Article

Where Are Potential Areas for Transit-Oriented Development (TOD)—Exploring the Demands for Built Environment for TOD Planning

Jiyang Xia 1,2 and Yue Zhang 2, *

1 College of Civil Engineering and Architecture, Zhejiang University, No. 866, Yuhangtang Road, Hangzhou 310058, China; xiajiyang0@zju.edu.cn
2 School of Art and Archaeology, Zhejiang University City College, No. 51, Huzhou Street, Hangzhou 310015, China
* Correspondence: zhangyue@zucc.edu.cn

Abstract: Sustainable urban growth advocates the implementation of transit-oriented development (TOD) to optimize urban spatial structure. The bilateral planning concept of TOD emphasizes the importance of discovering areas with existing TOD features but poor public transit service (potential TOD areas) and further introducing transit connectivity or conducting TOD policy in such areas to facilitate sustainable transportation. However, current studies that are devoted to discovering potential TOD areas remain scarce. In this study, we find that random forest (RF) is an optimal algorithm that can effectively identify potential TOD regions in Hong Kong. We propose an RF-mediated machine learning model (RF-TPI model) and reveal underlying mechanisms of specific indicators. After iteratively learning the typical features of TOD areas in Hong Kong, the developed RF-TPI model shows great capacity to identify potential TOD areas, with satisfactory model performances (accuracy score: 0.89, precision score: 0.81). Further investigation on manifestations of indicators by the SHapley Additive exPlanations (SHAP) interpreter demonstrates the intricate, significant nonlinear and threshold effects of distinct indicators. Conclusively, we highlight that random forest would be a prospective tool for identifying potential TOD areas to aid TOD strategy in urban sustainable endeavors.

Keywords: transit-oriented development (TOD); random forest; big data; land use; local effect

1. Introduction

Applying transit-oriented development (TOD) has been extensively recognized as a major approach to promoting sustainable and wholesome urban development [1–3]. TOD strategy is capable of integrating advisable land use and well-designed transit systems, thereby mitigating urban disease [4–6] (traffic congestion and air pollution, etc.) to create urban vibrancy.

A majority of current studies and practice regarding TOD planning focus on urban outskirts or greenfield, while emerging evidence prompts attention toward urban centers [7–10]. Most importantly, TOD has been denoted as “bilateral planning” [11–13] that should not only support the creation of enhanced transit-oriented urban sustainable growth, but also the introduction of transit to areas with already existing TOD features (e.g., diverse/high-intensity land use and high-density street network). Inner urban areas, in contrast, are more prone to possess such physical characteristics than peripheral regions. Moreover, infill development in downtown or high-density city areas have been proposed to be a more favorable way for TOD implement, because their superior existing densities in the vicinity can benefit stronger transit utilization [14–16]. Therefore, deciphering city areas with physical TOD features but lacking public transport connectivity (i.e., identify potential TOD areas) and answering manifestations of such characteristics are
required to aid justifiable TOD planning decisions, and hold great potential to be urban sustainability boosters.

The identification of potential TOD areas has provided location decision-making tools for TOD planning. Based on the 3Ds principle (density, diversity, and design) of TOD, Singh et al. first attempted to establish evaluation indicators of a potential TOD index and investigate the demand for TOD construction in various regions of Arnhem Nijmegen City using the spatial multicriteria analysis (SMCA) method [13]. Strong et al. constructed a theoretical framework of multiple-criteria tool after screening TOD indicators by an analytical hierarchical process (AHP) to answer how transit institutions select suitable areas to conduct TOD [17]. Using GIS and AHP methods, a decision-making tool was studied by Taki et al. to discover optimal TOD planning areas in Jakarta Metropolitan Region [18]. However, there are still some other issues that deserve consideration and remain unresolved. For example, TOD indexes and AHP approaches used in current studies performed weight setting in subjective ways, such as expert assignment, which would inevitably introduce artificial interference. In addition, the validity and reliability of these approaches lack verification. More importantly, analysis of the roles of relevant indicators (land use diversity and intensity, etc.) in determining potential TOD areas is currently ignored, while a mechanism of specific indicators is essential for future TOD planning to decide the optimal intensity range and type combination of land use in certain areas.

Hence, this study aimed to: (i) explore an effective model to identify urban potential TOD areas, and (ii) to dissect effects of each indicator in the process, accordingly, aiding agencies and professionals in making TOD planning decisions, including TOD location and land use optimization. To fulfil the objectives, a data-driven machine learning algorithm and a corresponding interpreter were used. Firstly, whether a specific area has a typical TOD feature was evaluated by random forest (RF). RF exhibits excellent performance in identifying urban functional areas and urban land use type, and has been adopted in a number of urban computer studies [19–21]. Endowed with a robust classification function and a superior algorithm (multiple decision trees randomly vote to comprehensively define the final result), RF is more prepared to avoid issues such as variable endogeneity and model overfitting [22]. However, RF has poor interpretation performance of specific action modes of variables. Thus, the SHapley Additive exPlanations (SHAP) interpreter, which can evaluate the global/local effects of variables based on their classic co-operative contribution, was adopted to expose the role of variables.

Herein, we constructed an RF-mediated potential TOD area identification model, designated as RF-TPI model, which has not been reported yet. To showcase the feasibility and application of the RF-TPI model, Hong Kong was taken as a case study. The RF algorithm and multisource urban data were utilized to excavate built environment characteristics of TOD regions. After training, the RF-TPI model was endowed with the capability to identify potential TOD areas in Hong Kong. Taken together, this work advanced our understanding of the roles of indicators in TOD area identification and provided insight that RF would be a promising tool for recognizing potential TOD areas.

2. Literature Review

So far, the field of TOD planning has shown significant progress in planning policy and land use optimization models, whereas studies concerning TOD sites selection is still scarce [8]. In general, research objects of discovering apposite regions for TOD construction can be divided into two categories, including a set of optional objects (generally multiple preset, to-be-built rail stations) and the entire urban area.

For judging the optimal area for TOD among the preset proposed station areas, representative research was conducted by the calculation/determination through the fuzzy analytic network process (FANP) method [23,24] and the location decision tool based on sieving TOD indicators and weights through the AHP approach [17]. Both studies helped transport agencies or personnel to select TOD sites to a certain extent, and proposed suggestions for land use optimization. Unfortunately, potential TOD areas within city
contexts were ignored. Since they are more extensive and undiscovered, finding them is of considerable importance to effectively support the TOD policy decision and advance urban redevelopment and sustainability by providing insights for TOD sites selection [4,25].

Kamruzzaman et al., Motieyan et al., Singh et al., and Taki et al. have made great contributions in exploring potential TOD areas at the whole city level [4,12,13,18], delineating that city areas with a high potential TOD level feature a high land use mix and facility diversity. Noteworthily, the smallest study units (census area [4], community [12], and metropolitan area [18]) in these studies are larger than the coverage of a TOD site (usually 10-minute walking range), making it difficult to provide more refined and effective results such as determining how many TOD sites are required in these areas and which area should be prioritized for TOD construction. Considering the practical service scope of TOD, Singh et al. divided the Arnhem Nijmegen City Region into several 300 m × 300 m grids as research objects to calculate the potential TOD index of each grid [13]. Encouragingly, they precisely and effectively disclosed locations with high TOD potential but poor transit connectivity.

The selection of indicators is a vital procedure in the topic of identifying TOD areas. The 3Ds principle (density, diversity, and design) of TOD is often used to guide indicator choice and ensure that indicators conform to TOD characteristics. However, due to the potential differences between TOD theory and practice and subjective influence of different stakeholders, relying on top-level theory to select indicators could be insufficient, such that practice-based indicator selection is important for TOD planning study. Specifically, in research subjects such as TOD effect and TOD typology, indicators mainly include land use diversity, facilities (commercial and residential), socioeconomic factors (population and housing prices), and urban design factors (accessibility). For TOD planning study, land use (land use type, high density, and high diversity) is the practice’s way and is one of the ultimate manifestations [4,12,26,27]. Additionally, working as the most critical factor that affects the implementation/transfer of TOD policies, land use has been widely recognized as a quantitative decision-making basis for evaluating and determining TOD areas [28]. Therefore, to perform identification of potential TOD areas, we exploited point-of-interest (POI) data to construct land use type indicators and land use diversity indicators, respectively. Meanwhile, building footprint data and open street map (OSM) data were used on behalf of the floor area ratio (FAR) and the street/intersection density indicator, respectively.

Accuracy/precision is the prerequisite for demonstrating the applicability of the method/model for TOD area identification. Unfortunately, the validation process of certain methods/models has been neglected in current studies. In general, the outputs (i.e., potential TOD area) of decision-making tools adopted in these studies are the sum of scores for various TOD indicators and weights. However, the effect of each indicator was assumed to be linear, plus inappropriate selection of indicators may occur due to potential differences between theory and practice. Thus, the validity of the approach/model requires examination. In order to remedy such a defect, the accuracy/precision score and AUC value were used to test the model validity in this work.

3. Study Area and Data

3.1. Study Area

Our case study to demonstrate the conception is Hong Kong. Hong Kong is one of China’s special administrative regions and an international metropolis in Asia, with high population density and compact urban form. Covering a total area of 1100 km² (Figure 1) and a population of over 7 million, the vast majority of Hong Kong residents live in built-up regions within an area of approximately 264 km². Hong Kong was chosen as the study area in this study because: (1) Hong Kong is documented as one of the earliest cities to launch a TOD practice and achieve success, and has been conclusively proven by scholars to appear to be a typical TOD mode [16,29,30], such that it is feasible to endow RF capability to recognize typical TOD features after a repeated iterative learning process, taking high
TOD level areas in Hong Kong as learning samples; and (2) Hong Kong has one of the most well-established railway systems worldwide [15]. Studies have shown that public transport occupies 90% of personal travel in Hong Kong [15], of which rail transport accounts for 43% of total public transport [30], revealing that rail transport is the backbone of both public transport and residents’ daily life. Thus, potential TOD areas identification and further public transit introduction are prospective to help improve Hong Kong’s rail transit and assist in urban renewal/redevelopment under the sustainable growth vision [8,16].

![Figure 1. The study area of Hong Kong, China.](image)

Based on the insight of Singh’s study, and the definition of TOD area (10-minute walking range), we divided Hong Kong into several 600 m × 600 m grids as the smallest study unit in this work. In addition, the division of 400 m × 400 m grids and 800 m × 800 m grids have been considered for comparisons and are shown in Figures S1 and S2 and Table S1, Supplementary Materials. Only 600 m × 600 m grids with the highest model accuracy/precision score were shown in the text.

### 3.2. Data Source and Process

In this study, multisource open urban data were chosen to construct built environment indicators (representing land use type and floor area ratio (FAR), etc.), thereby helping the RF-TPI model to learn and identify TOD features. The collection and processing of various data are as follows:

1. **POI:** It has been substantially manifested to reflect various activities in the city and characterize urban land use types [31–35]. Specifically, POI data were collected from amap.com (China’s largest web platform for mapping, navigation, and location-based services). After cleaning duplicate and useless data (e.g., the primary type of “Incidents and Events” and “Indoor facilities”), the collected POIs belong to 20 primary types and 262 mid types, with a total number of 209,976 in Hong Kong. Afterwards, the POIs were divided into 9 land use types in the preprocessing step according to POI mid types (https://lbs.amap.com/api/webservice/download, accessed on 21 June 2021) and the “Land Utilization in Hong Kong 2020” (https://www.pland.gov, accessed on 21 June 2021).

2. **Building data:** Hong Kong’s building data (including building footprint and floor data) were collected from amap.com. After geometry repairing and cleaning, Hong
Kong possessed 260,261 footprints. We further calculated the average floor area ratios (AFAR) of each 600 m × 600 m grid in Hong Kong.

(3) Street network data: Hong Kong’s street network data were obtained from openstreetmap.org, then cleaned and filtered in the ArcGIS platform. Afterwards, they were converted to raster images by a buffer zone tool, followed by using an ArcScan tool in the ArcGIS Desktop 10.5 (ESRI, Redlands, CA, USA) to generate the road centerline. Then, the topology tool was adopted to break the street network to produce intersection and exclude the breakpoint data in the complete line. All the above data were collected from June 2021 to August 2021.

4. Methodology
4.1. Methodological Framework

Figure 2 illustrates the methodological framework of identifying potential TOD areas in Hong Kong by the RF-TPI model and analyzing the effects of indicators. The scheme was proposed according to the following four procedures: (a) Data collection, cleaning and processing. (b) Construction of geodatabases: Hong Kong was divided by grids with different sizes in the ArcGIS geographic information platform, then the cleaned data was connected to each nearest neighbor grid. (c) Model training and evaluation: the rational training set and test set were established firstly. After continuous iterative learning and parameter adjustment, the optimal RF-TPI model with high recognition accuracy/precision for physical TOD characteristics was ultimately obtained. (d) Identification of potential TOD areas and interpretation of indicators: all study units in Hong Kong were examined by the RF-TPI model to determine whether TOD characteristics exist. Finally, the contributions of different indicators in the identification process were revealed by the SHAP interpreter.

Figure 2. Methodological framework of identifying potential transit-oriented development (TOD) areas by the RF-TPI model.
4.2. Variables

Dependent variable (label “1” or “0”): There were 96 metro stations in Hong Kong up to 2021. Notwithstanding TOD practice success in Hong Kong, not all station areas exhibit the physical characteristics of TOD. In terms of the research conclusions of TOD levels of Hong Kong subway stations [33], 13 station areas with low TOD levels (such as airport station, Disney station, etc.) were excluded in this study. Afterwards, grids where the remaining 83 station areas (with high TOD levels) are located were labeled as “1” and deemed to own TOD physical characteristics.

Independent variables (indicators): The 3Ds principle of TOD, the critical theoretical basis for selecting indicators, was adopted to construct 13 indicators (X1–X13) representing three dimensions (Density, Diversity, and Design) in this study. Especially, as discussed in the literature review section, land use indicators can provide practical assistance and insights for TOD planning. Thus, we established nine land use indicators according to Hong Kong’s land use classification, including industrial (X1), woodland/shrubland/grassland/wetland (X2), agriculture (X3), other urban or built-up land (X4), commercial (X5), water bodies (X6), transportation (X7), institutional/open space (X8), and residential (X9). The land use intensities of the above nine indicators were reflected by the numbers of their representative POIs. Overall, combined with AFAR (X10), which indicates the average floor area ratio within the study unit, X1–X10 jointly constituted indicators representing Density. Land use diversity index (X11), calculated according to the number of different land use types in the study unit, represented Diversity. Street density (X12) and intersection density (X13), which reflect the accessibility and block form, were defined as indicators representing transit-oriented design [30]. Addressed concretely, the trained RF-TPI model identified and judged whether each grid has typical TOD physical characteristics according to the 13 indicators. Information for the indicators is shown in Table 1.

Table 1. Descriptions and values of dependent and independent variables.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
<th>Observations (Obs)</th>
<th>Mean</th>
<th>St. Dev. *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Dependent variable = 1 high TOD level area, = 0 otherwise</td>
<td>Labeled with “1”: 83 Labeled with “0”: 4157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1</td>
<td>Industrial</td>
<td>4240</td>
<td>0.026</td>
<td>0.133</td>
</tr>
<tr>
<td>X2</td>
<td>Woodland/Shrubland/Grassland/Wetland</td>
<td>4240</td>
<td>0.067</td>
<td>0.580</td>
</tr>
<tr>
<td>X3</td>
<td>Agriculture</td>
<td>4240</td>
<td>0.025</td>
<td>0.215</td>
</tr>
<tr>
<td>X4</td>
<td>Other Urban or Built-up Land</td>
<td>4240</td>
<td>0.008</td>
<td>0.183</td>
</tr>
<tr>
<td>X5</td>
<td>Commercial</td>
<td>4240</td>
<td>0.267</td>
<td>0.385</td>
</tr>
<tr>
<td>X6</td>
<td>Water Bodies</td>
<td>4240</td>
<td>0.147</td>
<td>0.662</td>
</tr>
<tr>
<td>X7</td>
<td>Transportation</td>
<td>4240</td>
<td>0.153</td>
<td>0.315</td>
</tr>
<tr>
<td>X8</td>
<td>Institutional/Open Space</td>
<td>4240</td>
<td>0.135</td>
<td>0.283</td>
</tr>
<tr>
<td>X9</td>
<td>Residential</td>
<td>4240</td>
<td>0.147</td>
<td>0.330</td>
</tr>
<tr>
<td>X10</td>
<td>AFAR: The average FAR within each grid cell was calculated by: ( R = \frac{F}{A} ), where ( R ) is the average FAR within each grid cell, ( F ) is the total floor area of buildings, and ( A ) is the total area of each grid cell.</td>
<td>4240</td>
<td>0.028</td>
<td>0.082</td>
</tr>
</tbody>
</table>
### Table 1. Cont.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
<th>Observations (Obs)</th>
<th>Mean</th>
<th>St. Dev. *</th>
</tr>
</thead>
<tbody>
<tr>
<td>X11</td>
<td>Land use diversity index: The diversity of land use was calculated on the basis of the Shannon entropy index: $H = -\sum_{i=1}^{k} p_i \ln(p_i) / \ln(k)$, where $H$ represents the value of land use diversity, $p_i$ represents the percentage of POI numbers on type $i$, and $k$ represents the total number of POI types in a grid cell.</td>
<td>4240</td>
<td>0.677</td>
<td>0.800</td>
</tr>
<tr>
<td>X12</td>
<td>Street density: The ratio of total length of the road centerline to grid area.</td>
<td>4240</td>
<td>0.141</td>
<td>0.155</td>
</tr>
<tr>
<td>X13</td>
<td>Intersection density: The ratio of the number of intersections to the grid area.</td>
<td>4240</td>
<td>0.017</td>
<td>0.066</td>
</tr>
</tbody>
</table>

* St. dev. represents standard deviation. X1–X9 indicates the numbers of point-of-interest (POI) after weights assignment. Data of X10, X12, and X13 were normalized.

Weights assignment: The term frequency-inverse document frequency (TF-IDF) method was adopted to minimize the influence of counts imbalance caused by the lack of area information in POI data (e.g., the counts of certain POI type (commercial facilities) tend to be much larger than that of another type (water bodies)), and to highlight the different land use conditions in each grid. TF-IDF is a commonly used weighting technique for information retrieval and data mining [36,37]. Considering the overall numbers of POIs in a city, we endowed the corresponding weight to distinct POI types through TF-IDF. The weight of the POI type $j$ in grid cell $R_i$ was calculated as follows:

$$TFIDF_{ij} = TF_{ij} \times \log_2 \frac{N}{DF_j}$$

where $TF_{ij}$ is the proportion of POIs on type $j$ among all POIs in $R_i$, $N$ is the number of grid cells in Hong Kong, $DF_j$ is the number of grid cells contained in type $j$.

#### 4.3. Training Set and Test Set

In our study, 83 grids and 4157 grids were labeled with “1” and “0”, respectively. Thus, sample imbalance (83:4157) requires consideration to avoid reducing the effectiveness of model recognition. On the basis of previous literature and multiple attempts, 300 grids were randomly selected from 4157 “0”-labeled grids together with 83 “1”-labeled grids to form the Hong Kong data set (383 samples) (Table S2, Supplementary Materials). Next, 70% and 30% of the Hong Kong data set were randomly chosen as the training set and test set, respectively. As described, the training set was used by the RF-TPI model to learn TOD physical features and the test set was utilized to test the performance of model classifier after training.

#### 4.4. Modeling and Evaluation

A machine learning algorithm with powerful classification capability and accuracy/precision is required since 4240 grids in Hong Kong need to be classified and labeled. As reported, among the big data-driven classification machine algorithms (including XGBoost, Naïve Bayes, SVM, LSTM, and dense layer), RF is the strongest algorithm to execute identification of different urban land use types and functional areas through POI data [20,21,38,39]. The RF classifier consists of a series of decision trees, each of which is constructed by randomly extracting a subset of observations and features from the dataset, making every decision tree more independent and less correlated. Each decision tree in RF employs a range of yes/no questions based on a single or multiple indicators to partition the training data. At each node, the decision tree samples the features with replacement, and splits growth by calculating the amount of information contained in each feature. The output is obtained through aggregating the votes from the combined decision trees. Therefore,
RF can not only solve the drawback of contributions of weaker variables being covered by strong variables, but can also minimize the possibility of overfitting. To further examine the superiority of RF, we compared RF with other popular algorithms (i.e., decision tree (DT) and XGBoost) in this study.

To explore the validity of RF-TPI model recognition results, we used standard evaluation metrics against machine learning, including Precision, Recall, F1, Accuracy, and AUC value. To be exact, precision score (PS), also known as precision rate, represents the proportion of samples with true label “1” in the entire “1”-labeled outputs. A higher PS indicates the superior precision of model on identifying TOD features. Recall is the ratio of correctly identified samples with true label “1” in the test set. The higher recall score (RS) of the model illustrates that more areas with TOD features can be identified by it. F1 is a comprehensive consideration of both PS and RS. Accuracy score (AS) stands for the model’s correct recognition rate against all sample labels (i.e., both labels “0” and “1”) in the test set. Furthermore, due to the use of a binary classification task (judging labels “0” and “1”), we evaluated the validity of results recognized by the RF-TPI model via AUC value, which displays the proportion of negative instances that are misclassified as positive.

4.5. Explainer

SHAP is among the most effective machine learning interpreters and was used in this work to prevent a negative impact of a “black box” regarding machine learning (i.e., poor interpretability of model process) and analyze the specific manifestations of indicators in areas with TOD characteristics. The SHAP interpreter calculates the Shapley value of each indicator per grid (for detailed explanation, please refer to the work of Lundberg et al. [40,41] and Xiao et al. [42]), thereby benefiting the explanations of global/local effects of different indicators. We believe that this is favorable to provide more concrete and feasible assistance for TOD planning decisions. The Shapley value was calculated as follows:

\[
\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[ f_x(P_R^i \cup i) - f_x(P_R^i) \right]
\]

where \(\phi_i(f, x)\) is the Shapley value of indicator \(i\) for the single model voting. \(\mathcal{R}\) is the set of all indicator orderings. \(P_R^i\) is the set of all indicators that come before the indicator \(i\) in ordering \(R\). \(M\) is the number of indicators.

Moreover, the relative importance of each indicator was estimated as follows:

\[
I_i = \frac{1}{n} \sum_{k=1}^{n} \left| \phi_i^{(k)} \right|
\]

where \(I_i\) represents the importance of indicator \(i\). \(\phi_i^{(k)}\) is the indicator \(i\)’s Shapley value for each single model voting \(k\).

5. Results

5.1. Model Score

Table 2 shows the scores of different validators of three classifiers (RF, DT, and XGBoost). As described, we aimed to explore a model that enables one to effectively and precisely identify regions with TOD features. Thus, precision scores (PS) of these classifiers were mainly considered. As shown, RF exhibited higher PS and F1 than did DT and XGBoost, verifying the superiority of RF in determining TOD characteristics. This can be further demonstrated by the AUC value (0.916). Moreover, RF with respectable accuracy score (0.891) was advisable for recognizing whether a certain area is potential TOD region. Overall, RF was the optimal machine algorithm.
5.2. Hong Kong Potential TOD Levels

Hong Kong potential TOD levels represent the probability score (range from 0.00 to 1.00) of each grid exhibiting TOD characteristics, as judged by the RF-TPI model. After visualizing the score of each grid, the hot map of Hong Kong potential TOD levels was created, as shown in Figure 3. As displayed, a majority of high-value areas (red) are distributed in the southern area of Hong Kong and are near existing rail lines, showing a clustering pattern. Furthermore, from near to far along the rail lines, Hong Kong potential TOD levels demonstrated a trend from high (red) to low (green). Low-value areas (green) dominated the entire research area and presented a clustering pattern, whereas median-value areas (yellow) served as transition and connection between high-value areas and low-value areas and showed a certain linear trend.

![Figure 3. Potential TOD levels in Hong Kong.](image)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>PS</th>
<th>RS</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.891</td>
<td>0.810</td>
<td>0.773</td>
<td>0.791</td>
<td>0.916</td>
</tr>
<tr>
<td>DT</td>
<td>0.849</td>
<td>0.583</td>
<td>0.667</td>
<td>0.622</td>
<td>0.699</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.878</td>
<td>0.762</td>
<td>0.800</td>
<td>0.780</td>
<td>0.896</td>
</tr>
</tbody>
</table>

**Table 2. Validator scores of random forest (RF), decision tree (DT), and XGBoost classifiers.**

PS, RS, and AUC indicate precision score, recall score, and area under receiver operating-characteristic, respectively.

5.3. Hong Kong Potential TOD Areas

Hong Kong potential TOD areas were further identified by screening out grids with a potential TOD level score over 0.5. Specifically, in the binary classification task of the RF algorithm, only when the probability score of label “1” of each grid exceeds that of label “0” can the sample be judged as label “1”, indicating the score threshold is 0.5. Therefore, areas with a potential TOD level score of more than 0.5 were defined as potential TOD areas. As illustrated in Figure 4, except for area 14 (Islands District) and area 18 (Southern District), significant clusters of potential TOD areas (pink block) were shown across other regions in Hong Kong. These clusters are therefore recommended to decision makers (urban/regional planning and policy makers) for future consideration. On the one hand, after detailed assessment and screening of areas with poor transit connectivity in these clusters, introducing public transit services to them would achieve better efficacy due to their considerable demands of the built environment. On the other hand, prospective areas...
in the clusters that have been analyzed through technical and socioeconomic feasibility could be preferentially considered to contribute to urban renewal/redevelopment and sustainability guided by TOD policies.

![Figure 4. Potential TOD areas in Hong Kong identified by the RF-TPI model.](image)

### 5.4. Relative Importance of Indicators

The relative importance and ranking of 13 indicators in identifying potential TOD areas were visualized through calculating and outputting the mean Shapley value (MSV). As shown in the Figure 5a, AFAR (X10), street density (X12), commercial (X5), intersection density (X13), and institutional/open space (X8) were the top five significant indicators. AFAR (X10) exhibited considerable impact on the identification of potential TOD area, as verified by the fact that it possessed the highest MSV (0.094), whereas land use diversity index (X11), transportation (X7), and other urban and built-up land (X4) demonstrated comparable efficacy. In contrast, agriculture (X3), residential (X9), water bodies (X6), and woodland/shrubland/grassland/wetland (X2) showed negligible roles, with MSVs all below 0.009.

Next, we explored the specific contribution of each indicator in identifying potential TOD areas (i.e., global effects) via the SHAP interpreter. As indicated by the individual points on the x-axis in Figure 5a, the piled-up position represents a positive (Shapley values above 0) or negative (Shapley values below 0) impact. Furthermore, the different tail lengths on both sides of the x-axis indicate the influential degree, while the colors of these points illustrate the values of diverse samples. Figure 5b shows that the top six principal indicators (i.e., X10, X12, X5, X13, X8, and X11) are prone to have positive impacts on the recognition of TOD area, whereas other urban or built-up land (X4), industry (X1), and water bodies (X6) showed negative impacts. Specifically, the tail on the right side of the x-axis (consisted of red and blue points) for AFAR (X10) was longer than that for the left side (mainly consisted of blue points), implying that although demonstrating an overall positive impact, the various values of AFAR (X10) could influence the identification process in complex manners (i.e., high/low value resulted in a positive/double-sided effect). The same pattern can be seen in X13, X8, and X7. Moreover, it was revealed that residential (X9) exhibited bilateral effects (Figure 5b). Herein, these results elucidated that barely any indicators owned a simple relationship with the recognition results of the RF-TPI model.
5.5. Nonlinear Effects of Indicators

We further investigated the specific mode of action of each indicator per grid (i.e., local effects) in the RF-TPI-conducted potential TOD areas identification process by mapping the partial dependence plots (PDPs). PDPs graphs of the top eight indicators all depicted nonlinear patterns (Figure 6) (Figure S3, Supplementary Materials shows details of other indicators). It is noticeable that some indicators demonstrated obvious thresholds, and their impacts would attenuate/enhance as their values exceed the thresholds. This finding confirmed our expectation that the extensive utilization of a certain land use type could singularize or specialize the regional functions, thereby reducing the possibility of introducing TOD in such areas.

Specifically, the significant threshold of X10 demonstrated that in TOD planning, not only should one set a minimum value of FAR, but should also consider a rational maximum value (Figure 6). When X12 exceeded its threshold, potential TOD levels in areas with high-value X12 would be promoted. The value of X7 between 0.0 and 0.2 had a relatively positive effect on potential TOD level. However, its influence gradually diminished as its value (impact on model output).
value increased. A similar effect can be seen in X4 as well (Figure 6). This phenomenon indicated that X7 and X4 can be used as negative indicators to judge whether an area could be a potential TOD region, and that land use type of transportation and other urban or built-up land (e.g., utilities and cemeteries facilities) in the area should be reduced or avoided in TOD planning.

Taken together, these results declared the significant nonlinear and threshold effect of indicators, and the mixed and diverse land use situation in TOD areas, such that previous TOD index evaluations which considered the liner effects of indicators could be insufficient.

6. Conclusions and Discussion

TOD as an urban booster considerably promotes the realization of urban sustainable development vision. The introduction of TOD in potential sites within inner urban areas is conducive to the sustainability of land use and transit system. In this study, we constructed a random forest-based machine learning model to give information regarding potential TOD areas in Hong Kong. Through selection of indicators, division and construction of datasets, and the adjustment of model parameters, the optimal RF-TPI model was obtained. Our results showed that the optimal RF-TPI model displayed considerable performance (accuracy score at 0.89, precision score at 0.81, and AUC value at 0.92), and potently identified potential TOD areas in Hong Kong. These findings proved the viability of our hypothesis that, the recognition of the results of the RF-TPI model could effectively support relevant agencies/professionals in TOD site selection. Accordingly, by incorporating consideration of practical social or economic factors, along with output information produced from the RF-TPI model, the government could make favorable TOD planning decisions to foster urban sustainability.

The SHAP interpreter revealed that indicators could exhibit a distinct relative importance. Compared with urban land type proportion, street density and AFAR of the study unit have greater impacts on potential TOD level. In addition, the global effects of indicators on recognizing potential TOD areas were proven to be complex, including positive, negative, and bilateral impacts. Importantly, results of the local effect of indicators confirmed significant nonlinear effects, supplying several insights for TOD study and practice. First, the mode-of-action of various indicators require careful consideration in research on TOD evaluation/measurement. For instance, impacts of AFAR and street density on Hong Kong’s potential TOD level had obvious thresholds. Furthermore, several land use indicators (transportation and institutional/open space) showed diverse influences, as they have inflection points where their local effects changed (slope from positive to negative and threshold effect). This phenomenon indicated that a single land use type has limited impact on potential TOD level. Therefore, in TOD practice, it is necessary to comprehensively consider the interactions and potential synergistic effects of different land use types, which is consistent with the principle of urban land use diversity advocated by TOD.

There are some limitations and prospects in this study worthy of further discussion. First, though POI data can represent urban socioeconomic activities and has been widely used in urban studies, the lack of specific area information restrains its applicability in precise research. For example, current POIs do not strictly correspond to land use types, and their classification is undeveloped and not unified (e.g., commercial POI may appear in some land use types). Therefore, we expect the emergence of available and novel data types containing more urban information to consider TOD study. Second, TOD typology studies have suggested that diverse types of TOD may exist as TOD areas in a city exhibit different developmental levels, such that putting future research on a building training set based on different types of TOD areas (e.g., central/residential TOD area) could help the RF-TPI model to identify corresponding potential TOD areas in response to specific demands. Furthermore, results from the Hong Kong case study can be context-specific; thus, future work can entail extensive research across multiple cities to examine the robustness of the RF-TPI model.
In summary, the RF-TPI model, with a rationale for machine learning, effectively identifies potential TOD areas in Hong Kong, and is promising to be a generalizable framework applied in other cities to aid the execution of the TOD strategy in urban sustainable endeavors.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su14148364/s1, Figure S1: potential TOD areas in 400 m × 400 m Hong Kong grid map and validator scores of random forest (RF), decision tree (DT), and XGBoost classifiers. Figure S2: potential TOD areas in 800 m × 800 m Hong Kong grid map and validator scores of random forest (RF), decision tree (DT), and XGBoost classifiers. Figure S3: partial dependence plots of the remaining five indicators (including X1, X2, X3, X6, and X9) on potential TOD area identification. Table S1: validator scores of random forest (RF) model at different sizes of the smallest study unit (grid) in Hong Kong, Table S2: validator scores of random forest (RF) model at different ratios of Hong Kong training set and Hong Kong test set.

Author Contributions: Conceptualization, J.X. and Y.Z.; data curation, J.X.; formal analysis, J.X.; investigation, J.X.; methodology, J.X.; supervision, Y.Z.; validation, J.X.; writing—original draft, J.X.; writing—review and editing, J.X. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the editors and reviewers who provided insightful comments on improving the quality of this article.

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

References


23. Wey, W.-M. Smart Growth and Transit-Oriented Development Planning in Site Selection for a New Metro Transit Station in Taipei, Taiwan. *Habitat Int.* 2015, 47, 158–168. [CrossRef]


