remote sensing

MDPI

Article

Application of Nighttime Light Data Simulation Based on Multi-Indicator System and Machine Learning Model in Predicting Potentially Suitable Economic Development Areas: A Case Study of the Turpan–Hami Region

Guangpeng Zhang [1], Li Zhang [2], Yiyang Chen [2,3], Meng Chen [3], Jingjing Tian [4] and Yin Wu [5,*]

[1] College of Geography and Remote Sensing Science, Xinjiang University, Urumqi 830017, China; 107552203722@stu.xju.edu.cn
[2] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; zhangli@aircas.ac.cn (L.Z.); chenyiyang21@mails.ucas.ac.cn (Y.C.)
[3] University of Chinese Academy of Sciences, Beijing 100094, China; chenmeng243@mails.ucas.ac.cn
[4] College of Civil Engineering and Architecture, Xinjiang University, Urumqi 830047, China; 107556522288@stu.xju.edu.cn
[5] Xinjiang Uygur Autonomous Region Natural Resources Planning Research Institute, Urumqi 830011, China
* Correspondence: wu_yinn@126.com

Abstract: In recent years, the accelerated urbanization process in China has led to increased land resource constraints and unregulated expansion, imposing significant pressure on ecosystems and the environment. As a critical node along the Silk Road Economic Belt, the Turpan–Hami region has experienced rapid urban development under policy support but faces challenges in resource utilization efficiency and sustainable development. To address these challenges, this study innovatively combines nighttime light remote sensing data to quantify urban economic development intensity and integrates socioeconomic and natural environment indicators based on previous research. Four tree-based ensemble learning models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost)—were employed to predict potential urban economic development suitability zones and their suitability intensity. The results show that the CatBoost model performed the best in suitability prediction, revealing significant spatial disparities: high-suitability areas are concentrated in regions with superior resource conditions and well-developed infrastructure, whereas areas with terrain constraints and inadequate infrastructure exhibit lower suitability. An analysis of changes over historical periods (2010, 2015, and 2020) demonstrates a gradual expansion of high-suitability regions over time.

Keywords: urban economic development; machine learning; nighttime light data; SHAP analysis; Turpan–Hami region

check for updates

Academic Editor: Yuji Murayama

Received: 28 November 2024
Revised: 28 December 2024
Accepted: 30 December 2024
Published: 11 January 2025

Citation: Zhang, G.; Zhang, L.; Chen, Y.; Chen, M.; Tian, J.; Wu, Y. Application of Nighttime Light Data Simulation Based on Multi-Indicator System and Machine Learning Model in Predicting Potentially Suitable Economic Development Areas: A Case Study of the Turpan–Hami Region. *Remote Sens.* **2025**, *17*, 240. https://doi.org/10.3390/rs17020240

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

Land serves as the primary medium for human production and living [1]. In recent years, China's social and economic development has been rapid [2,3], and the pace of urbanization has continuously accelerated [4]. Due to the limited urban land resources, unplanned and improperly laid-out urban construction has placed pressure on urban ecosystems and inevitably affected the natural environment [5,6]. For instance, issues such as localized congestion due to saturated urban traffic networks, voltage instability and intermittent power outages due to electrical overload in certain areas, and air pollution in

specific urban districts have emerged [7,8]. These challenges also pose significant obstacles for urban management. Therefore, against the backdrop of the new urban development trends and the strategic needs of the nation, rational urban development planning through scientific methods is essential to alleviate the pressures on urban production and living environments caused by the current imbalance in urban land resource supply and demand. It is also necessary to improve land use efficiency and ensure stable and coordinated development within urban areas in the future [9]. Identifying and assessing potentially suitable zones for economic development has become a key pathway toward achieving sustainable urban development.

Research on identifying and assessing potentially suitable zones for economic development has become increasingly integrated and refined, gradually centering around regional characteristics. First, in response to these characteristics, researchers have placed greater emphasis on developing differentiated development plans tailored to each area's specific needs [10]. Considering the significant differences among cities in resources, climate, and economy, the current division of suitable areas requires not only a detailed analysis of regional characteristics but also a localized approach that integrates the natural and social conditions of each area [11]. In terms of factors considered in evaluations, existing studies have gradually expanded from single-factor assessments to multidimensional, multi-level indicator systems [12,13]. For example, social factors such as population density, industry distribution, and transportation accessibility can provide valuable references for urban development, while environmental factors such as ecosystem stability ensure that urban construction does not disrupt the surrounding ecological balance.

Machine learning models have been widely applied in the study of urban development and economic domains in recent years, particularly excelling in predicting urban growth, analyzing land use changes, and studying socioeconomic inequalities. These models effectively capture complex spatial and temporal dynamics, uncovering nonlinear relationships between land development, infrastructure distribution, and socioeconomic activities [14,15]. By analyzing urban expansion trends and socioeconomic dynamics, such studies not only provide scientific support for land use planning and policy-making but also help quantify the impact of urban function distribution on resource allocation and inequality [16]. Despite their clear advantages in predictive accuracy and handling complex relationships, challenges remain, including high computational demands and dependency on data quality. Overall, machine learning models serve as powerful tools for addressing complex urban issues and supporting regional policy-making, while highlighting the importance of improving model interpretability and integrating multi-source data [17,18].

Meanwhile, the application of remote sensing technology has further enhanced the capability for dynamic monitoring of land use [19]. By regularly acquiring large-scale remote sensing images, researchers can conduct long-term monitoring of land use and ecological changes. This technological approach not only meets the need for multi-source data in assessment methods but also promotes the development of a range of evaluation techniques based on remote sensing products [20]. In recent years, the application scope of remote sensing products has broadened beyond traditional land cover and vegetation assessments to include monitoring of urban heat island effects, pollution distribution, and ecosystem health [21].

With the in-depth application of various remote sensing indices, more specialized remote sensing methods focused on urban development are gradually being adopted. For instance, combining urban heat island monitoring with the Normalized Difference Vegetation Index (NDVI) allows researchers to better understand the evolution of urban thermal environments and propose recommendations for green infrastructure planning [22]. Additionally, urban density and infrastructure distribution maps generated from remote sensing

data provide reliable foundations for the delineation of suitable areas. These methods, developed through the application of remote sensing products, offer effective support for ecological protection and the rational allocation of resources in rapidly expanding urban areas [23,24].

As an important tool for urban monitoring, nighttime light remote sensing products have garnered increasing attention in recent years. Nighttime remote sensing primarily captures data on nocturnal lighting, directly reflecting the intensity and distribution of illumination across urban areas and quantifying elements such as economic activity, population density, and social vitality [25]. This unique advantage makes nighttime light data a valuable tool for assessing urban development levels, providing an intuitive representation of the concentration of nighttime economic activities and regional disparities [26]. Nighttime light remote sensing products allow for better detection of urban economic vitality and social dynamics at night, filling gaps that other remote sensing products may not effectively reveal [27].

In previous assessments of potentially suitable areas for urban development, the application of nighttime light remote sensing products has been relatively limited. However, the conditions reflected by these data represent the cumulative impact of multiple factors that have long influenced urban development in traditional evaluations [28]. These elements align well with the spatial layout, resource allocation, and regional vitality characteristics required for assessing and predicting suitable areas for urban development [29]. Therefore, this study explores the integration of machine learning models with nighttime light remote sensing data to predict and assess potentially suitable areas for urban development [30,31]. By integrating socioeconomic and natural environmental data and employing machine learning models to fit nighttime light data, this approach expands the application of nighttime remote sensing data in urban economic development research, enabling the rapid prediction of suitable economic development zones under specific natural and social conditions.

Turpan and Hami, located along the Silk Road Economic Belt in northwestern China, serve as crucial nodes connecting Central Asia. Under the impetus of the Belt and Road Initiative and the Western Development Strategy, the Tuha region has received substantial policy support, fostering continuous growth in infrastructure development and economic activities, which in turn has accelerated regional urbanization [32]. This rapid urbanization has created an urgent need for effective land resource development and sustainable use [33]. To address the land resource pressure and ecological protection needs arising from rapid urbanization, this study focuses on the cities of Turpan and Hami, integrating multi-source data and machine learning techniques to develop an assessment framework for predicting potentially suitable areas for urban economic development [34].

RF, XGBoost, LightGBM, and CatBoost exhibit unique advantages in handling different data characteristics. Random Forest reduces the risk of overfitting by constructing multiple decision trees and combining their outputs through majority voting [35,36]. XGBoost enhances traditional gradient boosting algorithms by incorporating regularization mechanisms, parallel computing, and efficient split-search strategies, making it particularly effective for high-dimensional and large-scale datasets [37]. LightGBM further optimizes the efficiency of gradient boosting by employing histogram-based decision tree algorithms and leaf-wise splitting strategies, which significantly reduce training time and memory usage, while demonstrating a superior performance on imbalanced datasets [38]. CatBoost is specifically designed to handle categorical features, leveraging innovative target encoding and ordered boosting techniques to effectively prevent data leakage and achieve an outstanding predictive performance without extensive hyperparameter tuning. These

four models each possess distinctive strengths, making them well-suited to the diversity and complexity of the data in this study [39].

In addition, these machine learning models exhibit significant advantages over traditional methods when handling complex data [40]. Traditional statistical methods often struggle to fully capture the intricate nonlinear relationships among features and their combined impact on the results [41]. In contrast, these four models can leverage feature importance analysis and built-in algorithmic mechanisms to deeply explore the relationships between features and quantify their contributions to the predictive outcomes [42]. At the same time, these models utilize efficient algorithmic optimizations, such as parallel computing and histogram-based algorithms, which significantly enhance the efficiency of processing large-scale data, thereby overcoming the limitations of traditional methods in terms of computational speed and scalability [43].

This study first utilized nighttime light remote sensing data to quantify the urban economic development intensity in the Turpan–Hami region. Based on multi-source data, a dataset was constructed to fit the nighttime light data, which was then used to predict potentially suitable areas for economic development. Subsequently, multiple machine learning models, including RF, XGBoost, LightGBM, and CatBoost, were employed to predict suitable areas. The accuracy of these models was assessed to identify the optimal model for further analysis. Finally, the SHAP model was used to provide interpretative insights into the contributions of each factor. By integrating nighttime light remote sensing data with socioeconomic and natural environmental factors, this study aims to provide an efficient predictive framework for regional economic planning under resource and ecological constraints. This framework holds significant theoretical and practical value in the context of rapid urbanization. Furthermore, through policy recommendations based on the analysis results, this study provides valuable references for achieving sustainable development goals in the Turpan–Hami region and offers insights for economic suitability assessments in other areas along the Belt and Road Initiative.

2. Materials and Methods

2.1. Study Area

The Turpan–Hami region is located in the eastern part of Xinjiang (41°12′–45°05′N, 87°16′–96°23′E), as shown in Figure 1. It primarily comprises the administrative areas of Turpan and Hami, with Turpan consisting of Gaochang District, Shanshan County, and Toksun County, and Hami consisting of Yizhou District, Yiwu County, and Barkol Kazakh Autonomous County. Covering an area of approximately 210,000 km^2, the region's elevation ranges from −194 m to 4881 m. The land use types include cultivated land, forest land, grassland, water bodies, construction land, and unused land [44]. The region is rich in coal resources [45]. The entire Turpan–Hami region is classified as an arid area, characterized by a typical continental arid climate with large diurnal temperature variations, low precipitation, and high evaporation rates. The region faces severe water scarcity, significant land desertification, sparse vegetation, a low ecological environment index, and a low land reserve resource index, indicating high ecological pressure. The relationship between land reserve resource development and the ecological environment is quite strained. However, due to its unique geographical and economic characteristics, the Turpan–Hami region plays a crucial role in the cooperation and development between China and Central Asia [46].

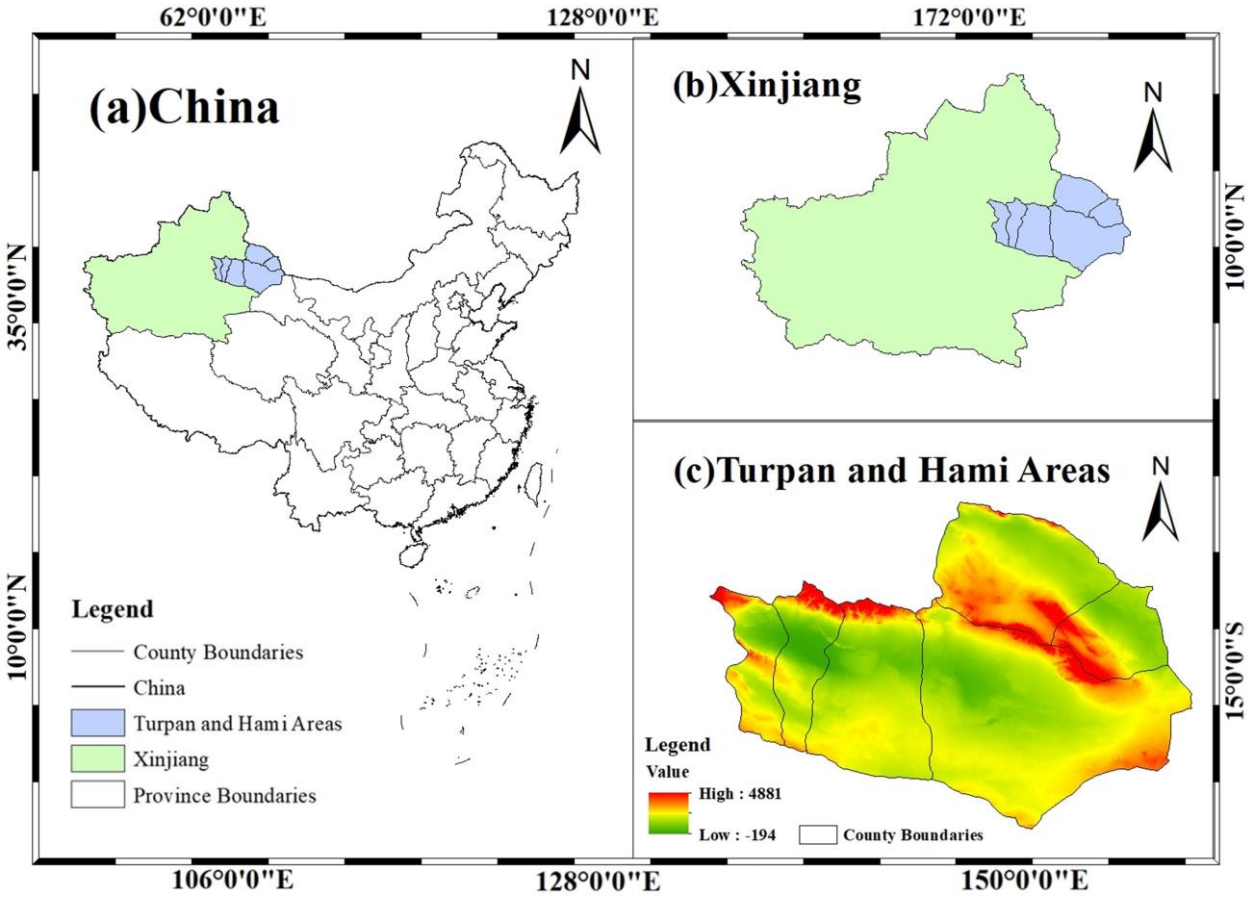


Figure 1. The area of study.

2.2. Construction of Evaluation Index System

The suitability of areas for urban economic development is influenced by multiple factors, including population, economic activity, transportation, topography, vegetation, and soil conditions. To accurately predict the spatial distribution of suitable areas for economic development, indicator selection should be based on these factors, incorporating a multidisciplinary perspective to achieve balanced growth between economic development and resource utilization. Drawing on existing research and considering the natural and social conditions of the Turpan–Hami region, this study constructed the evaluation indicator system shown in Table 1 to quantify urban economic development intensity.

Table 1. Evaluation index system.

Index	DATA Source	Justification	Reference
Elevation	NASA (SRTM dataset)	Elevation influences climatic conditions and the ease of resource access. Moderate elevation is generally more conducive to economic activities, facilitating construction and resource transportation, and thus aiding in the assessment of potentially suitable areas for economic development.	[47–51]
Slope	SRTM data calculations	Slope determines the difficulty of development; gentler slopes are suitable for large-scale construction and can reduce infrastructure development costs.	[47–50,52,53]

Table 1. *Cont.*

Index	DATA Source	Justification	Reference
Population Density	WorldPop dataset	Population distribution reflects the demand for regional economic development and the attractiveness of urban areas, providing insights into the urbanization process.	[51,52,54]
GDP	Resource and Environmental Science Data Platform 2017. DOI: 10.12078/2017121102 (2010, 2015, 2020)	GDP represents economic strength and helps assess the attractiveness of regional economic resources and development needs.	[51,55]
Road Density	gROADS dataset (2010) OpenStreetMap dataset (2015, 2020)	Road density determines resource mobility and regional economic connectivity, serving as a vital foundation for economic linkages.	[47,48,53–55]
Soil Texture	OpenLandMap Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m (Version v02) [56]	Soil texture determines land usability and carrying capacity, providing a reference for development activities.	[49,52]
NDVI	MOD13Q1 v006 dataset	NDVI reveals ecological sensitivity, ensuring the ecological sustainability of regional development.	/

2.3. Methods

This study utilizes nighttime light brightness values as variables to quantify the relative intensity of current urban development in the study area. A multi-criteria decision-making approach was employed to construct an evaluation indicator system for predicting potential urban development suitability zones. Leveraging the data processing capabilities of the Google Earth Engine (GEE) platform, training samples were initially created, and then, based on the constructed evaluation indicator system, nighttime light data and machine learning models were used to predict suitable areas. The main steps include the following: (1) creating training data for suitability assessment using the GEE platform; (2) predicting potentially suitable zones and assessing suitability intensity based on four machine learning models; and (3) analyzing the distribution of suitable zones according to model evaluation results and providing relevant policy recommendations. The technical route is shown in Figure 2.

2.3.1. Data Acquisition and Preprocessing

(1) Method for generating sample variable values based on nighttime light data

The "NPP-VIIRS-LIKE" nighttime light data product with a 500-m resolution from the AI Earth big data platform was used. These data are derived from two types of nighttime light remote sensing data (DMSP-OLS and NPP-VIIRS) through a cross-sensor correction method based on autoencoders. This process results in data that maintain a quality similar to that of NPP-VIIRS nighttime light data, overcoming the limitations of non-comparability between different nighttime light datasets and extending the usability of nighttime light data [57]. In this study, data from the year 2020 were selected. Within the study area, raster cells with certain brightness values from nighttime light data were randomly chosen as sample point coordinates and variable values. To achieve this, we used the natural breaks method in ArcGIS for auxiliary analysis. For the nighttime light data in the study area, there was a significant difference in the number of raster cells with noticeable brightness values

compared to those without. This method effectively distinguishes and extracts regions composed of raster cells with significant brightness differences, as shown in Figure 3a.

(2) Dataset Creation

First, we obtained the preprocessed data products on GEE based on the selected construction land suitability evaluation indicators. Among the seven constructed indicators, the road density was derived from vector road data through line density analysis to obtain raster data. Due to the varying resolutions, we resampled all data to a 500-m resolution during preprocessing to improve the evaluation accuracy. We then extracted the corresponding band values of the data products to construct the band features of the training samples, integrating them to create a new raster image with multi-band features within the study area. Finally, 3000 sample points were randomly generated on GEE to ensure uniform distribution within the extracted area, as shown in Figure 3b. The values of multiple bands at the corresponding locations were extracted from the raster images based on the coordinates of each sample point, resulting in band features for each sample point. Additionally, the resulting sample data were randomly split, with 70% of the sample points used to train the machine learning model and 30% used to validate the model's performance.

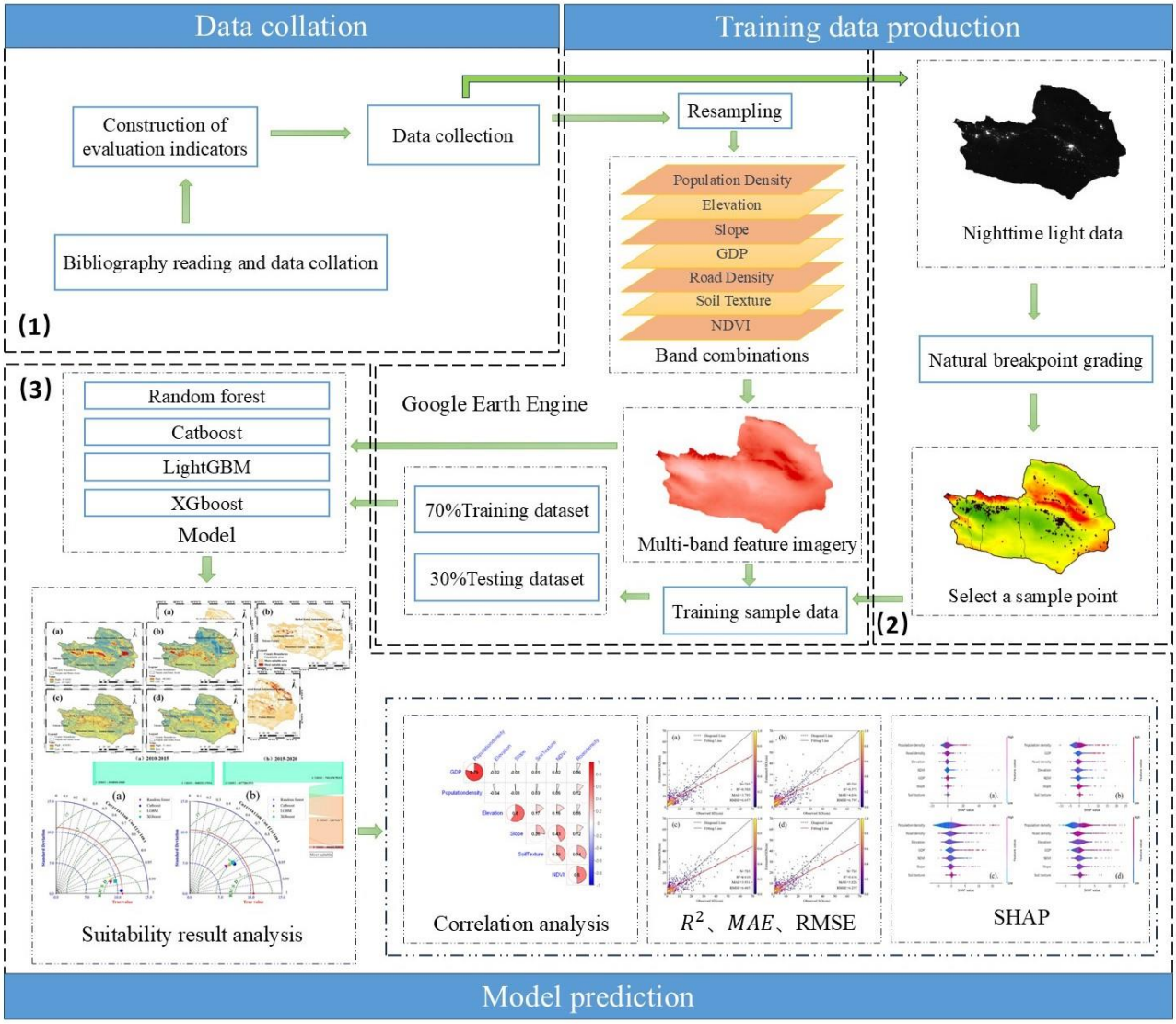


Figure 2. Technology roadmap.

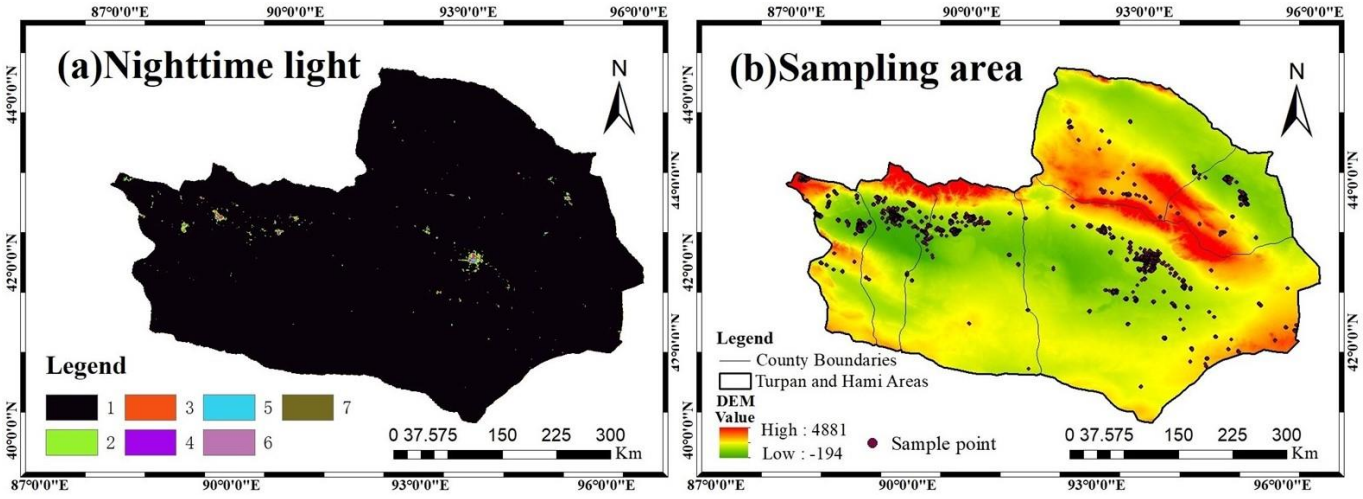


Figure 3. (**a**) Grade distribution map of nighttime light data in the study area; (**b**) random sampling point distribution.

2.3.2. Evaluation Method for Construction Land Suitability Intensity

In this study, four machine learning models—RF, XGBoost, LightGBM, and CatBoost—were comprehensively applied to predict potentially suitable areas for urban development. These models demonstrate significant advantages in handling large-scale, multi-dimensional data, providing efficient and accurate support for suitability prediction. By integrating multiple decision trees and gradient optimization algorithms, these models enhance prediction accuracy and stability while accommodating data diversity. This approach establishes a scientific basis for identifying potential suitable areas under complex data conditions, as detailed below.

(1) RF

Random Forest (RF) is a robust machine learning algorithm that is insensitive to overfitting noise. It consists of multiple decision trees, where each node represents a judgment condition for a feature variable, and each leaf node represents a category or a value depending on the application scenario of the model [58]. The model reduces variance by constructing each decision tree through random selection of samples and features. Each decision tree is trained based on different samples and features, resulting in some variability among the trees. In this method, the training data are divided into several subsets and trained independently through bootstrapping. The final prediction of the Random Forest is obtained by voting or averaging the predictions of all individual trees [59]. The formula is shown in Equation (1):

$$y(x) = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (1)$$

where y represents the output result, $T_i(x)$ denotes the i-th training set, and N is the number of independent training sets.

(2) XGBoost

The XGBoost model is a gradient boosting algorithm framework widely used in the field of machine learning. Its core idea is to iteratively train multiple weak learners and combine them to achieve strong predictive power. The algorithm enhances basic classifiers into robust classifiers, with accuracy stemming from multiple decision trees. It improves and optimizes the original gradient boosting algorithm, featuring efficiency, flexibility, and scalability. Additionally, XGBoost employs a block structure for parallel learning, pre-sorting data and storing it in a compressed column format [60,61].

(3) LightGBM

LightGBM employs a gradient boosting algorithm to iteratively optimize the loss function and fit the data continuously. In each iteration, the residuals between the predicted values of the new model and the actual values are used to train the next tree. The final result is obtained by summing the predictions of multiple trees [62]. The objective of LightGBM regression is to minimize the squared loss, which is the square of the difference between the actual values and the predicted values. The loss function can be expressed as Equation (2):

$$L(y, F) = \frac{1}{2}\sum_{i=1}^{n}(y_i - F(x_i))^2 \quad (2)$$

where y_i is the actual value of the i-th sample, and $F(x_i)$ is the predicted value of the i-th sample by the model. In each iteration, the model needs to update its parameters to reduce the loss. Gradient boosting uses the gradient descent method to solve for the negative gradient of the loss function and then updates the model parameters. For squared loss, the update rule can be expressed as Equation (3):

$$F^{(t)}(x_i) = F^{(t-1)} + \eta \cdot h_t(x_i) \quad (3)$$

where $F^{(t)}(x_i)$ is the predicted value of sample x_i after the t-th iteration, η is the learning rate, and $h_t(x_i)$ is the negative gradient of sample x_i for the t-th tree.

(4) CatBoost

CatBoost is another gradient boosting algorithm that uses symmetrical trees as base learners and employs a novel method to modify the gradient estimation in the classical algorithm, known as ordered boosting. This method can overcome prediction shifts caused by gradient bias, further enhancing the model's generalization ability. Additionally, the algorithm handles categorical features well by adding prior distribution values to the original Greedy Target-based statistics method. This effectively reduces the impact of noise and low-frequency categorical data on data distribution [63]. If a set of sample sequences is ordered as $\theta = \{\sigma_1, \ldots, \sigma_n\}$, then in the optimized formula, at the k-th iteration, the feature $x_{\sigma_{p,k}}$ of the p-th sample is as follows:

$$x_{\sigma_{p,k}} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] \cdot Y_{\sigma_j} + \beta \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + \beta} \quad (4)$$

where P is the added prior distribution term, and β is the weight of the prior value.

2.3.3. Model Evaluation

In the evaluation of model performance, multiple assessment metrics were utilized in this study to verify the accuracy and reliability of the prediction results. For the four machine learning models, we employed three commonly used metrics—R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE)—to assess model fit, prediction accuracy, and error magnitude, facilitating a quantitative analysis of each model's performance in predicting potentially suitable areas for urban development. Furthermore, to interpret the impact of each feature on the prediction outcomes, this study introduced Shapley Additive Explanation (SHAP) values to quantify and visualize the contribution of individual features to the model predictions, thereby revealing the driving factors behind model decisions. Details are as follows:

(1) R^2, MAE, RMSE

To ensure the accuracy and reliability of the results in the construction land suitability prediction task, we evaluated the performance of the four machine learning models constructed in this study. During the evaluation process, we used 30% of the data to validate the trained models. Specifically, we input the sample points from the validation data into the models and compared the predicted values with the actual values in the samples. We employed three common evaluation metrics, namely R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), to assess the models' fit, predictive accuracy, and precision. The relevant formulas are shown in Equations (5)–(7):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

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

where SS_{res} is the sum of squared residuals, representing the total difference between the model's predicted values and the actual values, and SS_{tot} is the total sum of squares, representing the total difference between the actual values and the mean of the actual values. These formulas provide us with quantitative metrics to measure model performance, aiding in the accurate evaluation of the four models' performance in the construction land suitability prediction task.

(2) SHAP

This study uses the SHAP model to explain the contribution of each feature to the model's predictions. Based on the concept of Shapley values from cooperative game theory, SHAP is widely used to interpret the prediction results of machine learning models. The main objective of the SHAP model is to provide an explanation of the model's prediction results, revealing the reasons behind the model's decisions. By calculating each feature's contribution to the model's predictions, the SHAP model can be used to quantify the impact of each feature on the final prediction result [64].

3. Results

3.1. Single Factor Index Results

In this study, we generated thematic maps for the research area based on the constructed indicators, as shown in Figure 4. The spatial characteristics of each dataset within the study area can be observed. The GDP and population density are primarily concentrated in the eastern and western urban areas of the Turpan–Hami region, indicating the high concentration of economic activity in these cities. This suggests that a small number of areas account for the majority of the economic output, reflecting the uneven population distribution in the Turpan–Hami region. Furthermore, high-altitude areas are mainly located in the central part of Hami and the northern part of Turpan, while the difference in NDVI values between the eastern and western areas is also quite significant. Additionally, the regional variability in soil texture is more pronounced compared to other characteristic indicators.

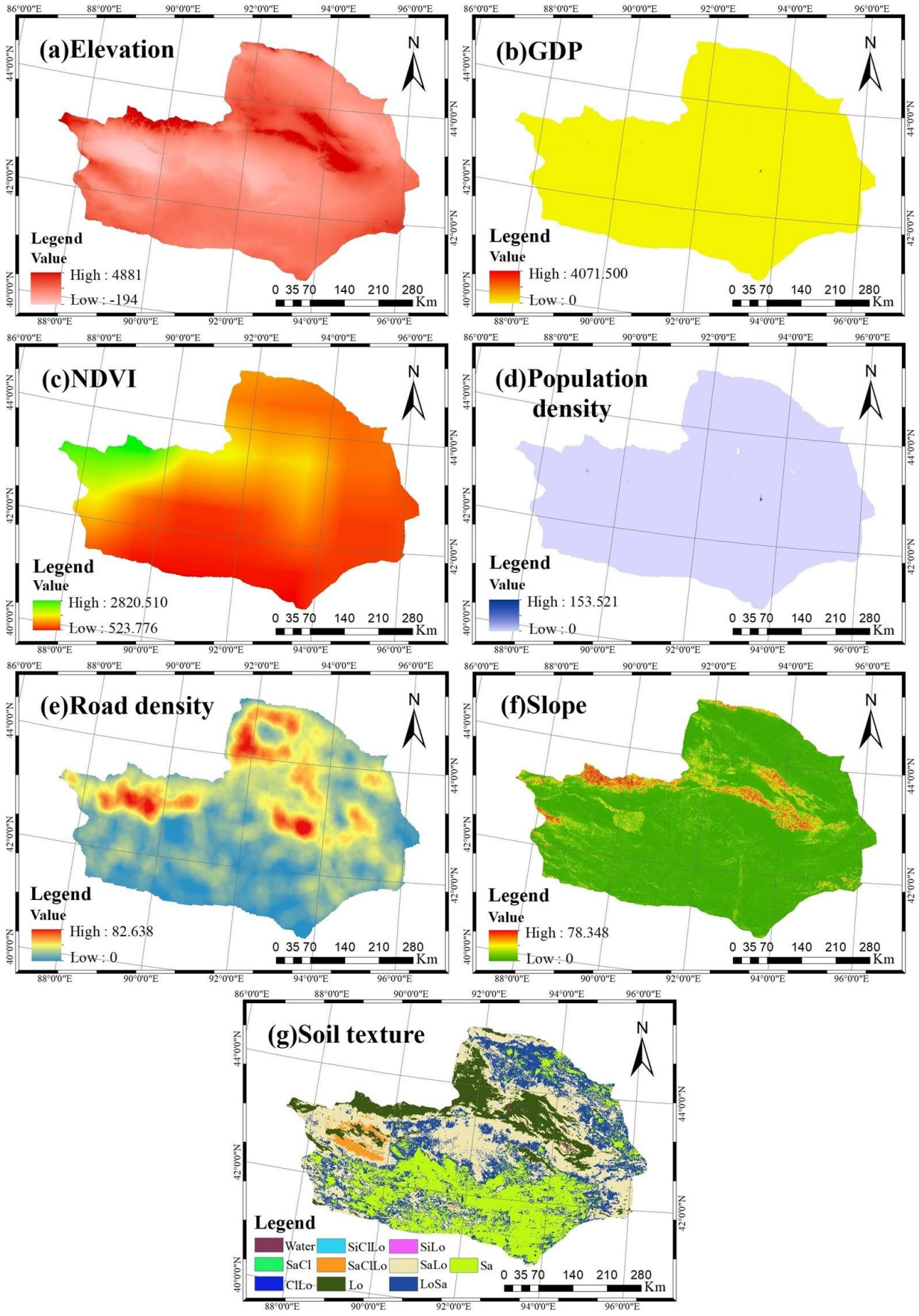


Figure 4. Thematic map of evaluation index data in the study area. (**a**) Elevation; (**b**) GDP; (**c**) NDVI; (**d**) population density; (**e**) road density; (**f**) slope; (**g**) soil texture.

3.2. Model Evaluation Results

When applying the Bayesian optimization algorithm, we set the number of parameter combinations to 100, recording the loss values for each of the four models with each combination. When the loss values stabilized, the optimal parameter combination at that point was output, as shown in Table 2.

Table 2. Four optimal combinations of model parameters.

Model	The Best Combination of Parameters
RF	max_depth = 19, max_features = 2, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 200
XGBoost	colsample_bytree = 0.736, gamma = 0.579, learning_rate = 0.061, max_depth = 16, min_child_weight = 1, n_estimators = 300, subsample = 0.507
LightGBM	colsample_bytree = 0.772, learning_rate = 0.062, max_depth = 11, min_child_samples = 1, n_estimators = 280, num_leaves = 23, reg_alpha = 0.277, reg_lambda = 0.578, subsample = 0.612
CatBoost	colsample_bylevel = 0.925, depth = 9, l2_leaf_reg = 4.687, learning_rate = 0.237, min_child_samples = 14, n_estimators = 300, subsample = 0.527

We plotted scatter plots for the training and validation sets as shown in Figures 5 and 6, respectively, and the evaluation results are summarized in Table 3. On the training set, XGBoost achieved the best R^2 value (0.992) and the lowest MAE (0.370) and RMSE (0.737). Both LightGBM and CatBoost outperformed RF with higher R^2 values (0.964 and 0.970, respectively) and lower MAE (1.031 and 0.903) and RMSE (1.562 and 1.433) values. On the validation set, CatBoost achieved the best R^2 (0.691) and the lowest RMSE (5.922), with a relatively low MAE (2.560). Additionally, XGBoost and LightGBM outperformed RF, with R^2 values of 0.668 and 0.674, MAE values of 2.524 and 2.594, and RMSE values of 6.142 and 6.082, respectively. Overall, the ensemble models (XGBoost, LightGBM, and CatBoost) demonstrated a significantly better predictive performance compared to RF. However, RF also showed an acceptable performance, with R^2 values greater than 0.6 on both the training and validation sets. This indicates that machine learning models have significant potential and advantages in predicting potentially suitable areas for urban economic development.

From the analysis of the Taylor diagram (Figure 7), the four models demonstrate a relatively strong performance on both the training and test sets. CatBoost performed the best on the training set, achieving the highest correlation coefficient, standard deviation closely matching the true values, and the lowest root mean square deviation (RMSD), indicating excellent fitting capability. On the test set, CatBoost, LGBM, and XGBoost all exhibited good generalization abilities, with correlation coefficients and standard deviations close to the true values, as well as low RMSD, resulting in an overall balanced performance. In contrast, the overall performance of Random Forest was relatively modest, showing reasonable fitting ability on the training set but a significant drop in performance on the test set, with lower correlation coefficients and standard deviations compared to the other models, indicating weaker generalization. Overall, CatBoost demonstrated the strongest fitting capability on the training set, while CatBoost, LGBM, and XGBoost exhibited similarly strong generalization abilities with minimal differences.

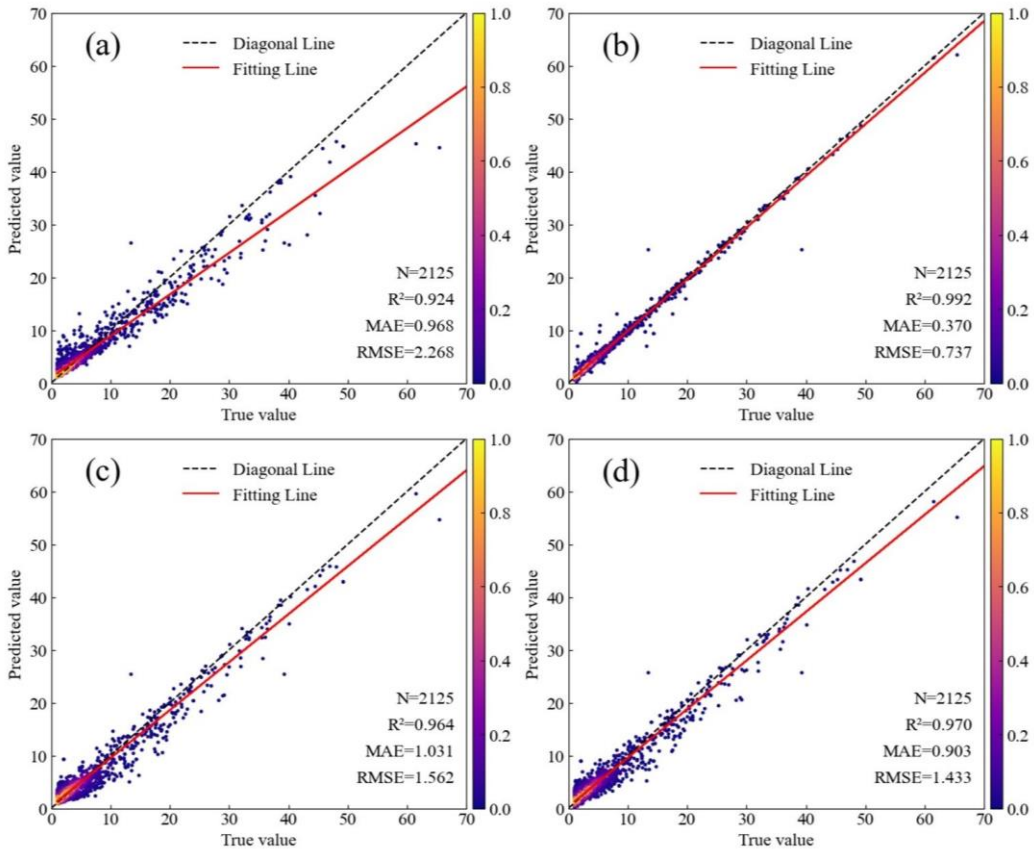


Figure 5. Scatter plots of the four models for the training set. (**a**) RF; (**b**) XGBoost; (**c**) LightGBM; (**d**) CatBoost.

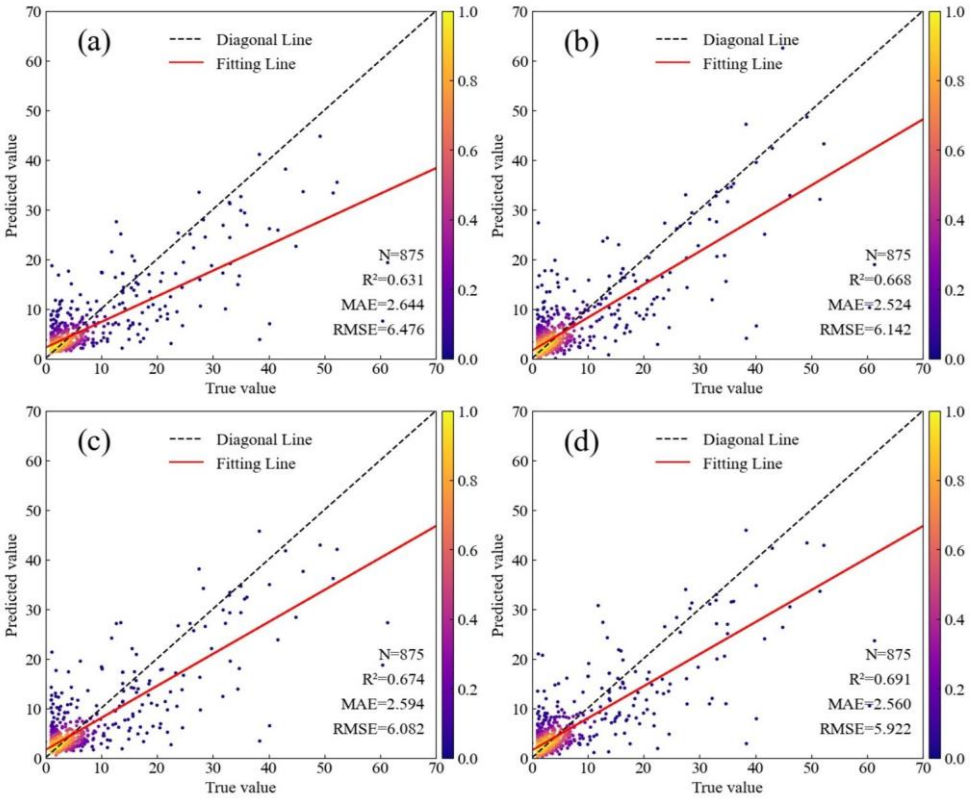


Figure 6. Scatter plots of the four models for the validation set. (**a**) RF; (**b**) XGBoost; (**c**) LightGBM; (**d**) CatBoost.

Table 3. Statistics for model evaluation.

	Training Dataset				Validation Dataset			
	RF	**XGBoost**	**LightGBM**	**CatBoost**	**RF**	**XGBoost**	**LightGBM**	**CatBoost**
R^2	0.924	0.992	0.964	0.970	0.631	0.668	0.674	0.691
MAE	0.968	0.370	1.031	0.903	2.644	2.524	2.594	2.560
RMSE	2.268	0.737	1.562	1.433	6.476	6.142	6.082	5.922

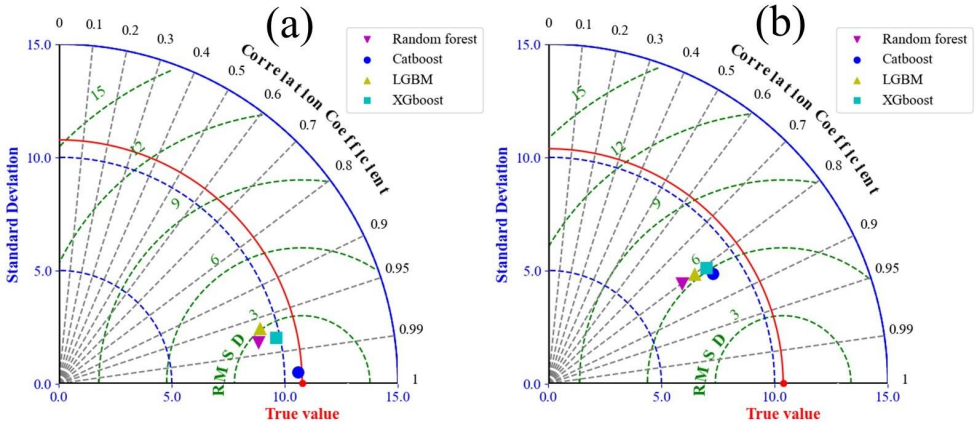


Figure 7. Taylor diagram. (**a**) Training set; (**b**) validation Set.

3.3. Prediction Results for Potentially Suitable Areas for Urban Economic Development

By training and predicting with four models, we obtained four sets of results and produced thematic maps of potentially suitable areas for urban economic development, as shown in Figure 8, where the relative values indicate the level of suitability. The evaluation results show that the CatBoost model (Figure 8d) demonstrated the best predictive performance and generalization ability. Overall, there are significant regional differences in the potentially suitable areas for urban economic development in the Turpan–Hami region, as shown in Figure 6d. The central, eastern, and some northern areas show higher suitability due to their favorable environmental conditions that are suitable for urban development. Conversely, the western and some north-central areas exhibit relatively lower suitability, likely influenced by topographical constraints and infrastructure limitations. Specifically, the areas with the highest suitability are mainly concentrated in and around existing urban areas, particularly in the eastern part of Barkol County and the northern part of Yiwu County. These areas, being close to existing infrastructure and transportation networks, possess the potential for further development. Considering the local development characteristics and resource conditions, moderate development could be undertaken in these high-suitability areas in the future, including the construction of residential areas, factories, and wind power facilities utilizing renewable resources such as wind energy.

3.4. Spatiotemporal Characteristics and Dynamic Transition Analysis of Potentially Suitable Areas for Urban Economic Development in the Turpan–Hami Region

In this study, the CatBoost model was employed to predict potential urban economic development areas at three historical time points: 2010, 2015, and 2020. Subsequently, the results were classified into three suitability levels—"unsuitable", "moderately suitable", and "most suitable"—using the natural breaks method and were visualized through thematic and dynamic transition maps (Figures 9 and 10). These results clearly reveal the distribution and evolution of urban economic development suitability levels in the Turpan–Hami region over the past decade, reflecting the characteristics and potential of regional urbanization.

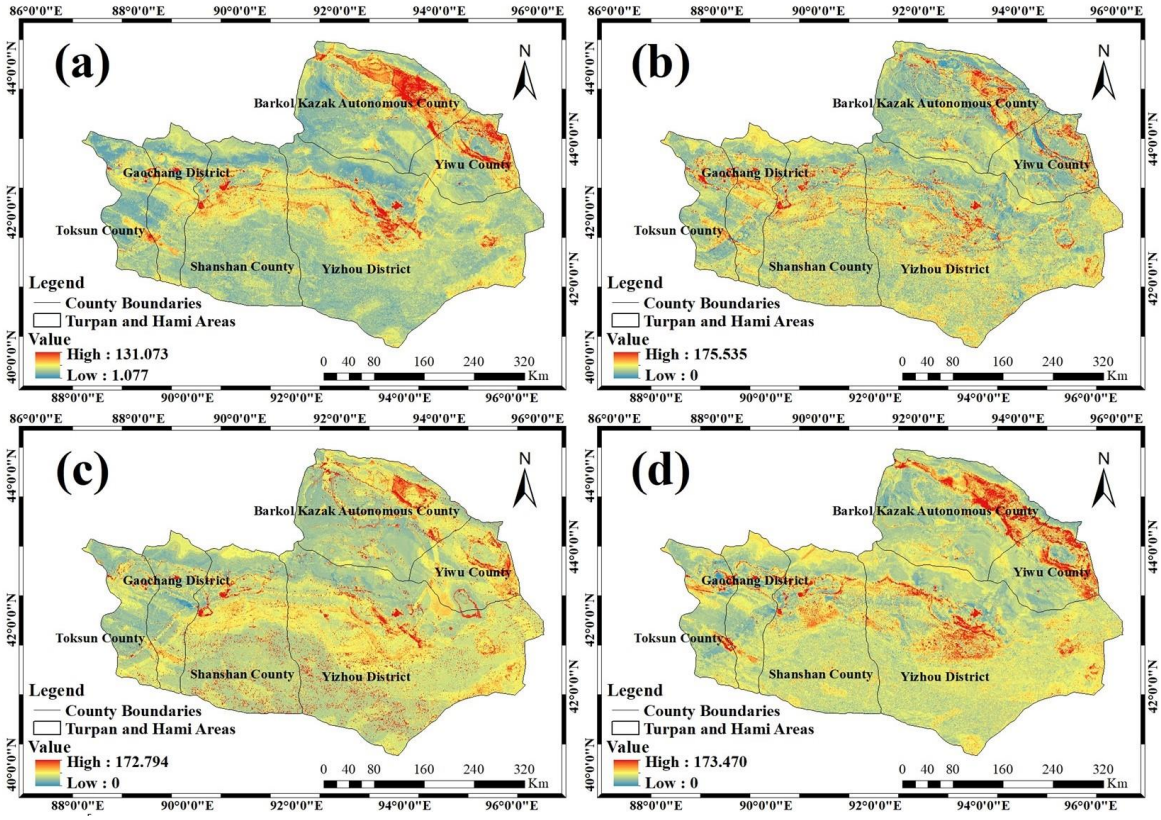


Figure 8. Prediction results of potentially suitable areas for economic development. (**a**) RF; (**b**) XGBoost; (**c**) LightGBM; (**d**) CatBoost.

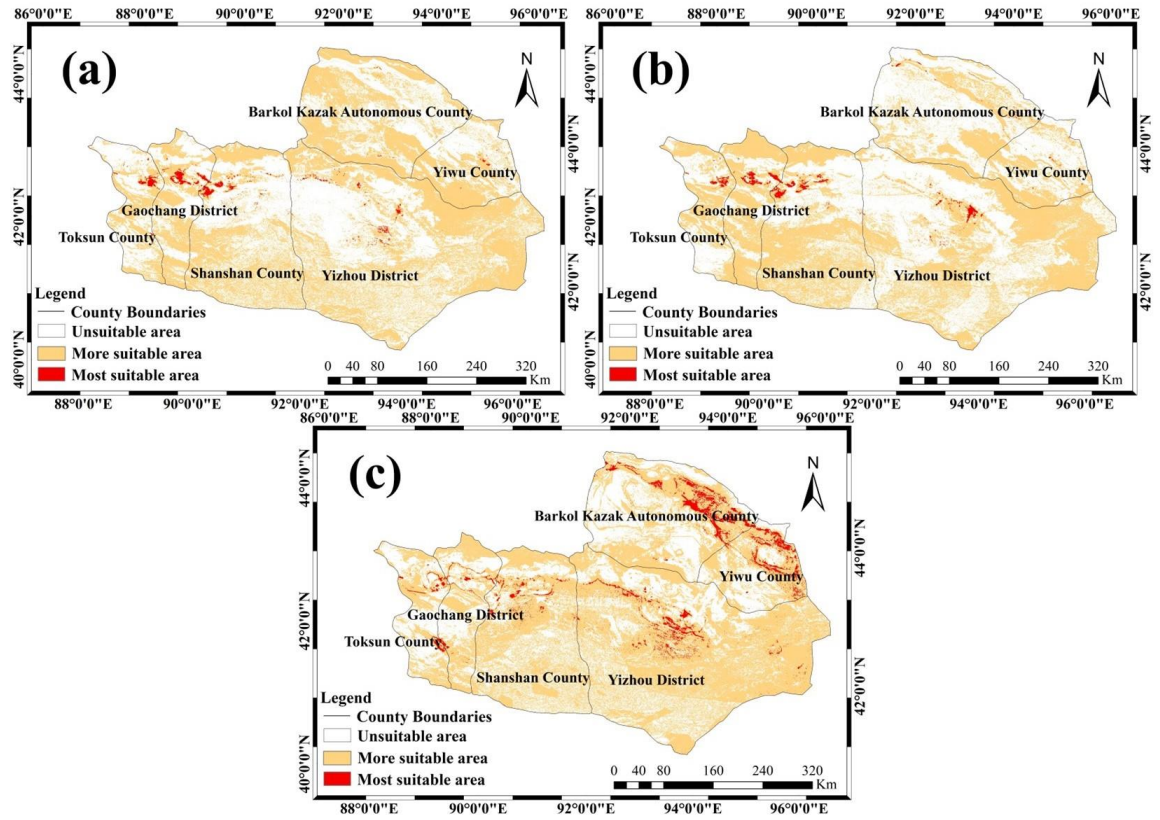


Figure 9. Potentially suitable areas for urban economic development in the Tuha area: (**a**) 2010; (**b**) 2015; (**c**) 2020.

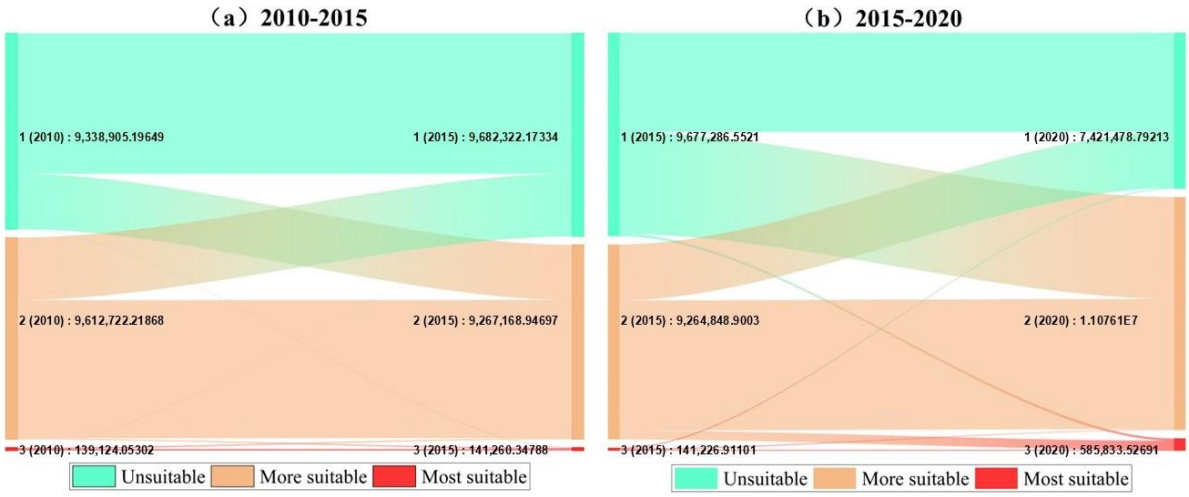


Figure 10. Map of dynamic transformations map between the different grades of suitability in 2010–2020.

Figure 9 shows that the high-suitability areas for urban economic development in the Turpan–Hami region are primarily concentrated around existing built-up zones, particularly in 2010 and 2015, where these high-suitability areas were closely adjacent to cities and built-up areas. This trend suggests that early urban economic development was mainly focused within and around established urban areas, with limited expansion due to topographical and resource constraints. However, by 2020, high-suitability areas expanded significantly, extending not only around existing built-up zones but also northward into the Turpan–Hami region. This expansion reflects the further release of urban economic development potential in the region, particularly driven by the Belt and Road Initiative.

Figure 10 presents the dynamic transitions in urban economic development suitability levels from 2010 to 2020, analyzing the proportion of mutual conversions between different levels. From 2010 to 2015, transitions were frequent between unsuitable areas and moderately suitable areas, indicating a significant shift from unsuitable to moderately suitable areas, which aligns with the gradual improvement of urban infrastructure. From 2015 to 2020, moderately suitable areas further transitioned into most suitable areas, suggesting an accelerated trend of suitability improvement, with the region's development potential gradually being realized. However, direct transitions from unsuitable to most suitable areas remained relatively rare, highlighting the constraints imposed by resources, the environment, and infrastructure, and indicating that achieving a leap from 'unsuitable' to 'most suitable' remains a significant challenge.

4. Discussion

4.1. Relative Importance of Evaluation Indicators

This study employed the SHAP model to interpret the importance and contributions of different variables (including both positive and negative contributions) in the machine learning prediction process. Figure 11 presents the SHAP values for the four models, ranked by their mean absolute SHAP values from high to low. We found that "population density" and "GDP" were the key features across all models, indicating that these two indicators played a dominant role in predicting economic development potential. "Population density" directly influences labor supply and market demand, while "GDP", as a comprehensive indicator of economic activity, serves as a critical driving force for regional development. In addition, "road density" showed a significant impact in the RF, LightGBM, and CatBoost models, highlighting the importance of transportation infrastructure in regional economic

development potential. This influence likely stems from its role in enhancing resource mobility and improving the efficiency of economic activities.

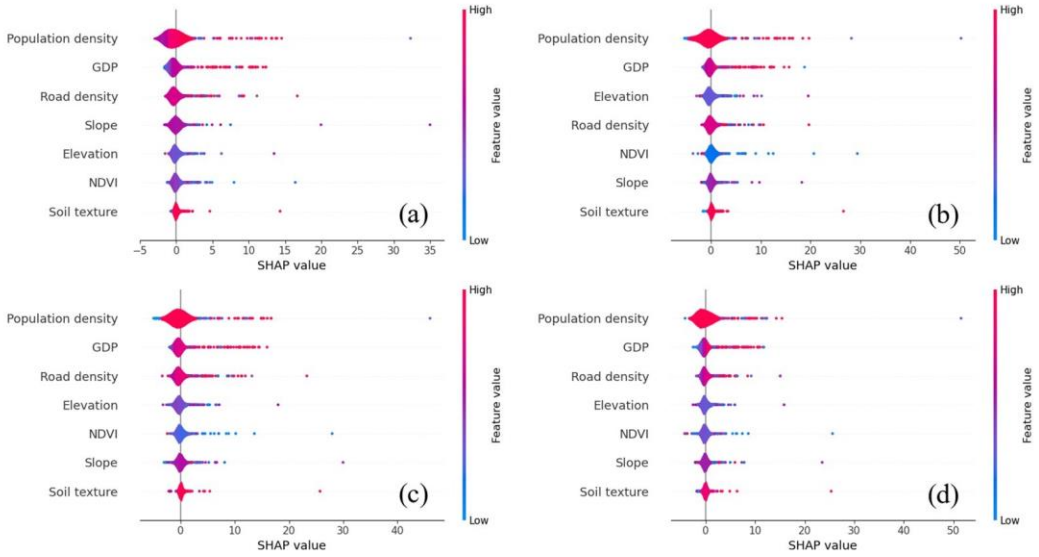


Figure 11. SHAP results for models. (**a**) RF; (**b**) XGBoost; (**c**) LightGBM; (**d**) CatBoost.

Conversely, the "soil texture" feature had the lowest importance across all four models, suggesting its limited direct impact on economic development potential. This result reflects that natural factors may carry less weight than socioeconomic factors in economic prediction models. Moreover, the SHAP value analysis further revealed the positive and negative contributions of different indicators to predictions. For instance, the SHAP values of "population density" and "GDP" indicated significant positive contributions, while the contribution of "road density" varied across models, possibly due to differences in how models assign weights to specific features. Overall, the analysis shows that socioeconomic factors (such as population density, GDP, and road density) predominantly drive predictions of economic development potential, while natural factors (such as soil texture) play a relatively minor role. This finding further emphasizes the central role of socioeconomic factors in influencing the suitability of construction land and predictions of economic development potential, providing critical insights for regional development policymaking.

4.2. Correlations of Evaluation Indicators

In the context of suitability evaluations, the construction of evaluation indicators is typically based on previous research, with dynamic adjustments made to the existing evaluation indicator system according to the social and natural conditions of the study area. This holds true whether the assessment framework is based on dynamic processes [50] or static processes [47,48]. To explore the correlation of each influencing factor, we conducted a correlation analysis of the constructed feature indicators, as shown in Figure 12. The indicators "GDP" and "population density", as well as "elevation" and "slope", exhibited strong and positive correlations compared to other features, with correlation coefficients of 0.79 and 0.6, respectively. In contrast, "GDP" had the weakest correlation with "slope" and "soil texture". This also indirectly reflects the link between regional population density and economy in the process of social development.

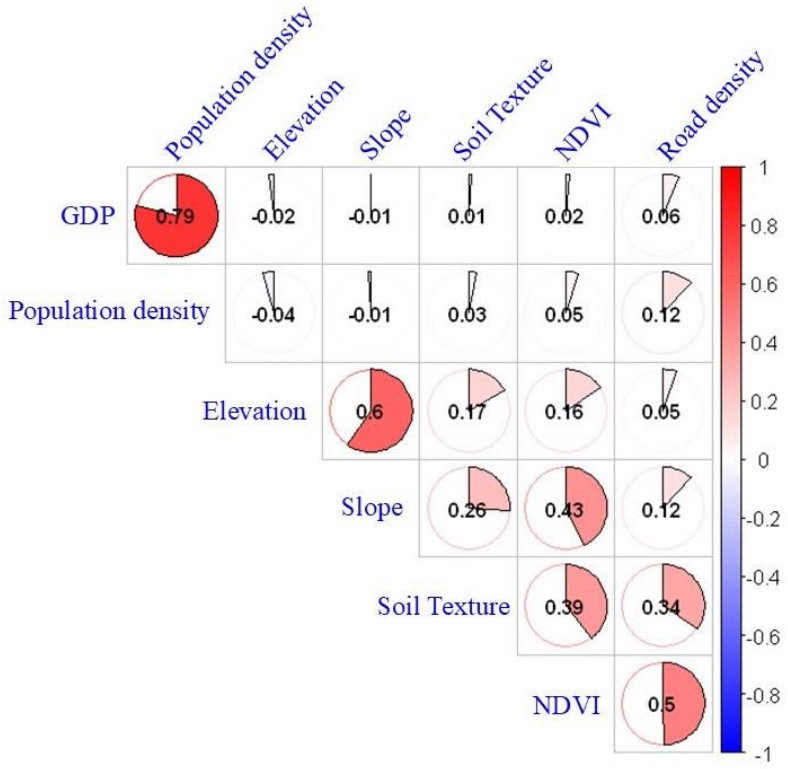


Figure 12. Evaluation index relevance results.

4.3. Performance Limitations, Model Analysis, and Challenges in Practical Applications

Although this study employed four machine learning models (RF, XGBoost, LightGBM, and CatBoost) to predict suitable areas for economic development and achieved good fitting performances on the training set, the suboptimal performance on the test set revealed some potential issues [65]. These challenges may be attributed to the imbalance of regional data, inconsistencies across multi-source datasets, and insufficient representativeness of the training samples. For example, significant variations in feature distributions (such as GDP and population density) across certain areas may limit the models' predictive accuracy in regions with low feature values [66].

Furthermore, each model demonstrated unique strengths and weaknesses: RF is robust and resistant to overfitting but struggles with efficiency in handling large-scale datasets and capturing extreme values [67]; XGBoost shows strong optimization capabilities and flexibility but is sensitive to class imbalance and has high parameter complexity [68,69]; LightGBM performs well in high-dimensional and large-scale data processing but is prone to overfitting in scenarios with small samples or limited feature data [15]; CatBoost excels in handling categorical features and generalization but demands significant computational resources and involves a complex parameter tuning process [70]. Additionally, these models may face challenges in practical applications, such as data quality issues (e.g., noise and missing values), as well as instability in cross-regional or cross-temporal predictions, while their interpretability remains a key concern in real-world scenarios.

Future research could focus on improving data preprocessing techniques, optimizing the distribution of training samples, and combining multi-model ensemble strategies with automated parameter tuning techniques to enhance model performance and robustness. Incorporating more diverse external data (e.g., socioeconomic, climatic, or infrastructure-related data) could enrich the feature space and improve model adaptability. Furthermore, integrating more advanced deep learning models tailored to spatiotemporal data may further enhance the predictive accuracy [71]. Finally, conducting analyses at finer spatiotemporal scales could allow for more precise assessments of dynamic regional suitability,

providing a scientific basis for developing localized and actionable policy recommendations. These potential research directions will help address existing challenges while expanding the applicability of the models in predicting economic development suitability, ultimately supporting more effective and sustainable urban planning strategies.

4.4. Development Recommendations Based on Prediction Results

Based on the social context and urban development potential of the Turpan–Hami region, the following policy recommendations are proposed to promote sustainable development.

(1) Enhancing Infrastructure Development, Focusing on High-Potential Areas

As suitability levels increase, the government should prioritize infrastructure investment in the most suitable areas (e.g., the highly suitable northern regions), particularly in transportation hubs, logistics facilities, water and electricity supplies, and communication networks. This will enhance regional connectivity, attract investment, and support the growth of emerging high-suitability areas.

(2) Optimizing Land Use Planning with a Focus on High-Suitability Zones

Detailed land use planning should be conducted in the most suitable and moderately suitable areas to avoid disorderly expansion and ensure that development activities are concentrated in high-suitability zones. For example, the areas marked in red on the 2015 and 2020 maps should serve as core zones for urban and industrial development, while ecological protection zones should be preserved in areas of lower suitability.

(3) Differentiated Development of Low-Suitability Areas to Promote Balanced Regiona Growth

While low-suitability areas may not be ideal for large-scale industrial or urban development, they hold potential for ecological tourism, sustainable agriculture, and ecological restoration. For example, low-suitability areas can leverage their natural conditions and cultural features to develop eco-tourism projects, driving economic growth while preserving the environment. Furthermore, policy support can encourage small-scale sustainable agricultural practices, such as drought-resistant crops, to improve local livelihoods and enhance the economic resilience of these regions.

(4) Promoting Development of Industries Compatible with Natural Conditions

Based on the observed evolution of suitability, high-suitability areas in the northern and central regions can be prioritized for the development of renewable energy, cultural tourism, and modern agriculture to enhance regional economic vitality. Particularly for the emerging economic zones within high-suitability areas in 2020, policy support can be leveraged to accelerate the growth of regional specialized economies.

In conclusion, the sustainable development of the Turpan–Hami region can be gradually achieved through differentiated and targeted policies. By integrating the dynamic changes in suitability distribution, scientifically planning infrastructure development, land use, ecological protection, and the growth of specialized industries will help to fully leverage regional advantages [72]. Through sound policy guidance, the Turpan–Hami region can not only achieve higher economic suitability but also play a significant role in the Belt and Road Initiative, promoting high-quality regional economic growth [73].

5. Conclusions

Nighttime light data comprehensively reflect spatial layouts, resource allocation, and regional vitality, providing a novel method for rapidly predicting suitable areas for urban economic development [27,74]. This study integrated a multidimensional indicator system with four tree-based ensemble learning models to innovatively assess potentially suitable

areas and the intensity of suitability for urban economic development in the Turpan–Hami region. The results indicate that all models performed excellently in suitability prediction, with the CatBoost model demonstrating exceptional predictive capability and robustness. The evaluation revealed significant spatial variations in suitability across the Turpan–Hami region: higher suitability was observed in the central and eastern regions and parts of the northern regions due to favorable environmental and resource conditions, while lower suitability was noted in the western and some north-central regions, likely constrained by topographical challenges and inadequate infrastructure.

Using suitability predictions generated by the CatBoost model for 2010, 2015, and 2020, along with dynamic analysis of temporal changes, this study uncovered the expansion trends of suitable zones for urban economic development in the Turpan–Hami region. By 2020, high-suitability zones not only expanded around existing built-up areas but also extended significantly northward. This trend indicates that the region's urban development potential has been further unlocked over time, driven by the synergistic effects of regional economic growth, infrastructure improvements, and policy support. These findings provide a scientific basis for guiding land use allocation and resource management in the Turpan–Hami region amidst rapid urbanization, while aligning with sustainable development objectives in urban planning. Based on the characteristics of suitability expansion, the study offers quantitative insights for future land management and urban expansion, aiding policymakers in optimizing the distribution of construction land while ensuring ecological protection.

The research method proposed in this study, which integrates a multi-indicator system with machine learning models, demonstrates broad applicability and significant potential for expansion. Firstly, this method does not rely on region-specific data or characteristics, allowing for flexible adjustment of the indicator system based on the specific conditions of the study area, thereby accommodating diverse natural, social, and economic contexts. For example, by replacing or adding indicators relevant to local economic structures, ecological conditions, or infrastructure characteristics, this method can be applied to evaluate the development suitability of other types of cities, such as coastal cities, resource-dependent cities, or ecologically sensitive areas. Moreover, the high flexibility and scalability of machine learning models enable their application across various spatial scales (from urban to regional to national levels) and at different stages of development (e.g., during rapid urbanization or in mature urban renewal phases). This method can also be enhanced by incorporating higher spatiotemporal resolution remote sensing data, further improving its adaptability and predictive accuracy. Therefore, the proposed method is not only applicable to urban development assessments in the arid regions of northwestern China but also provides scientific decision-making support for other regions, particularly in the context of the rapidly evolving global urbanization process.

Author Contributions: Conceptualization, G.Z., L.Z., Y.C. and Y.W.; methodology, G.Z., L.Z., Y.C. and Y.W.; software, G.Z. and M.C.; validation, G.Z., L.Z., Y.C. and Y.W.; formal analysis, G.Z., L.Z., Y.C. and Y.W.; investigation, G.Z. and J.T.; resources, Y.W.; data curation, G.Z. and Y.W.; writing—original draft preparation, G.Z.; writing—review and editing, G.Z., L.Z., Y.C. and Y.W.; visualization, G.Z.; supervision, J.T.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Third Xinjiang Scientific Expedition Program (Grant No. 2022xjkk1100).

Data Availability Statement: The research data used in this study are available from the corresponding authors on request.

Acknowledgments: We would like to express our gratitude to the following organizations for providing the remote sensing products used in this study: the National Geomatics Center of China; the Resources and Environmental Sciences Data Platform; and the teams of Yu Bailang from East China Normal University, Chen Zuoqi from Fuzhou University, and Shi Kaifang from Southwest University. We also extend our thanks to the Google Earth Engine platform for facilitating various aspects of our analysis.

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

References

1. Lange, A.; Piorr, A.; Siebert, R.; Zasada, I. Spatial differentiation of farm diversification: How rural attractiveness and vicinity to cities determine farm households' response to the CAP. *Land Use Policy* **2013**, *31*, 136–144. [CrossRef]
2. Zhu, X. Understanding China's Growth: Past, Present, and Future. *J. Econ. Perspect.* **2012**, *26*, 103–124. [CrossRef]
3. Qi, Y.; Yang, Y.; Jin, F. China's economic development stage and its spatio-temporal evolution: A prefectural-level analysis. *J. Geogr. Sci.* **2013**, *23*, 297–314. [CrossRef]
4. Chen, M.; Liu, W.; Tao, X. Evolution and assessment on China's urbanization 1960–2010: Under-urbanization or over-urbanization? *Habitat Int.* **2013**, *38*, 25–33. [CrossRef]
5. Bek, M.A.; Azmy, N.; Elkafrawy, S. The effect of unplanned growth of urban areas on heat island phenomena. *Ain Shams Eng. J.* **2018**, *9*, 3169–3177. [CrossRef]
6. Kaur, H.; Garg, P. Urban sustainability assessment tools: A review. *J. Clean. Prod.* **2019**, *210*, 146–158. [CrossRef]
7. Kessides, I.N. Chaos in power: Pakistan's electricity crisis. *Energy Policy* **2013**, *55*, 271–285. [CrossRef]
8. Lu, J.; Li, B.; Li, H.; Al-Barakani, A. Expansion of city scale, traffic modes, traffic congestion, and air pollution. *Cities* **2021**, *108*, 102974. [CrossRef]
9. Liu, R.; Zhang, K.; Zhang, Z.; Borthwick, A.G.L. Land-use suitability analysis for urban development in Beijing. *J. Environ. Manag.* **2014**, *145*, 170–179. [CrossRef] [PubMed]
10. Stimson, R.J.; Stough, R.R.; Roberts, B.H. *Regional Economic Development: Analysis and Planning Strategy*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
11. Ettlinger, N. The localization of development in comparative perspective. *Econ. Geogr.* **1994**, *70*, 144–166. [CrossRef]
12. Huang, I.B.; Keisler, J.; Linkov, I. Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. *Sci. Total Environ.* **2011**, *409*, 3578–3594. [CrossRef] [PubMed]
13. Kwak, Y.; Deal, B.; Heavisides, T. A large scale multi criteria suitability analysis for identifying solar development potential: A decision support approach for the state of Illinois, USA. *Renew. Energy* **2021**, *177*, 554–567. [CrossRef]
14. Tsagkis, P.; Bakogiannis, E.; Nikitas, A. Analysing urban growth using machine learning and open data: An artificial neural network modelled case study of five Greek cities. *Sustain. Cities Soc.* **2023**, *89*, 104337. [CrossRef]
15. Kim, Y.; Safikhani, A.; Tepe, E. Machine learning application to spatio-temporal modeling of urban growth. *Comput. Environ. Urban Syst.* **2022**, *94*, 101801. [CrossRef]
16. Nyamekye, C.; Kwofie, S.; Ghansah, B.; Agyapong, E.; Boamah, L.A. Assessing urban growth in Ghana using machine learning and intensity analysis: A case study of the New Juaben Municipality. *Land Use Policy* **2020**, *99*, 105057. [CrossRef]
17. Fan, C.; Xu, J.; Natarajan, B.Y.; Mostafavi, A. Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. *Comput.—Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2013–2029. [CrossRef]
18. Jianan, Z.; Tiejian, H.; Qun, Y. Based on Gated Recurrent network analysis of advanced manufacturing cluster and unified large market to promote regional economic development. *Comput. Ind. Eng.* **2024**, *197*, 110575. [CrossRef]
19. Kadhim, N.; Mourshed, M.; Bray, M. Advances in remote sensing applications for urban sustainability. *Euro-Mediterr. J. Environ. Integr.* **2016**, *1*, 7. [CrossRef]
20. Wu, Q.; Li, H.-q.; Wang, R.-s.; Paulussen, J.; He, Y.; Wang, M.; Wang, B.-h.; Wang, Z. Monitoring and predicting land use change in Beijing using remote sensing and GIS. *Landsc. Urban Plan.* **2006**, *78*, 322–333. [CrossRef]
21. Zhou, D.; Xiao, J.; Bonafoni, S.; Berger, C.; Deilami, K.; Zhou, Y.; Frolking, S.; Yao, R.; Qiao, Z.; Sobrino, J.A. Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives. *Remote Sens.* **2018**, *11*, 48. [CrossRef]
22. Chun, B.; Guldmann, J.-M. Spatial statistical analysis and simulation of the urban heat island in high-density central cities. *Landsc. Urban Plan.* **2014**, *125*, 76–88. [CrossRef]
23. Dutta, D.; Rahman, A.; Paul, S.; Kundu, A. Spatial and temporal trends of urban green spaces: An assessment using hyper-temporal NDVI datasets. *Geocarto Int.* **2022**, *37*, 7983–8003. [CrossRef]
24. Kuffer, M.; Pfeffer, K.; Persello, C. Special issue "remote-sensing-based urban planning indicators". *Remote Sens.* **2021**, *13*, 1264. [CrossRef]

25. Fragkias, M.; Lobo, J.; Seto, K.C. A comparison of nighttime lights data for urban energy research: Insights from scaling analysis in the US system of cities. *Environ. Plan. B Urban Anal. City Sci.* **2017**, *44*, 1077–1096. [CrossRef]
26. Li, D.; Zhao, X.; Li, X. Remote sensing of human beings–a perspective from nighttime light. *Geo-Spat. Inf. Sci.* **2016**, *19*, 69–79. [CrossRef]
27. Keola, S.; Andersson, M.; Hall, O. Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. *World Dev.* **2015**, *66*, 322–334. [CrossRef]
28. Bennett, M.M.; Smith, L.C. Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sens. Environ.* **2017**, *192*, 176–197. [CrossRef]
29. Chen, Z.; Wei, Y.; Shi, K.; Zhao, Z.; Wang, C.; Wu, B.; Qiu, B.; Yu, B. The potential of nighttime light remote sensing data to evaluate the development of digital economy: A case study of China at the city level. *Comput. Environ. Urban Syst.* **2022**, *92*, 101749. [CrossRef]
30. Yang, M.; Wang, S.; Zhou, Y.; Wang, L.; Zeng, C. A study of urbanization in China based on regionalization using DMSP/OLS night light imagery. In Proceedings of the 2009 Joint Urban Remote Sensing Event, Shanghai, China, 20–22 May 2009; pp. 1–7.
31. Lao, J.; Wang, C.; Wang, J.; Pan, F.; Xi, X.; Liang, L. Land use simulation of Guangzhou based on nighttime light data and planning policies. *Remote Sens.* **2020**, *12*, 1675. [CrossRef]
32. Liu, H.; Fang, C.; Miao, Y.; Ma, H.; Zhang, Q.; Zhou, Q. Spatio-temporal evolution of population and urbanization in the countries along the Belt and Road 1950–2050. *J. Geogr. Sci.* **2018**, *28*, 919–936. [CrossRef]
33. Koroso, N.H.; Zevenbergen, J.A. Urban land management under rapid urbanization: Exploring the link between urban land policies and urban land use efficiency in Ethiopia. *Cities* **2024**, *153*, 105269. [CrossRef]
34. Gómez, J.A.; Patiño, J.E.; Duque, J.C.; Passos, S. Spatiotemporal modeling of urban growth using machine learning. *Remote Sens.* **2019**, *12*, 109. [CrossRef]
35. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for big data: An interdisciplinary review. *J. Big Data* **2020**, *7*, 94. [CrossRef]
36. Demir, S.; Sahin, E.K. Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. *Acta Geotech.* **2023**, *18*, 3403–3419. [CrossRef]
37. Qiu, Y.; Zhou, J.; Khandelwal, M.; Yang, H.; Yang, P.; Li, C. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Eng. Comput.* **2022**, *38*, 4145–4162. [CrossRef]
38. Sun, X.; Liu, M.; Sima, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. *Financ. Res. Lett.* **2020**, *32*, 101084. [CrossRef]
39. Jabeur, S.B.; Gharib, C.; Mefteh-Wali, S.; Arfi, W.B. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technol. Forecast. Soc. Change* **2021**, *166*, 120658. [CrossRef]
40. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1–21. [CrossRef]
41. Song, M.; Chen, D. A comparison of three heuristic optimization algorithms for solving the multi-objective land allocation (MOLA) problem. *Ann. GIS* **2018**, *24*, 19–31. [CrossRef]
42. Ganaie, M.A.; Hu, M.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
43. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ. -Comput. Inf. Sci.* **2023**, *35*, 757–774. [CrossRef]
44. Liu, Z.; Wang, J.; Ding, J.; Xie, X. Analysis of spatial–temporal evolution trends and influential factors of desert-oasis thermal environment in typical arid zone: The case of Turpan–Hami region. *Ecol. Indic.* **2023**, *154*, 110747. [CrossRef]
45. Han, S.; Zhang, J.; Zhou, Y.; Bai, S.; Huang, L.; Wang, C.; Huang, W. Formation and accumulation of lower Jurassic tight gas sands field in Kekeya area of Tuha Basin, northwestern China. *J. Nat. Gas Sci. Eng.* **2016**, *29*, 101–109. [CrossRef]
46. Kou, Y.; Chen, S.; Zhou, K.; Qiu, Z.; He, J.; Shi, X.; Zhou, X.; Zhang, Q. Spatiotemporal Patterns and Coupling Coordination Analysis of Multiscale Social–Economic–Ecological Effects in Ecologically Vulnerable Areas Based on Multi-Source Data: A Case Study of the Tuha Region, Xinjiang Province. *Land* **2024**, *13*, 282. [CrossRef]
47. Ustaoglu, E.; Aydınoglu, A.C. Suitability evaluation of urban construction land in Pendik district of Istanbul, Turkey. *Land Use Policy* **2020**, *99*, 104783. [CrossRef]
48. Huang, R.; Nie, Y.; Duo, L.; Zhang, X.; Wu, Z.; Xiong, J. Construction land suitability assessment in rapid urbanizing cities for promoting the implementation of United Nations sustainable development goals: A case study of Nanchang, China. *Environ. Sci. Pollut. Res.* **2021**, *28*, 25650–25663. [CrossRef] [PubMed]
49. Akbari, M.; Neamatollahi, E.; Neamatollahi, P. Evaluating land suitability for spatial planning in arid regions of eastern Iran using fuzzy logic and multi-criteria analysis. *Ecol. Indic.* **2019**, *98*, 587–598. [CrossRef]
50. Peng, J.; Ma, J.; Du, Y.; Zhang, L.; Hu, X. Ecological suitability evaluation for mountainous area development based on conceptual model of landscape structure, function, and dynamics. *Ecol. Indic.* **2016**, *61*, 500–511. [CrossRef]

51. Kang, Z.; Wang, S.; Xu, L.; Yang, F.; Zhang, S. Suitability assessment of urban land use in Dalian, China using PNN and GIS. *Nat. Hazards* **2021**, *106*, 913–936. [CrossRef]
52. Bamrungkhul, S.; Tanaka, T. The assessment of land suitability for urban development in the anticipated rapid urbanization area from the Belt and Road Initiative: A case study of Nong Khai City, Thailand. *Sustain. Cities Soc.* **2022**, *83*, 103988. [CrossRef]
53. Saxena, A.; Jat, M.K. Land suitability and urban growth modeling: Development of SLEUTH-Suitability. *Comput. Environ. Urban Syst.* **2020**, *81*, 101475. [CrossRef]
54. Foroozesh, F.; Monavari, S.M.; Salmanmahiny, A.; Robati, M.; Rahimi, R. Assessment of sustainable urban development based on a hybrid decision-making approach: Group fuzzy BWM, AHP, and TOPSIS–GIS. *Sustain. Cities Soc.* **2022**, *76*, 103402. [CrossRef]
55. Wang, Y.; Jin, C.; Lu, M.; Lu, Y. Assessing the suitability of regional human settlements environment from a different preferences perspective: A case study of Zhejiang Province, China. *Habitat Int.* **2017**, *70*, 1–12. [CrossRef]
56. Hengl, T. *Soil Texture Classes (USDA System) for 6 Soil Depths (0, 10, 30, 60, 100 and 200 cm) at 250 m*; Zenodo: Geneve, Switzerland, 2018.
57. Chen, Z.; Yu, B.; Yang, C.; Zhou, Y.; Yao, S.; Qian, X.; Wang, C.; Wu, B.; Wu, J. An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration. *Earth Syst. Sci. Data* **2021**, *13*, 889–906. [CrossRef]
58. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
59. Ahmad, M.W.; Reynolds, J.; Rezgui, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *J. Clean. Prod.* **2018**, *203*, 810–821. [CrossRef]
60. Ramraj, S.; Uzir, N.; Sunil, R.; Banerjee, S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *Int. J. Control. Theory Appl.* **2016**, *9*, 651–662.
61. Barzegar, R.; Razzagh, S.; Quilty, J.; Adamowski, J.; Kheyrollah Pour, H.; Booij, M.J. Improving GALDIT-based groundwater vulnerability predictive mapping using coupled resampling algorithms and machine learning models. *J. Hydrol.* **2021**, *598*, 126370. [CrossRef]
62. Ke, G. LightGBM A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
63. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041. [CrossRef]
64. Wen, X.; Xie, Y.; Wu, L.; Jiang, L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid. Anal. Prev.* **2021**, *159*, 106261. [CrossRef]
65. Yang, Y.; Lv, H.; Chen, N. A survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **2023**, *56*, 5545–5589.
66. El Naqa, I.; Ruan, D.; Valdes, G.; Dekker, A.; McNutt, T.; Ge, Y.; Wu, Q.J.; Oh, J.H.; Thor, M.; Smith, W. Machine learning and modeling: Data, validation, communication challenges. *Med. Phys.* **2018**, *45*, e834–e840. [CrossRef]
67. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
68. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
69. Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.* **2023**, *31*, 3360–3379. [CrossRef]
70. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6666.
71. Jiang, Z. Deep Learning for Spatiotemporal Big Data: A Vision on Opportunities and Challenges. *arXiv* **2023**, arXiv:2310.19957.
72. Qian, J.; Peng, Y.; Luo, C.; Wu, C.; Du, Q. Urban land expansion and sustainable land use policy in Shenzhen: A case study of China's rapid urbanization. *Sustainability* **2015**, *8*, 16. [CrossRef]
73. Wang, J.; Lin, Y.; Glendinning, A.; Xu, Y. Land-use changes and land policies evolution in China's urbanization processes. *Land Use Policy* **2018**, *75*, 375–387. [CrossRef]
74. Ma, T.; Zhou, C.; Pei, T.; Haynie, S.; Fan, J. Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities. *Remote Sens. Environ.* **2012**, *124*, 99–107. [CrossRef]

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.