The land price increase prediction in Osaka city in Japan

Takahiro Yoshii 8/11/2020

**Summary**

In this report, I investigate the land price increase over the last five years with geographical features, accessibility, and venue information using Random forest regression. I found there are 4 factors to increase the land price, accessibility, city center, number of venues, and wealthy neighborhood. And I got high predictability (0.7222 ).

**Introduction**

During the COVID pandemic, there are many people lost their job or faced decrease of their salary. In order to increase one's income, real estate investment is one option. Under COVID pandemic situation, some people decide to release their property due to financial difficulties so that this is the best timing to invest on real estate! When you are really investing on real estate, you would like to know where will rise its value most. Here we focus on Osaka city because it is expected to grow where Olympic and Expo are scheduled. Originally Osaka area is well known for its historical and commercial value in Japan.

**Data**

Data source(sorry some data sources are in Japanese)

1.Land price data in 2020,2015

(https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L01-v2\_5.html)

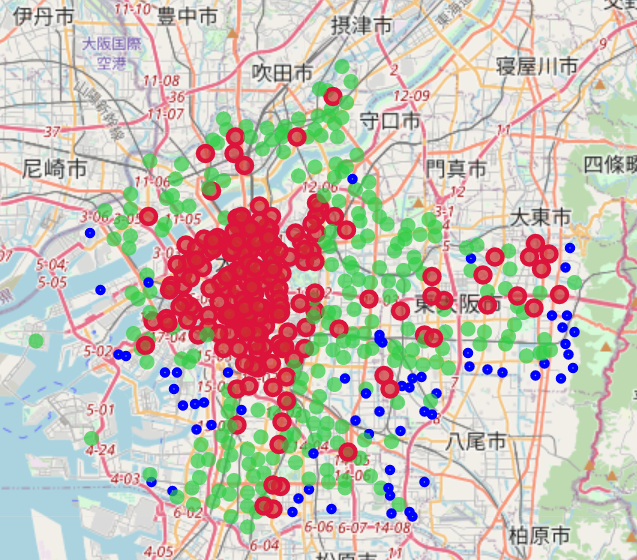
2.venue information(https://ja.foursquare.com/city-guide)

My target is to predict the price change ratio during 2015 to 2020 which I obtained information from Data.1 Land price is mainly determined by its environment.　For venue information I obtained from Data1 and Data2 By using Foursquare API, I obtain information of venues around the location. (in this case we chose 500m which is around 10 min walk). And additional information such as distance from the closest station is obtained from Data1

Data cleansing

In the data cleansing section, I got one hot coding based on venue categories and the total number of venues registered in "Foursquare" because the land price and the number of venues seem to be connected. Also the kinds of shops should be related to the land price.

Data exploration1(Geographical insight)



Land price change

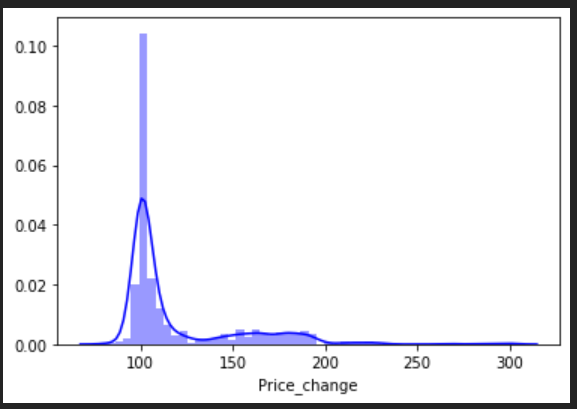
●<100%

●100%<105%

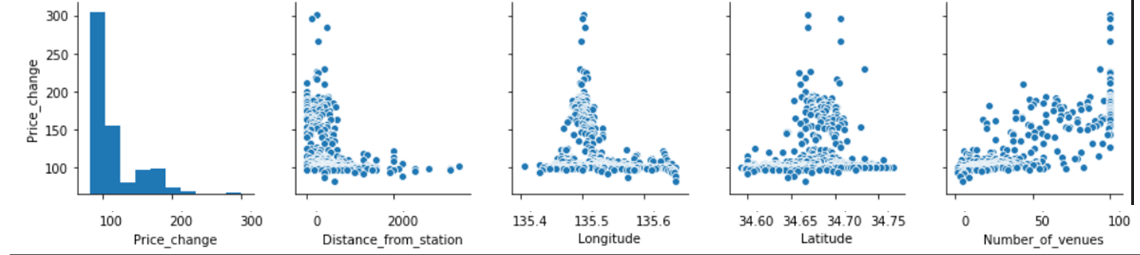
●105%<

I plotted the location as follows: Blue circle: The land price is the same or cheaper than 2015 Green circle: The land price is the same or up to 5% increase compared to 2015 which is about 75 percentile of the whole land price range. Red circle: The land price is more than a five percent increase compared to 2015 I notice the center of the city is mainly red and surrounded by green circles. I assume the city center is growing so that its land price increased. And people chose to spread around for a more reasonable price so that the land price outside of city center areas increased as well. Blue circle areas are mainly on the southeast and west side. The west side is mainly the port area which suggest fishery and shipping losing their importance. The decrease in the land price in the southeast suggests the city is expanding to the north side where is flat land.

Data exploration2(Numerical insight)



This is the histogram of the land price change from 2015. There are two peaks. One is around 103% which means the land price slightly increased the same as five years ago. Another is around 170% which means the land price increased by 70% since 2015.



This is a pair plot of the land price change, distance from station, Longitude, Latitude, and the number of venues. From the pair plot, I noticed three things.

1.The land price tends to increase when the distance from a station.

People prefer the place close to a station. However, it has a limited effect on the land price increase possibly because the size of the station matters. People prefer the big station over the small if the land price is the same.

2.City center tends to increase the land price.

According to the latitude and the longitude versus price change plot, locations that have the middle in latitude and longitude tend to have a higher land price increase that coincides with the folium plot I presented previously Locations with more venues tend to have a higher land price change.

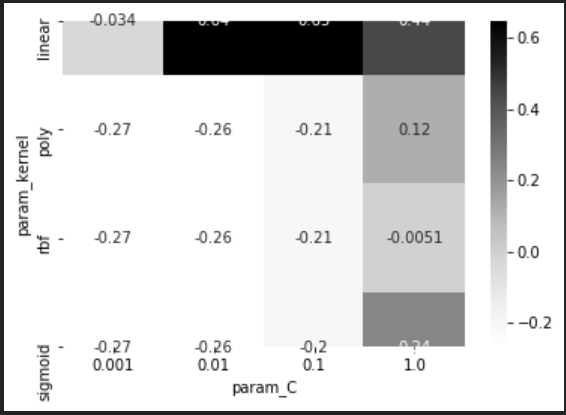
3.The relation between a price change and the number of venues seems to have exponential relation.

I assume a bigger price change means more people stay or visit the location. There are two possibilities, one is there are originally some attractive sites which became more popular in 5 years so more people visit as a result, the land price increased. Another is the more people attracted venues to open around the location.

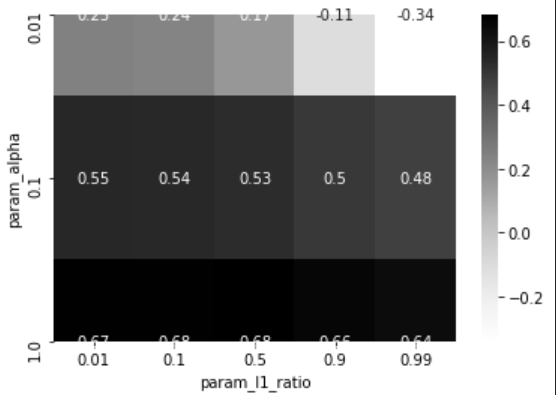
**Methodology**

In this report, I compared three major algorisms, Support vector machine(SVM), Elastic net regression, Random forest regression from sklearn. As preprocessing, I used a standard scaler to standardize the explanatory variables from sklearn. Hyper parameters were optimized using GridsearchCV form sklearn.

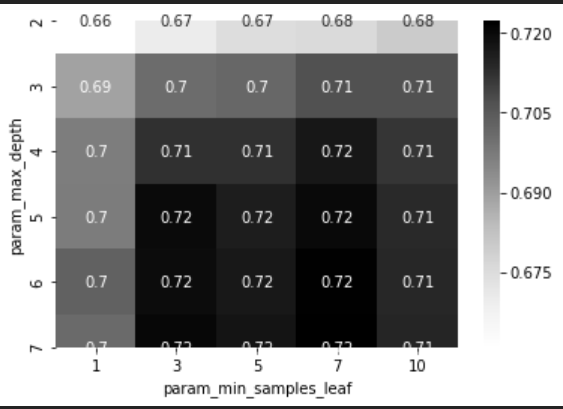
**Results**



This shows heat map of hyperparameter optimization for SVM. The test set score is 0.6201. The best parameters are C equals 0.1 and kernel is linear. And the best cross-validation score is 0.6469.



This shows heat map of hyperparameter optimization for Elastic net regression. The test set score is 0.6976. The best parameters are alpha is 1, L1 ratio is 0.5. And the best cross-validation score is 0.6810



This shows heat map of hyperparameter optimization for Elastic net regression. The test set score is 0.7016. The best parameters are max depth is 7, minimum samples per leaf is 7. And the best cross-validation score is 0.7222

|  |  |  |  |
| --- | --- | --- | --- |
|  | SVM | Elastic net | Random forest |
| the best cross-validation score | 0.6469 | 0.681 | 0.7222 |

Table. Comparison of three algorithms

As a result, the random forest regressor achieved the best cross-validation score among the three. So I adopted the random forest regressor as a prediction algorithm. For further detail. I checked all feature importance.

|  |  |  |
| --- | --- | --- |
| Feature | importance | Accumulated importance |
| Distance\_from\_station | 0.687 | 0.687 |
| Longitude | 0.048 | 0.735 |
| Latitude | 0.047 | 0.782 |
| Number\_of\_venues | 0.041 | 0.823 |
| ATM | 0.025 | 0.848 |
| American Restaurant | 0.024 | 0.872 |
| Antique Shop | 0.014 | 0.886 |
| Arcade | 0.012 | 0.898 |
| Art Gallery | 0.007 | 0.905 |
| Art Museum | 0.006 | 0.911 |

Table. The importance of top 10 features

In the result section, I conclude Random Forest regression was the best algorithm and four factors affect the land price increase.

The biggest is the distance from the closest station implies accessibility to the location is the biggest concern when you choose the land. Having better accessibly attracted more people to visit and business for over the last five years.

2nd is longitude and latitude. it means the city center is the key factor as well as we can see from the folium map. There is a tendency for people to concentrate on originally popular places to visit or invest.

3rd factor is the total number of venues. It implies the land price increase occurred in commercial areas where more people register venue information to folium. ATM and arcade appeared in the top 10 features that support this idea. I should mention that the land price increase occurred, more specifically, in the shopping area.

An interesting fact is that American restaurants, antique shops, and art-related facilities are also some impacts on the land price increase. I think those venues are for wealthy people. It implies the concentration of wealth happened, in short, rich people became richer and the land price increased in some way.

**Conclusion**

In this report, I investigate the land price increase over the last five years with geographical features, accessibility, and venue information using Random forest regression. I found there are 4 factors to increase the land price, accessibility, city center, number of venues, and wealthy neighborhood. And I got high predictability( 0.7222 ). I will use this insight to invest next time.