heat Mass Transf.* **2025**, *160*, 108383. [\[CrossRef\]](#)
40. Lin, Y.-L.; Xiong, J.-J.; Xing, H.; Ning, X. Research on Carbon Emission Prediction Method for Highway Construction Based on XGBoost SVR Combination Model. *J. Cent. South Univ.* **2024**, *55*, 2588–2599. [\[CrossRef\]](#)
41. Li, Y.-T. *Prediction of Automobile CO<sub>2</sub> Emission Based on SHAP and BS-XGBoost*; Chongqing Jiaotong University: Chongqing, China, 2023.
42. Yang, D.-F.; Wang, X.-M.; Han, R.-N. Nonlinear and Synergistic Effects of the Built environment on Street Vitality: The Case of Shenyang. *J. Urban Plan.* **2023**, *5*, 93–102. [\[CrossRef\]](#)
43. Yu, W.-J.; Chen, H.-W.; Qi, H.-L.; Pan, Z.-L.; Li, H.-W.; Hu, D.-B. LSTM-XGBoost Based RR Intervals Time Series Prediction Method in Hypertensive Patients. *Chin. J. Med. Instrum.* **2024**, *48*, 392–395. [\[CrossRef\]](#)
44. Liu, X.-Z.; Xie, Z.-L.; Zhang, Y.; Huang, J.; Kuang, L.-R.; Li, X.-J.; Li, J.; Zou, Y.-X.; Xiang, T.-Y.; Yin, N.-Y.; et al. Machine learning for predicting in-hospital mortality in elderly patients with heart failure combined with hypertension: A multicenter retrospective study. *Cardiovasc. Diabetol.* **2024**, *23*, 2503–2509. [\[CrossRef\]](#)
45. Santos, R.-M.; Guedes, A.; Gendriz, S.-I. SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 316–341. [\[CrossRef\]](#)
46. Alireza, F.; Roya, E.; Hamid, N. Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods. *Diagnostics* **2023**, *13*, 2391. [\[CrossRef\]](#)
47. Shi, W.; Xiang, Y.; Ying, Y.; Jiao, Y.; Zhao, R.; Qiu, W. Predicting Neighborhood-Level Residential Carbon Emissions from Street View Images Using Computer Vision and Machine Learning. *Remote Sens.* **2024**, *16*, 1312. [\[CrossRef\]](#)
48. Wang, Z.-Q.; Ren, J.-G.; Han, S.; Li, W.-C. Interpretable machine learning-based carbon emission prediction in the materialization stage of buildings. *J. Saf. Environ.* **2024**, *24*, 2454–2466. [\[CrossRef\]](#)
49. Ewing, R.; Tian, G.; Goates, J.-P.; Zhang, M.; Greenwald, M.J.; Joyce, A.; Kircher, J.; Greene, W. Varying influences of the built environment on household travel in 15 diverse regions of the United States. *Urban Stud.* **2015**, *52*, 2330–2348. [\[CrossRef\]](#)
50. Chen, T.-Q.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
51. Meng, Y.; Yang, N.-H.; Qian, Z.-L.; Zhang, G.-Y. What makes an online review more helpful: An interpretation framework using XGBoost and SHAP values. *J. Theor. Appl. Electron. Commer. Res.* **2020**, *16*, 466–490. [\[CrossRef\]](#)
52. Laphrom, W.; Se, C.; Champahom, T.; Jomnonkwao, S.; Wipulanusat, W.; Satiennam, T.; Ratanavaraha, V. XGBoost-SHAP and Unobserved Heterogeneity Modelling of Temporal Multivehicle Truck-Involved Crash Severity Patterns. *Civ. Eng. J.* **2024**, *10*, 2476–3055. [\[CrossRef\]](#)
53. Zhu, X.-H.; Lu, K.-F.; Peng, Z.-R.; He, H.-D.; Xu, S.-Q. Spatiotemporal variations of carbon dioxide (CO<sub>2</sub>) at Urban neighborhood scale: Characterization of distribution patterns and contributions of emission sources. *Sustain. Cities Soc.* **2022**, *78*, 103646. [\[CrossRef\]](#)
54. Wang, R.-F.; Wang, M.-M. Multi-scale analysis of surface thermal environment in relation to urban form: A case study of the Guangdong-Hong Kong-Macao Greater Bay Area. *Sustain. Cities Soc.* **2023**, *99*, 104953. [\[CrossRef\]](#)
55. Li, H.-Y.; Yang, T.; Du, Y.; Tan, Y.-N.; Wang, Z.-F. Interpreting hourly mass concentrations of PM<sub>2.5</sub> chemical components with an optimal deep-learning model. *J. Environ. Sci.* **2025**, *151*, 125–139. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Rosenfeld, A.; Arthur, H.; Akbari, H.; Bretz, S.; Fishman, B.L.; Stamper-Kurn, D.M.; Sailor, D.; Taha, H. Mitigation of Urban Heat Islands: Materials, Utility Programs, Updates. *Energy Build.* **1995**, *22*, 255–265. [\[CrossRef\]](#)
57. Wang, X.-D. Residential Building Carbon Emission Analysis and Accounting. *Green Build.* **2022**, *14*, 4. [\[CrossRef\]](#)
58. Urban Morphology Lab. Building Tomorrow Sustainable Cities. 2015. Available online: <http://www.doc88.com/p-818902890885.html> (accessed on 25 November 2024).
59. Zhou, X.-F. *Research on the Impact of Urban Morphology on Urban Climate in the Built up Zone*; Huazhong University of Science and Technology: Wuhan, China, 2013.

60. Wang, H.; Wu, J.-L.; Zhang, A.-L.; Nie, X. Analysis of the Effects of Land Use Intensity on Carbon Emissions: A Case Study of Nanning City, Guangxi Zhuang Autonomous Region. *Ecol. Econ.* **2016**, *32*, 42–46+99. [[CrossRef](#)]
61. Masnawi, M.-R. The New Millennium and the New Urban Paradim: The Compact City in Practice. In *Achieving Sustainable Urban Form*; E&FN Spon: London, UK; New York, NY, USA, 2000.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.