Predicting Land Values of Tokyo Metropolitan Area with Regression

Kiyohiko Nishi

March 2020

Submitted in partial fulfillment of the requirements of the Coursera course,

"Applied Data Science Capstone"

1. Introduction

1.1. Background

Owing one's own land in live is a hard decision to make. There are a lot of things to be considered from advantage of neighborhood - convenient to commute, calm and peaceful, etc. – to disadvantage of the neighborhood – high crime rate, high risk of natural hazard, and so on. Generally speaking, however, the most important factor which affects our decision is the land price.

As for land price in Tokyo metropolitan area, where it is considered expensive area to live compared to other areas in Japan, increase tendency has still been seen despite of economic stagnation that has caused drop in land price except Tokyo area.

In this circumstance, a simple tool providing support for making decisions is strongly expected.

1.2. Problem to be solved

A model which is able to predict proper land price by analysis of the data of land price in Tokyo metropolitan area is prepared. Then, as a verification of the model, the most affordable place to live is suggested.

2. Data

2.1. Data Sources

Following sources are used in this study.

Land Attribution Data	Data Source	Description
Area name Latitude Longitude Land price (JPY/m2) Nearest station name Distance to nearest station (m)	National Land Numerical Information: Publication of Land Price Data http://nlftp.mlit.go.jp/ksj/index.html Provided by Ministry of Land, Infrastructure, Transport and Tourism.	Various GIS data of each prefecture is provided. Shape file format, GML, and GeoJSON format is available. Dataset is able to be downloaded as zip file by each prefecture, containing all relevant files. Multiple attribution data is included in a dataset, such as regulation and land usage, other than those used in this study.
Number of stations in neighborhood	Foursquare location data https://foursquare.com/	Data is retrieved using Foursquare API endpoint at the 'search' endpoint with

	Provided by Foursquare Labs Inc.	query as “station’ and limitation in category id which is relevant to train/metro station.
Distance to Tokyo station (Km) Distance to city hall of the capital of each prefecture (Km)	Coordinates of each landmark: OpenStreetMap data Obtained by Nominatim geocorder of Geopy module https://wiki.openstreetmap.org/wiki/Nominatim Distance between two position: calculated by Haversine Formula https://en.wikipedia.org/wiki/Haversine_formula	Obtaining latitude and longitude of each landmark and calculated the distance between each position in the dataset using Haversine Formula.

2.2. Data Description

Overview of the dataset is as follows. In total, 1,019 points of data is used to analize.

	City	Latitude	Longitude	Stations	Station	DistToSt	ToCityHall	ToTokyoST	LandValue
0	武蔵野	35.710699	139.584039	0	吉祥寺	960	9.998406	17.018053	560000
1	調布	35.670318	139.586601	0	仙川	1000	9.735226	16.498822	351000
2	三鷹	35.678447	139.587580	0	三鷹台	2000	9.487364	16.371411	301000
3	武蔵野	35.700362	139.587689	3	吉祥寺	900	9.467431	16.504097	653000
4	武蔵野	35.704320	139.588387	0	吉祥寺	800	9.470161	16.505738	595000

3. Method

3.1. Data Understanding

3.1.1. Description of obtained data

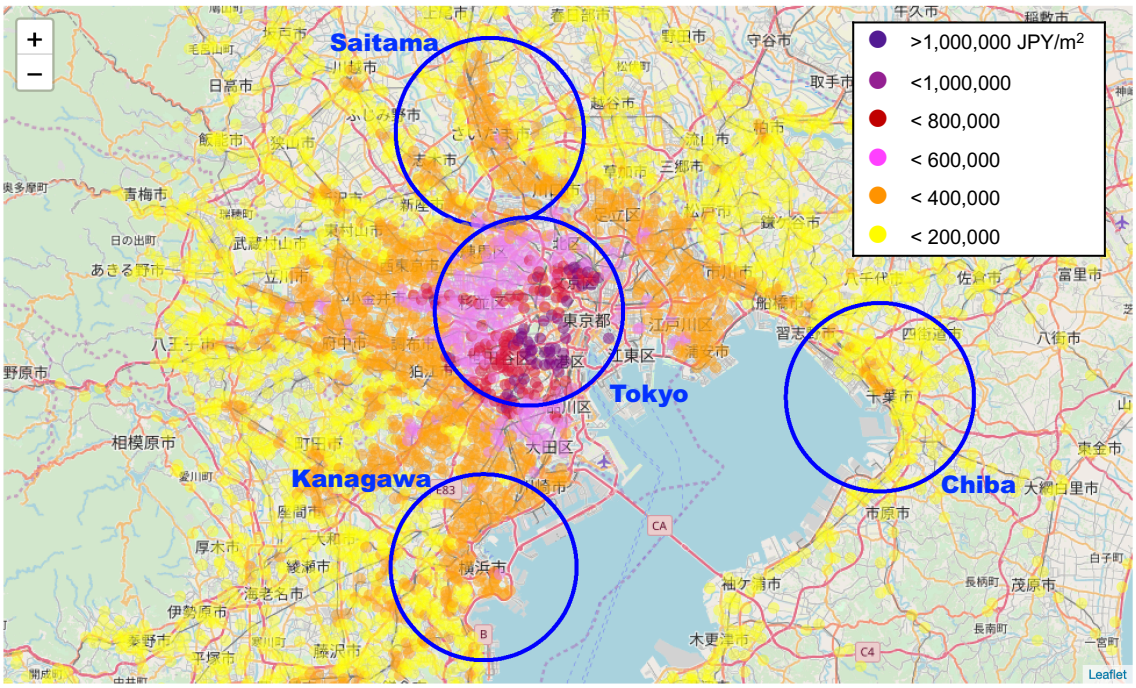
Before data cleaning, obtained land price data (JPY/m2) is summarized as shown below.

	Saitama	Chiba	Tokyo	Kanagawa	Total
Conunt	1,017	949	1,369	1,296	4,631
Min	10,300	5,100	5,700	18,300	5,100
Max	515,000	349,000	1,920,000	667,000	1,920,000

3.1.2. GIS Map

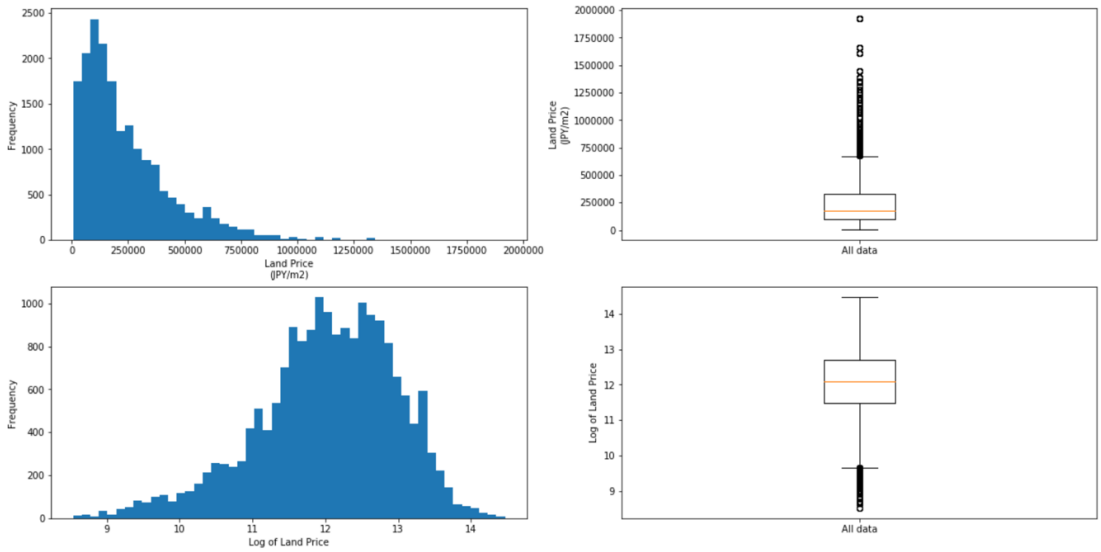
Land price data is visualized on a map with Folium module. Ranked by land price, it is indicated by marker color. At a single glance, it is found that land price is getting lower along with the distance from center of Tokyo.

The center of four circle on the map is the city hall of the capital of each prefecture considered in this research, and the circle indicates 10 km distance range. It seems there are some tendency that land price is higher at the center of each area.



3.1.3. Data distribution

As shown in the graphs below, distribution of land price is strongly skewed to right. Then taking logarithms of land price, it looks like normal distribution. Therefore, for further analysis, log transformed land price data will be taking into account.



3.2. Data Preparation

3.2.1. Distance data

Since the distance from Tokyo and the capital city hall of each prefecture seems to affect the land price formation, distance data is calculated. Setting Tokyo station as the landmark of Tokyo, coordinate data of each landmark id obtained with *Geopy* module. Then, distance between each point is calculated with Haversine Formula. Obtained data is combined to the other dataset as follows.

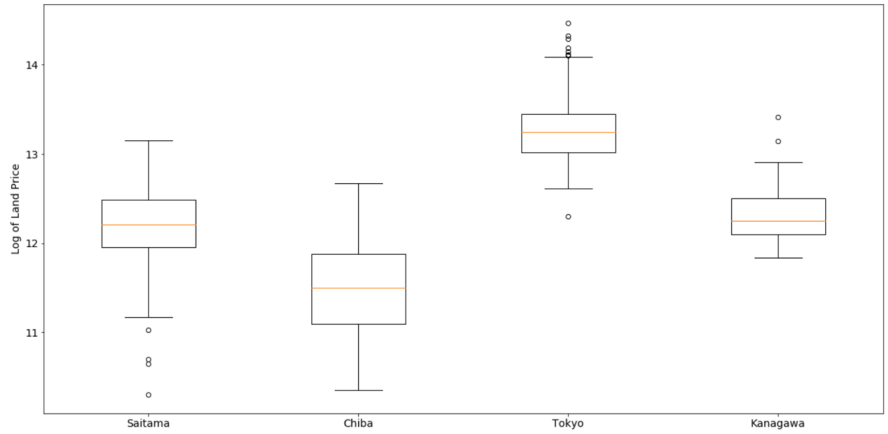
	Saitama City Hall	Chiba City Hall	Tokyo City Hall	Yokohama City Hall	Tokyo Station
Latitude	35.8618	35.6047	35.6896	35.4438	35.6812
Longitude	139.6452	140.1065	139.6921	139.6374	139.7671

	City	Station	DistToSt	LandValue	Latitude	Longitude	ToCityHall	ToTokyoST
0	小鹿野	秩父	14000	25700	36.022335	139.003067	60.542365	78.802873
1	小鹿野	秩父	12000	21500	36.012412	139.021927	58.599773	76.784894
2	秩父	御花畑	900	60100	35.989768	139.077127	53.131900	71.218056
3	秩父	御花畑	750	59900	35.995823	139.077471	53.284929	71.516544
4	秩父	秩父	800	53200	36.003158	139.081972	53.129681	71.566058

3.2.2. Filtering

Narrowing the points of interest within 10km distance from the city hall of each capital, data set is filtered by the value of distance to the city hall. As a result, description of dataset filtered and comparison among each prefecture is shown in a box plot graph as below. It clearly shows Tokyo is the most expansive, and Chiba seems to be affordable compared to other prefectures.

	City	Station	DistToSt	LandValue	Latitude	Longitude	ToTokyoST	ToCityHall
count	267	267	267.000000	267.000000	267.000000	267.000000	267.000000	267.000000
unique	20	50	NaN	NaN	NaN	NaN	NaN	NaN
top	川口	蕨	NaN	NaN	NaN	NaN	NaN	NaN
freq	63	22	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	1395.955056	210765.917603	35.856829	139.650725	22.810849	6.231632
std	NaN	NaN	870.202274	76077.778756	0.041599	0.054207	4.743259	2.612952
min	NaN	NaN	300.000000	29800.000000	35.785634	139.537278	13.963422	0.288330
25%	NaN	NaN	750.000000	155000.000000	35.824432	139.610477	18.850386	4.086466
50%	NaN	NaN	1100.000000	201000.000000	35.850724	139.652052	22.404159	6.692085
75%	NaN	NaN	1800.000000	265000.000000	35.885886	139.694676	26.290577	8.312827
max	NaN	NaN	5600.000000	515000.000000	35.949160	139.753421	32.532434	9.982989



3.2.3 Additional attribute

It's well known fact that proximity to train/metro station is one of the factors which adds premium to land value; that is why such data is included in the original data set publicly provided. In relation to this, the number of available stations in its neighborhoods is thought to be worth to consider. Therefore, using Foursquare API 'search' endpoint, number of results is counted in each point when query is 'station' and category id is limited to train/metro/tram station.

3.2.4 Dataset for analysis

From the dataset summarized in 2.2 section, geometry data and qualitative data is removed. Then, log transformed value of land price is added to the dataset. Concatenating the dataset of each prefecture, a data frame for analysis is prepared as below.

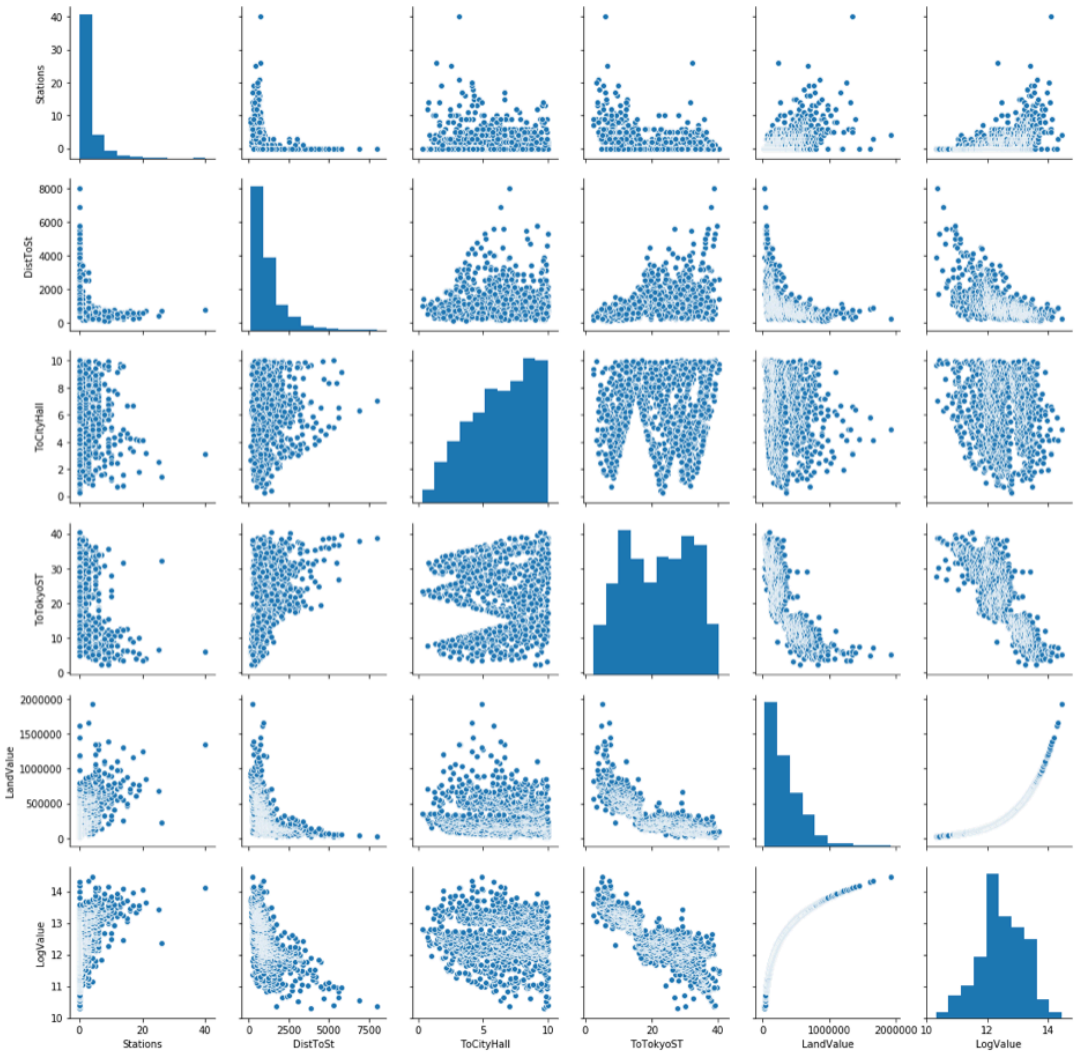
	Stations	DistToSt	ToCityHall	ToTokyoST	LandValue	LogValue
0	0	3900.0	9.679512	27.572543	29800.0	10.302264
1	0	8000.0	7.039010	38.625657	31500.0	10.357743
2	0	1700.0	9.821689	30.219071	32000.0	10.373491
3	0	5300.0	9.997029	38.890954	32800.0	10.398184
4	0	3100.0	9.840500	37.759704	36800.0	10.513253
...	...	...	...	...	...	...
1014	9	310.0	6.370946	5.811658	1390000.0	14.144814
1015	0	720.0	4.176302	7.417678	1450000.0	14.187074
1016	0	840.0	5.832020	4.778504	1610000.0	14.291745
1017	3	900.0	4.124020	7.274164	1660000.0	14.322328
1018	4	250.0	4.901122	5.065084	1920000.0	14.467836

1019 rows × 6 columns

3.3. Modeling

3.3.1. Correlation of each variables

Scatter plot matrix is shown with *pairplot* object of *seaborn* module. It is suggested that there are negative correlation between land price and distance to Tokyo/nearest station.

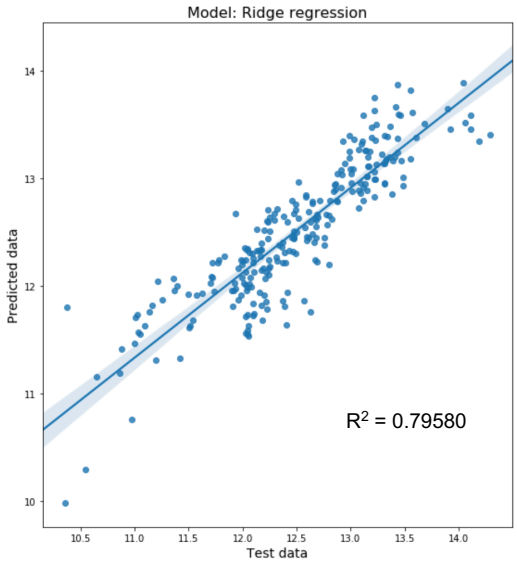


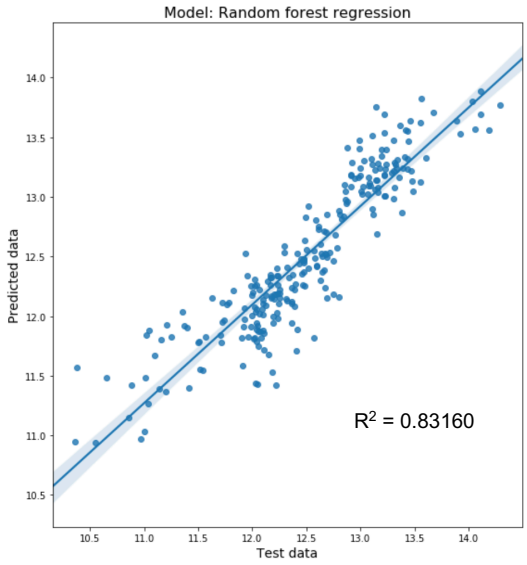
3.3.2. Machine learning

Two methods of machine learning, ridge regression and random forest regression, are considered to conduct regression analysis with relevant object in *scikit-learn* module. Parameters of each object are set to default value except *n_estimator*s object which is set to 100. Data splitting is conducted with *train_test_split* object in *scikit-learn* module, and the ratio of test data and train data is set to 1:3.

3.3.3. Evaluation

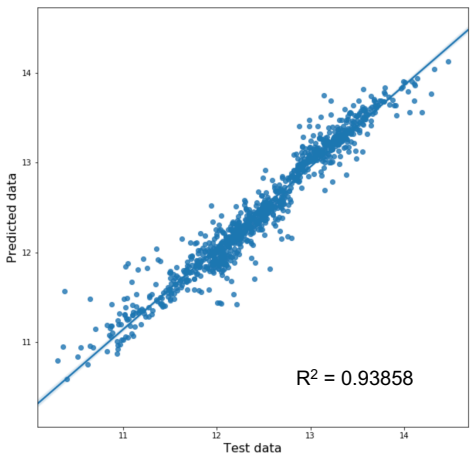
Using a model built with train data, prediction as for test data variables is executed. The predicted R Squared value is similar between two models, and both models are considered well created to conduct further prediction. But random forest regression is chosen as a model because it provides better prediction in both end of lower and higher of price data.





3.4. Deployment

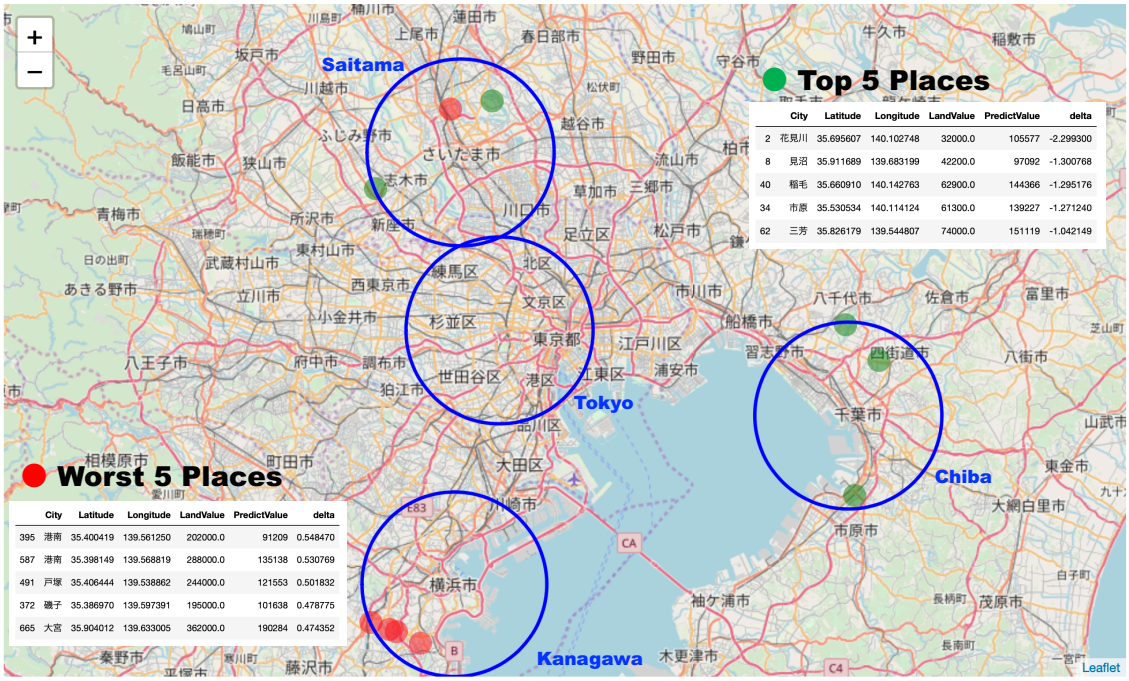
Land price prediction is conducted with random forest regression described above, and exponential conversion is conducted to obtained predicted data for comparison to original price date. Difference between original price value and predicted value is described as relative change to original price value. Final dataset obtained is shown below.



	City	Latitude	Longitude	Stations	Station	DistToSt	ToCityHall	ToTokyoST	LandValue	PredictValue	delta
2	花見川	35.695607	140.102748	0	八千代台	1700.0	9.821689	30.219071	32000.0	105577	-2.299300
8	見沼	35.911689	139.683199	0	大宮	5600.0	6.530860	26.834771	42200.0	97092	-1.300768
40	稲毛	35.660910	140.142763	0	四街道	2300.0	6.796555	33.866578	62900.0	144366	-1.295176
34	市原	35.530534	140.114124	0	八幡宿	1000.0	8.572809	35.415164	61300.0	139227	-1.271240
62	三芳	35.826179	139.544807	0	みずほ台	1600.0	9.887371	25.899016	74000.0	151119	-1.042149
...	...	...	...	...	...	...	...	...	...	...	...
665	大宮	35.904012	139.633005	0	大宮	1100.0	4.847476	27.701503	362000.0	190284	0.474352
372	磯子	35.386970	139.597391	0	洋光台	1200.0	7.326629	36.159718	195000.0	101638	0.478775
491	戸塚	35.406444	139.538862	0	戸塚	920.0	9.906969	36.919170	244000.0	121553	0.501832
587	港南	35.398149	139.568819	5	上永谷	600.0	8.078220	36.258996	288000.0	135138	0.530769
395	港南	35.400419	139.561250	3	下永谷	450.0	8.475124	36.389084	202000.0	91209	0.548470

4. Result

Top 5, where land price is under-evaluated and the suggested places to own land to live, and Worst 5, where land price is over-evaluated, are summarized in the table below and marked on the map.



Top 5 Places

	City	Latitude	Longitude	LandValue	PredictValue	delta
2	花見川	35.695607	140.102748	32000.0	105577	-2.299300
8	見沼	35.911689	139.683199	42200.0	97092	-1.300768
40	稲毛	35.660910	140.142763	62900.0	144366	-1.295176
34	市原	35.530534	140.114124	61300.0	139227	-1.271240
62	三芳	35.826179	139.544807	74000.0	151119	-1.042149

Worst 5 Places

	City	Latitude	Longitude	LandValue	PredictValue	delta
395	港南	35.400419	139.561250	202000.0	91209	0.548470
587	港南	35.398149	139.568819	288000.0	135138	0.530769
491	戸塚	35.406444	139.538862	244000.0	121553	0.501832
372	磯子	35.386970	139.597391	195000.0	101638	0.478775
665	大宮	35.904012	139.633005	362000.0	190284	0.474352

5. Discussion

Four out of five in Top 5 is located in Chiba prefecture and four out of five in Worst 5 is in Kanagawa prefecture. From various observation through this research, it could be generally said that Chiba prefecture id under evaluated and Kanagawa prefecture is over evaluated.

6. Conclusion

Publicly disclosed land price data in Tokyo Metropolitan Area is examined in order to suggest the most affordable place to own land to live. Comparison to predicted land price data obtained by machine learning model with random forest regression shows Top 5 under evaluated places, and it is found out that Chiba prefecture is generally under evaluated and affordable.

EoD