Analysing Driving Factors of Land Value Based on Big Data in New York

CAI Zhongyu + HUANG Suqi + SHANG Rui + YANG Ziwei

Spring 2022

URBA6002 Urban Big Data Analysis

Content

1. Introduction

1.1. Background

Growing urbanization has many benefits and drawbacks, with the most criticized high housing and land prices. The most important factor in real estate development is the location, and the supply is also an essential factor in the price of land.

With the development of machine learning techniques that have made predictions based on big data more accessible in recent years, many scholars have started to study predicting land prices from different perspectives, aiming to uncover the other influential factors behind the price of land.

1.2. Research Objectives and Framework

The United States is a country of immigrants with a particular ethnic complexity. And New York City (NYC), as one of the major cities in the United States, has a liberal East Coast atmosphere that makes this a culturally inclusive place. This research selects NYC as the study location, known for its high land price, to uncover the relationship between land price and crucial factors.

The research objective is to evaluate the optimal models and identify the most critical factors on land price per square foot in NYC by utilizing non-linear machine learning models based on data that contains over 500 features collected from various perspectives.

The research framework is shown in Figure1.1. First of all, collecting big data from multiple accesses and conducting data preprocessing. Then, five machine learning models are adopted, with parameter optimization and feature selection. After that, creating the visualization and spatial analysis of the selected factors through ArcGIS, The methodological framework can effectively uncover the relationship between land value and crucial factors according to experimental results.

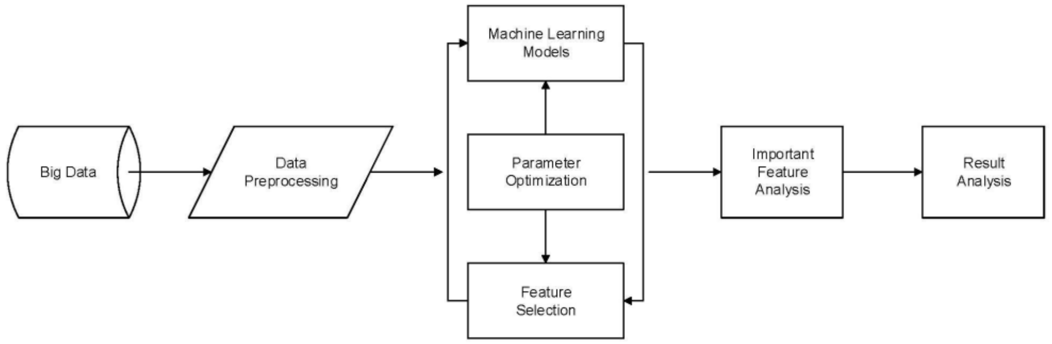

Figure 1.1 Research framework

2. Literature Review

The land price is a crucial reflection of the urban economy, and the values are of great interest to real estate companies as well as the government. Therefore, understanding the effect of other factors on the land price is meaningful for both the real estate market and urban planning.

2.1. Factors Review

Some related articles have considered the relationship between land prices and different factors. Some articles considered the factors of the land attributes. For example, Hu *et al*. (2016) investigated the relationship between land values and land attributes in Wuhan City, China. They showed that the floor ratio area was an essential feature of the land price. Other articles considered the personal characteristics of the residents. Mirkatouli *et al*. (2018) investigated the individual characteristics, especially residents' income level, and educational level, that significantly influenced land values. And there are also some articles considering the economic level. Wen and Goodman (2013) examined a strong connection between economic level and land values using the two-stage least-squares method. But there are still many other factors that should be considered, such as the number of surrounding POIs, the density of the road network, betweenness centrality of roads, greenery level, and so on. And the number of features considered in articles was very limited, with only a few dozen variables. In order to investigate whether other unnoticed variables can affect land value and predict the land values more accurately, this article collected a large number of features from land attributes, demographic characteristics, economic level, and urban environment to perform machine learning models.

2.2. Machine Learning Models Review

Machine learning models mostly rely on data-driven model selection to reveal the features with the greatest explanatory power (Vespignani, 2009). The machine learning models include Lasso (Tibshirani, R. 2011), Random Forest (Breiman, 2001), Support Vector Machine (Noble, 2004), Gradient Boosting Decision Tree (Stojić *et al*., 2019), K Nearest Neighbours (Song *et al*., 2017), Multiple Linear Regression (Linear regression, 2018), and Multilayer Perceptron (Velo *et al*., 2014).

To be specific, to investigate the relationship between land values and other factors, there has been a lot of research based on large-scale data and machine-learning-based models. Simlai (2021) used least squares-based machine learning models, such as MLR and LASSO, and other conventional regression models to estimate housing values in California. The results showed that the machine learning models performed better than the traditional regression models. Ceh *et al*. (2018) estimated apartment prices using RF and OLS methods and showed better performance using the RF model. Singh *et al*. (2020) used housing price data to predict prices using RF, gradient boosting, and LASSO machine learning models.The study showed that the prediction accuracy of gradient boosting was the highest. Ma *et al.*(2020) used some kinds of models, such as MLR, SVR, KNN, and RF, to Predict housing prices in Montreal. The result showed that the tree-based model performed best. And Deep Learning, such as MLP, is a rising topic in recent years and is also suitable for predicting with large-scale data. The MLR, RF, GBDT, KNN, and MLP models are selected to predict the land price in this study.

3. Data Collection and Pre-processing

3.1. Data Collection

There is six categories of collected data: land attribute, demographic, economic, public safety, transportation, and POI, which are 504 features in total. The relevant characteristics of the dataset are shown in Table 3.1.

Table 3.1 Data source

Category	Dataset	Time Period	Geographic Unit	Data Source
Land Attribute	PLUTO	2021	Tax Lot	https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
Demographics	Population	2020	Census Tract	https://data.census.gov/cedsci/table?g=0400000US36%241000000
	Number of Fertility	2020		
	Number of Disability	2020		
	Number of Ethnicity	2020		
	Number of Housing Units	2020		
Economic	Class of Worker	2020		
	Commuting	2020		
	Employment	2020		
	Income and Poverty	2020		
Public Safety	Motor Vehicle Collisions	2022	/	https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
Transportation	Road Network	2022	/	https://data.cityofnewyork.us/City-Government/road/svwp-sbcd
	Street Rating	2022	/	https://data.cityofnewyork.us/Transportation/Street-Rating/mxi3-5xz5
POI	POI	2022	/	https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj

3.2. Data Preprocessing

3.2.1. Target Variable

The target variable is the unit land price from land attribute data to mitigate the impacts brought by the assessed land price, which is recorded at the tax lot level in the collected dataset PLUTO, a land attribute dataset. Therefore, the assessed land price is divided by lot area to calculate the land price per square foot (land price/sf). And the vacant, zero, or infinite value in the land price is deleted since the land price is the label of this study. After that, adopting the three-sigma rule to remove the outlines that ensure the study's effectiveness. Tax lots with land value/sf that fall out of [μ-3, μ+3] are viewed as outliers and are deleted. Here μ is the mean value and is the standard deviation. After this, we obtain the land value/sf of 840118 tax lots.

This study selects the census block as the unified geographic unit. Therefore, the next step is to group the land price based on the median value to transfer the geographic unit from the tax lot to the census block, obtaining the land value/sf of 31975 census blocks.

Furthermore, most machine learning models assume variables obey the normal distribution. But the result of unit land price shows a significant right skewness (Figure 3.1 (a)). Thus, this study takes a log of the target variable, allowing more significant values than the median to be reduced by a certain percentage, resulting in normally distributed data (Figur 3.1 (b)) that enhances the prediction capacity.

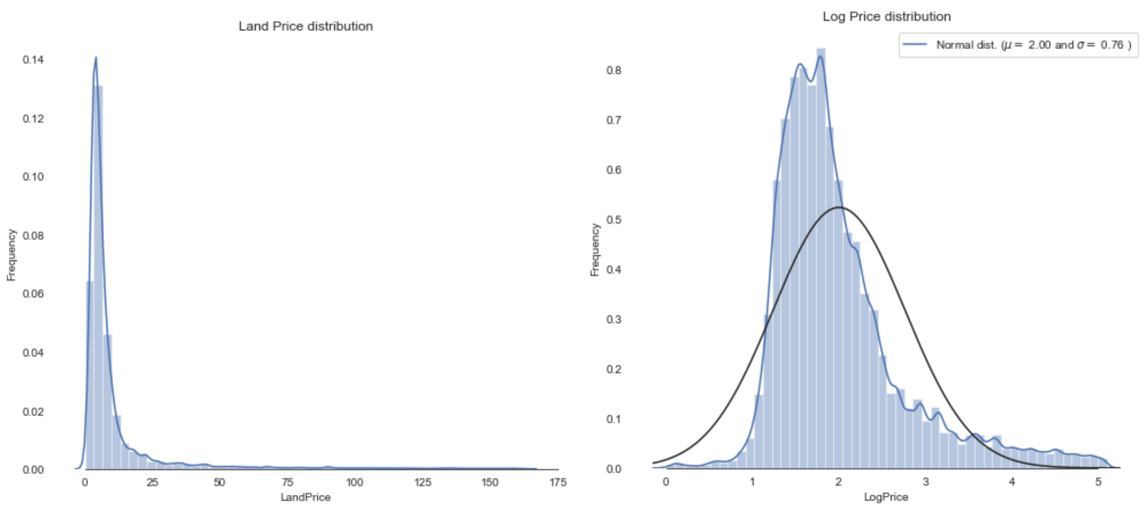

Figure 3.1 (a) *Figur3.1 (b)*

Figure 3.1 (a) Unit land price; (b) Log of unit land price

3.2.2. Feature Variables

Four steps are adopted for feature preprocessing, considering the data with more than five hundred potential impact features and data availability. First of all, remove all the columns with missing values to maintain data objectivity. Unifying geographic units is required since the features are collected from various accesses with different geographic units. In the process of integrating the unit of tax lot of the land attribute data to the unit of census block, this study uses the median value for numerical features and the mode value for categorical features to represent the value of each census block. In terms of the geographic and economic data, the unit of them is the census tract. The research applies the median value or the density of the census tract to indicate the value of each census block in the census tract. For the public safety, transportation, and POI data, use the spatial join tool of ArcGIS to merge data into the census block unit. After that, calculate each required feature. The last step is to normalize all feature data using the Z-score method to eliminate extreme data's adverse effects.

Other preprocessing methods are being applied due to the diversity of data's characteristics. For the POI dataset, the study calculates the density of all POI (Figure 3.2 (a)) and the density of different kinds of POI, respectively, to represent the accessibility of different facilities in the census block. In addition, it is referred to as Shannon's diversity index (SHDI) to demonstrate the diversity of POI, as shown in Figure 3.2 (b). The Shannon index has been a popular diversity index in the ecological field (Spellerberg & Fedor, 2003). Its connotation can also be applied to POI diversity. The steps of calculating are as follows. Firstly, divide NYC into fishnets of the same size, and examine the nearest POIs within 100 meters. Then use formula (1) to calculate the diversity of each kind of POI.

$$H' = -\sum_{i=1}^{R} p_i \ln p_i \qquad (1)$$

Where p_i is the proportion of different POI in each unit. After that, sum the values up, and spatial join them into census blocks using the median value. It can be seen that it is Manhattan that has the highest density and diversity of POI.

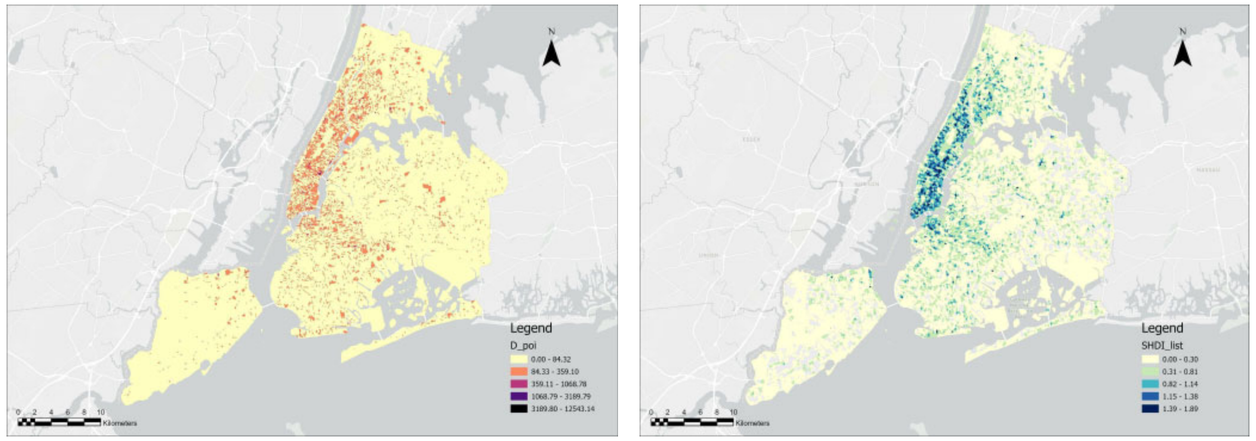

Figure 3.2 (a) *Figure 3.2 (b)*

Figure 3.2 (a) Density of POI; (b) Shannon's diversity index (SHDI)

In terms of transportation data, road density and betweenness centrality are considered to represent traffic conditions. Betweenness counts the number of geodesic paths that pass through a vertex (Cooper, 2015). The study calculates betweenness in sDNA and uses it to represent spatial flows in a real situation.

The feature data after preprocessing contains 32,326 rows and 503 columns.

4. Machine learning models

The project used five machine learning and deep learning models to predict the land value in NYC, which are Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Multiple Linear Regression (MLR), K-Nearest Neighbors (KNN), and Multiple Layer Perceptron (MLP) respectively.

4.1. Experimented algorithms

4.1.1. Random Forest

Random Forest (RF) is an ensemble method in which the basic idea is multiple models (trees) operating as a committee will outperform any individual models, producing improved results. There are two significant categories of ensemble methods: Boosting and Bagging. The main difference between those two methods could be concluded in two aspects:

1. Whether to put back the sample during the random sampling process?

2. Whether the sample weight will be changed during the training process?

RF model belongs to the bagging ensemble method, in which different training data subsets are drawn with replacement, and each model has an equal weight. Therefore, the RF model demonstrates better performance when dealing with unbalanced datasets and solves the over-fitting problem with less discrimination of data noise as a token for consideration.

4.1.2. Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is another ensemble method that belongs to Boosting. Different from the RF model, each GBDT tree is developed iteratively to fit the residual of trees that came before it. Specifically, each tree receives an extra weight, and the incorrect prediction will be assigned more weight in the sequence training process. Such characteristics allow GBDT performs better in reducing bias and be more flexible on the loss function.

4.1.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a supervised machine learning designed for classification problems, assuming similar cases exist in close proximity. It utilizes the output of K-nearest neighbours as the output of one testing data instead of learning the mapping from the training dataset. Due to the above characteristics, KNN is fast in computation but insensitive to outliners. And KNN also can be used for regression.

4.1.4. Multiple Linear Regression

Multiple Linear Regression (MLR) extends the linear approach to modelling the relationship between one dependent variable and more than one independent variable.

4.1.5. Multilayer Perceptron Regression

Multilayer Perceptron (MLP) is a deep learning model based on a neural network. At least three layers of nodes compose an MLP model in typical situations. The nodes are connected by relevant weight to achieve the minimal difference between the network output and the intended output. It needs to state that the input of MLP should be one-dimensional data, requiring flattening the tabular data before infusion.

4.2. Model Comparison

After importing 503 features into models, 75% of the data is separated for training, while 25% is for validation. The preliminary performance of different models is shown in Figure 4.1 And the R-Square Value (R^2) and Mean Squared Error (MSE) are the metrics used to evaluate the model performance.

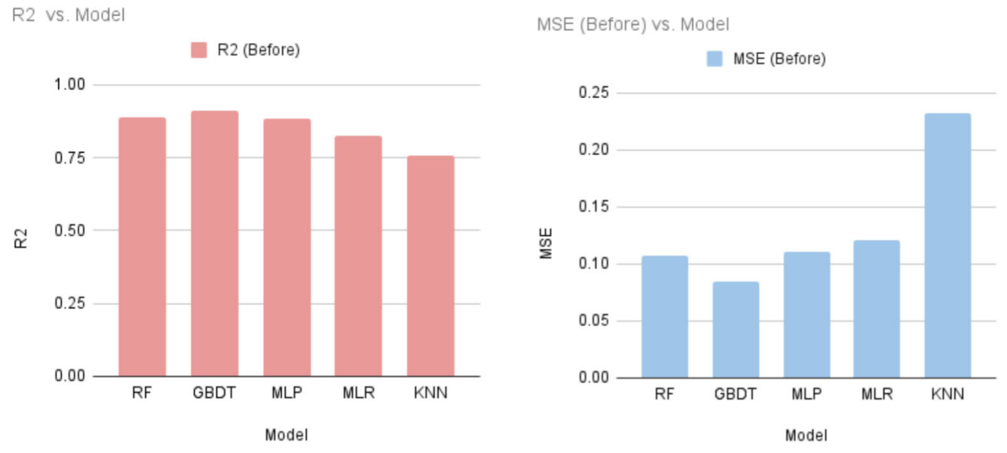

Figure 4.1 R-Square Value and Mean Squared Error of 5 Models

The result indicates that the best prediction model is the GBDT model, of which R^2 is 0.910 and MSE is 0.084. Right after the GBDT model, the RF and MLP model also performs well with an R^2 of 0.887 and 0.84. And the R^2 of the MLR model is 0.826. On the contrary, the worst model for predicting land price is the KNN model, with an R^2 of just 0.758. In summary, compared with the MLR regression and KNN regression models, the ensemble models are more suitable for modelling problems with a large number of different features, such as predicting the land price.

4.3. Parameter Optimization

For parameter optimization, this study only adjusts the parameters of the RF model by GridSearchCV. The number of trees and the max_depth are two important parameters in RF. And this study sets the optimal number of trees in the range of 100 to 1000 with an interval of 100, while the max_depth was searched in the range of 5 to 15 with an interval of 2. As a result, the best parameter is that the number of trees is 900, and the max_depth is 13. After parameter tuning, the R^2 value of the RF model has increased to 0.90088.

4.4. Feature Selection

After comparing the model performance, the study also wants to identify the most important features and assess the effects of important features on models. Recursive feature elimination (RFE) is a method for excluding irrelevant features. Furthermore, test data could be accompanied by recursive feature elimination with cross-validation (RFECV). It can return the number of features that performs the best result on the validation set (Vychegzhanin *et al.*, 2019). And the RFECV method is time-consuming and needs a high-performance computer. Due to the limitations of computer capabilities, the team has conducted the RFECV method, setting the step is 100, and 203 features were selected. However, after importing the 203 features to different models, the evaluation metrics do not change better. It is supposed the step of RFECV is too large to select the precise features. Due to the massive amount of data and limited time, the project teams proposed an alternative way of selecting variables instead of RFECV.

Finally, the study uses the function of feature_importances based on the random forest to select important features. After ranking the features from most important to least important, the different number of features are imported to the models, from 50 to 200 and 5 increments each time. The results show that the most important 60 features perform the best results. The top 60 important features are shown in Table 4.1.

Table 4.1 Selected 60 Features

Category	Feature
Land Attribute	Builtfar
	Commfar
	Facilfar
	Residfar
	Bldgdepth
	Yearbuilt
	Retailarea
	Numfloors
	Landuse_01
	Landuse_02
	Landuse_05
	Schooldist_01
	Schooldist_02
	Schooldist_03
	Schooldist_04
	Schooldist_11
	Schooldist_31
Demographic	Popunder18_percent
	Asian Non-Hispanic Percent
	Black Non-Hispanic Percent
	Hispanic/Latino (Of Any Race) Percent
	Non-Hispanic Of Two Or More Races Percent
	Some Other Race, Non-Hispanic Percent
	White Non-Hispanic Percent
	Population 25 Years And Over, Bachelor's Degree Or Higher
	Fertility: Women 15 To 50 Years, Less Than High School Graduate
	Fertility: Women 15 To 50 Years, Some College Or Associate's Degree
	Fertility: Women 15 To 50 Years, Graduate Or Professional Degree
	Direct-Purchase Health Insurance Alone Or In Combination, 65 Years And Over
	Disability: Some Other Race Alone
	Disability: White Alone, Not Hispanic Or Latino
	Total Civilian Noninstitutionalized Population: White Alone
	Total Civilian Noninstitutionalized Population: Black Or African American Alone
	Total Civilian Noninstitutionalized Population: Asian Alone
	Total Civilian Noninstitutionalized Population: Some Other Race Alone
	Total Civilian Noninstitutionalized Population: White Alone, Not Hispanic Or Latino
	Direct-Purchase Health Insurance Alone Or In Combination, 65 Years And Over
	Private Health Insurance Alone
	Population 16 Years And Over, Asian Alone
	Population 20 To 64 Years, Male
	Population 16 Years And Over, White Alone, Not Hispanic Or Latino
	Population 16 Years And Over, White Alone
	Labor Force Participation Rate: Population 25 To 64 Years, Bachelor's Degree Or Higher
Economic	Median Household Income
	Median Household Income, Asia Alone
	Median Household Income, Black Or African American Alone
	Median Household Income, Householder 25 To 44
	Median Household Income, Householder Over 65
	Median Household Income, White Alone Not Hispanic
	Average Household Size
	Occupied_unit_percent
	$75,000 Or More
	Car, Truck, Or Van - Drove Alone, 35,000 To $49,999
	Public Transportation (Excluding Taxicab), $75,000 Or More

	Walked: $75,000 Or More
Public Safety	Accidentcounts
Transportation	Bthn
	Road_density
POI	D_poi
	Shdi_list

4.5. Result Comparison

After running the models with the optimal parameters and the selected 60 features, it can be seen from Table 4.2 that, except for GBDT, the metrics of other models have improved. KNN improved the most, with R^2 increasing by 7.39% and MSE decreasing by 23.28%. However, GBDT is still the best model for land price prediction in NYC, although the R^2 of GBDT decreased to 0.906. Because the max_depth is set up the same for objective comparison, while the GBDT usually requires a smaller max_depth than RF.

Table 4.2 Model results comparison after feature selection

Model	R^2 (Before)	R^2 (After)	R^2 (Change %)	MSE (Before)	MSE (After)	MSE (Change %)
RF	0.887	0.901	+1.58	0.107	0.093	-13.08
GBDT	0.910	0.906	-0.44	0.084	0.088	+4.76
MLP	0.884	0.889	+0.57	0.111	0.110	-0.01
MLR	0.826	0.842	+1.94	0.121	0.149	-23.1
KNN	0.758	0.814	+7.39	0.232	0.178	-23.28

5. Feature Analysis

To further study the important factors affecting house prices, the top important 20 features are selected using two methods and presented. One of the methods is Recursive Feature Elimination (RFE), and another is the feature importance method based on Random Forest. The selected top 20 features are demonstrated respectively in Table 5.1 and Table 5.2.

5.1. Feature Selection using Recursive Feature Elimination

The important features selected can be divided into four categories, including land attributes, demographics, economic factors, and road transport conditions. Half of the most influential factors are relevant to land attributes. Among them, the Floor Area Ratio (FAR) of land plays an important role in determining land value, and an increase in the mean FAR of every plot would directly result in an increase in land price (Moon, 2019). FAR of three different land-use types regulates the development density for commercial, residential, and facility plots. Demographic and economic-related factors also have a high correlation with land value. For instance, the land value of the corresponding parcels could be affected by the total number of noninstitutionalized Asians and Whites. Besides, the proportion of pregnant women in a given age group with a high level of education is another influencing factor. The accessibility of plots may also affect the value of the land.

Table 5.1 Selected 20 Features

Category	Feature
Land Attribute	Builtfar
	Commfar
	Facilfar
	Residfar
	Bldgdepth
	Year built
	Landuse_01
	Landuse_02
	Schooldist_02
	Schooldist_03
Demographic	Popunder18_percent
	Total Civilian Noninstitutionalized Population, Asian Alone
	Total Civilian Noninstitutionalized Population, White Alone, Not Hispanic Or Latino
	Fertility: Women 15 To 50 Years, Graduate Or Professional Degree
Economic	Monthly Housing Costs 3,000 Or More
	Median Household Income, Asia Alone
	Earning $75,000 Or More
	Earning $75,000 Or More, Walk To Work
Transportation	Betweenness
	Road Density

5.2. Feature Importance

The feature importance measures of Random Forest provide a method for calculating the score of the feature and a relative ranking of all features. Table 5.2 demonstrates the top 20 important features and the corresponding scores. Those 20 variables are essentially the same as those selected by RF-RFE. The features related to land attributes have the largest impact on land price compared to other categories of variables, followed by the total population with certain characteristics. Specifically, the feature that has the greatest impact on land price is schooldist_02, and Figure 5.4.3 demonstrates that land in school district 2 is located in Manhattan with the highest land values (Ma *et al.*, 2020). In summary, the features selected by RFE and feature importance is similar, which means both methods are reliable for selecting important features.

Table 5.2 Top 20 Features with Highest Score Calculated by Feature Importance

Index	Features	Importance
1	Schooldist_02	0.46082
2	Landuse_01	0.17954
3	Builtfar	0.04913
4	Earning $75,000 Or More, Walk To Work	0.02786
5	Commfar	0.01976
6	Schooldist_03	0.01823
7	Fertility: Women 15 To 50 Years, Graduate Or Professional Degree	0.01201
8	Total Civilian Noninstitutionalized Population, White Alone, Not Hispanic Or Latino	0.01146
9	Civilian Noninstitutionalized Population, Private Health Insurance Alone	0.01062
10	Landuse_02	0.00828
11	Residfar	0.00661
12	Earning $75,000 Or More	0.00643
13	Total Civilian Noninstitutionalized Population, Asian Alone	0.00577
14	Median Household Income,Asia Alone	0.00572
15	Facilfar	0.00555
16	Bldgdepth	0.00500
17	Yearbuilt	0.00472
18	Schooldist_01	0.00441
19	Betweenness	0.00430
20	Road_density	0.00417

5.3. Spatial Characteristics of Target Variable - Land Price

As for the spatial distribution of the target variable, it can be seen from Figure 5.1 that places with high land prices are clustered in the Manhattan borough. In addition, the Hot Spot Analysis of the land price converted by taking log is performed in ArcGIS in order to identify the spatial clusters of high values and low values regarding land price, as shown in Figure 5.2. There is no doubt that the high values are concentrated in Manhattan, and an extremely small number of parcels are in the surrounding boroughs, while the rest of the land is the cold spot for the land price.

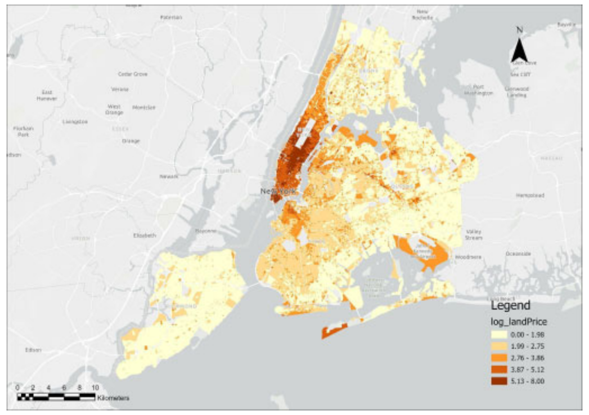

Figure 5.1 The Distribution of Land Price

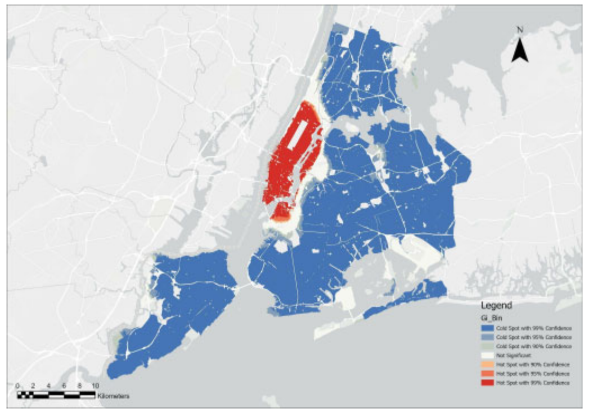
N
Legend
Gi_Bin
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
0 2 4 6 8 10
Kilometers

Figure 5.2 Hot Spot Analysis of Log Land Price

5.4. Important Features Analysis

5.4.1. Floor Area Ratio (FAR)

Figure 5.3 shows the spatial distribution of four different types of FAR in New York, including built FAR, residential FAR, commercial FAR, and facility FAR, respectively. After comparing Figure 5.1 and Figure 5.3 (a), it can be seen that the land may be more valuable with an increase in built FAR and vice versa. Figure 5.3 (b) presents the local bivariate relationship between built FAR and land value, and the result shows most areas, particularly in plots of eastern NYC with lower land prices, have a positive linear relationship with land price. From these four graphs shown below, it can be seen that the overall distribution of built FAR is similar to the distribution of residential and facility FAR. The distribution of residential FAR and facility FAR are even more similar, which indicates that land with higher residential FAR has been built with more service facilities accordingly. While high commercial FAR lands are primarily on the south side of Manhattan borough, high residential and facility FAR result in higher land values on the north side of Manhattan.

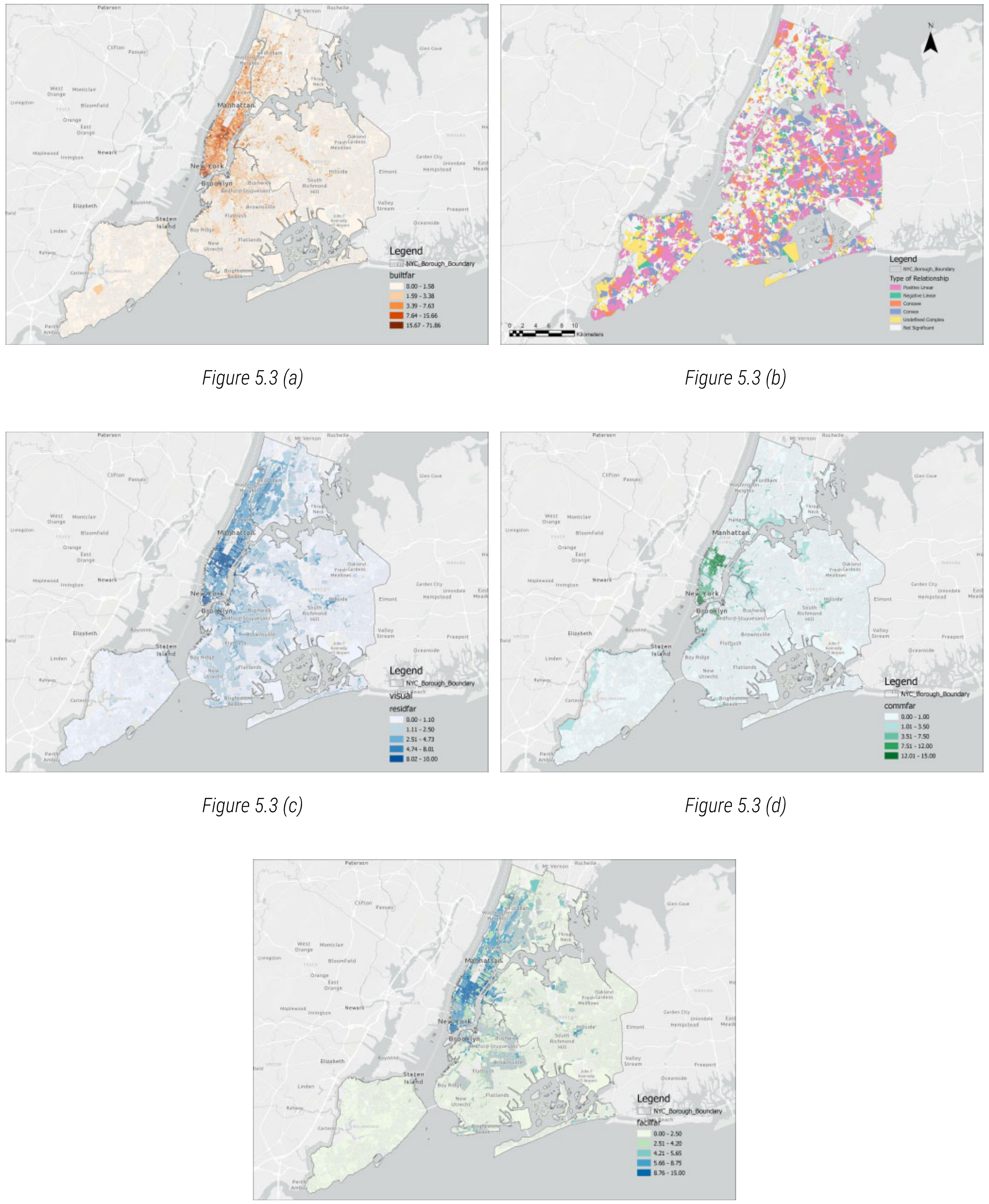

Figure 5.3 (a)

Figure 5.3 (b)

Figure 5.3 (c)

Figure 5.3 (d)

Figure 5.3 (e)

Figure 5.3 (a) The Distribution of Built FAR; (b) Relationship between Land Price and Built FAR; (c) The Distribution of Residential FAR; (d) The Distribution of Commercial FAR; (e) The Distribution of Facility FAR

5.4.2. Road Network

The accessibility of land and traffic conditions are also factors that affect land value, and Figure 5.4 demonstrates areas in Brooklyn and Queens that have higher road density and greater traffic capacity. With the exception of Manhattan, which has the highest land prices, the land price would increase when traffic capacity is increased.

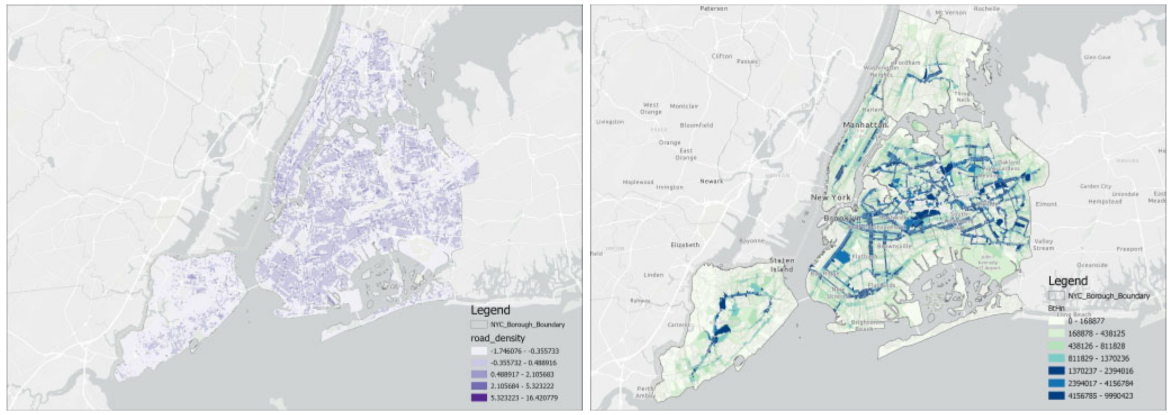

Figure 5.4 (a)

Figure 5.4 (b)

Figure 5.4 (a) Road Density Map; (b) Potential Spatial Flows (Betweenness)

5.4.3. School District

The third factor selected is three crucial school districts, their locations are shown in Figure 5.5. It can be seen that the important school districts related to land value are located in the southern part of Manhattan, where the land price and the residential FAR are also the highest.

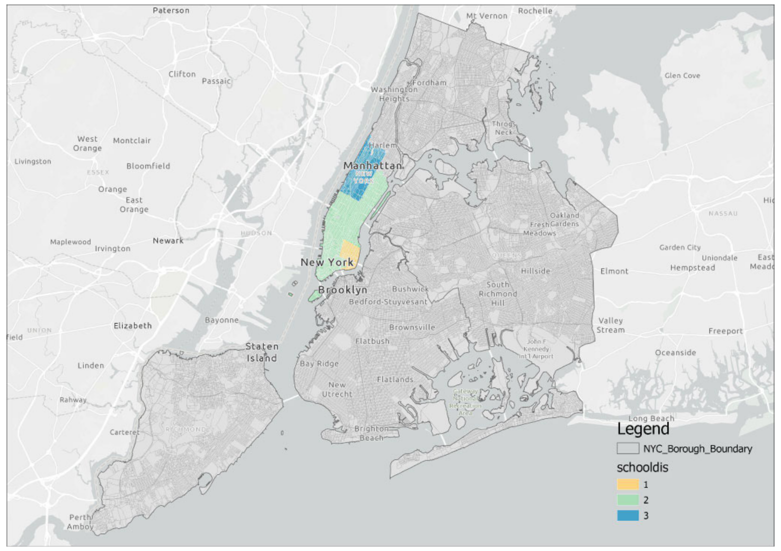

Figure 5.5 The Distribution of Important School Districts

5.4.4. Building Classification

Figure 5.6 demonstrates the distribution of different types of residential buildings, with yellow area indicating one- or two-family buildings and blue area indicating multi-family walk-up buildings. The value of the land on which one- or two-family houses are located is lower than that of the land on which the multi-family apartment building is located, which may be due to the geographical location. Such Manhattan and the north side of Brooklyn, are highly developed areas with limited land, further increasing the land price.

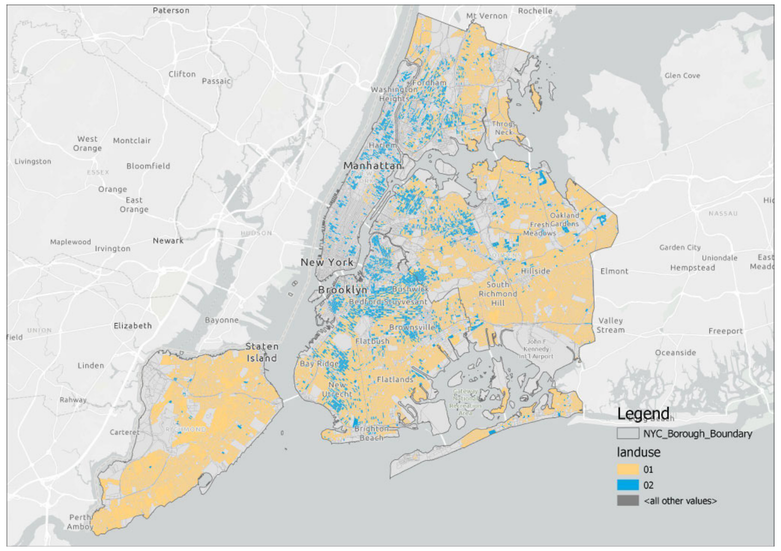

Figure 5.6 The Distribution of Two Kinds of Building Types

6. Conclusion and Limitation

The study collects hundreds of diverse data to measure their influence on land price, including land attributes, demographic, economic, public safety, transportation, and POI. Five non-linear machine learning models are applied to predict the land value in NYC, including Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Multiple Linear Regression (MLR), K-Nearest Neighbors (KNN), and Multiple Layer Perceptron (MLP). Among these models, GBDT performs best, while the result of KNN is the worst.

In addition, there are two ways to assess the importance of each feature. Recursive Feature Elimination (RFE) is applied to choose the most influential features to land price, and Feature Importance is also used to calculate the score of every feature. After selecting the important features, most models' performances have improved. The important 60 features include the characteristics of each category. And the features related to the land attribute have more influence on land price, such as FAR. Some interesting but unexpected features are also uncovered, such as fertility, and the distribution of Asians, inferring the

demographic factor also correlated to spatial differences in the land price. With the explosive growth of all kinds of city-related data and the practical application of machine learning in various fields, there are more new perspectives and directions for urban studies, such as land value prediction, which would be very meaningful.

In terms of the potential limitations, the median value is used as the value of a continuous variable at the census block level, and the mode value is used to integrate the values of the categorical variables into each census block unit in the process of unifying geographic units, which result in an inaccuracy between the actual value. And RFECV did not perform the best result due to the limited computer capacities. In addition, the study didn't conduct parameter tuning for all models. Those limitations should be improved in further research.

Word Count: 4849

References

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Breiman, L.(2001). Random forests. Mach. Learn. 45 (1), 5-32. https://doi.org/10.1023/A:1010933404324.Bourassa, S. C., Hoesli, M., Merlin, L., & Renne, J. (2021). Big data, accessibility and urban house prices. Urban Studies, 58(15), 3176–3195. https://doi.org/10.1177/0042098020982508

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. ISPRS International Journal Of Geo-Information, 7(5), 168. https://doi.org/10.3390/ijgi7050168

Cooper, C. H. V. (2015). Spatial localization of closeness and betweenness measures: A self-contradictory but useful form of network analysis. International Journal of Geographical Information Science, 29(8), 1293–1309. https://doi.org/10.1080/13658816.2015.1018834

E. Simlai, P. (2021). Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data. Journal Of Property Research, 38(4), 305-336. https://doi.org/10.1080/09599916.2021.1890187

Hu, S., Yang, S., Li, W., Zhang, C., & Xu, F. (2016). Spatially non-stationary relationships between urban residential land price and impact factors in Wuhan city, China. Applied Geography, 68, 48-56. https://doi.org/10.1016/j.apgeog.2016.01.006

Kim, J., Won, J., Kim, H., & Heo, J. (2021). Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea. Sustainability, 13(23), 13088. https://doi.org/10.3390/su132313088

Linear regression, 2018. Wikipedia. Retrieved from. https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=863781366.

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. Land Use Policy, 94, 104537. https://doi.org/10.1016/j.landusepol.2020.104537

Mirkatouli, J., Samadi, R., & Hosseini, A. (2018). Evaluating and analysis of socio-economic variables on land and housing prices in Mashhad, Iran. Sustainable Cities And Society, 41, 695-705. https://doi.org/10.1016/j.scs.2018.06.022

Moon, B. (2019) 'The effect of FAR (floor area ratio) regulations on land values: The case of New York', Papers in Regional Science. Wiley Online Library, 98(6), pp. 2343–2354.

Noble, W. S. (2004). Support vector machine applications in computational biology. InB. Schoelkopf, K. Tsuda, & J. P. Vert (Eds.), Kernel Methods in Computational Biology (pp.71–92). MIT Press.

Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. International Journal Of System Assurance Engineering And Management, 11(S2), 208-219. https://doi.org/10.1007/s13198-020-00946-3

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26-34. https://doi.org/10.1016/j.neucom.2017.04.018

Spellerberg, I., & Fedor, P. (2003). A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. Global Ecology and Biogeography, 12(3), 177-179.

Stojić, A., Stanić, N., Vuković, G., Stanišić, S., Perišić, M., Šoštarić, A., & Lazić, L. (2019). Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. Science Of The Total Environment, 653, 140-147. https://doi.org/10.1016/j.scitotenv.2018.10.368

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal Of The Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Velo, R., López, P., & Maseda, F. (2014). Wind speed estimation using multilayer perceptron. Energy Conversion And Management, 81, 1-9. https://doi.org/10.1016/j.enconman.2014.02.017

Vespignani, A. (2009). Predicting the behavior of techno-social systems. Science, 325(5939), 425–428. https://doi.org/10.1126/science.1171990

Vychegzhanin, S. V., Razova, E. V., & Kotelnikov, E. V. (2019). What Number of Features is Optimal: A New Method Based on Approximation Function for Stance Detection Task. Proceedings of the 9th International Conference on Information Communication and Management, 43–47. https://doi.org/10.1145/3357419.3357430

Wen, H., & Goodman, A. (2013). Relationship between urban land price and housing price: Evidence from 21 provincial capitals in China. Habitat International, 40, 9-17. https://doi.org/10.1016/j.habitatint.2013.01.004