IBM – Coursera

Capstone project

Analysis on real estate price in Shanghai

Frank LI – 2019

I. Introduction:

This is a report for Data Science Specialization. I am going to explore the neighborhoods of Shanghai in order to extract the correlation between the real estate value and its surrounding amenities. The owners or agents advertise properties are closed to convinent facilities like café bar, restaurants or supermarkets, etc.; Can the surrounding venues affect the price of a house? If so, what types of venues have the most affect, both positively and negatively? The target audience for this report are: - Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price. - Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price. - Apartment sellers who can better tail their advertisements.

2. The question to solve: This project will try to explore the neighborhoods of Shanghai to see: - if the surrounding venues can effect the price of real estates? - what kind of surrounding facilities, and to what extend, can effect the price? - if we can use the surroundings to estimate the value of an accommodation over the average price of one area? And to what degree of confidence? The result can be useful for real-estate agent, potential home buyers, who can roughly estimate the value of a target house over the average.

II. Data description:

The main data used for this project as below:

- The average price by neighborhoods in Shanghai.
- The venues in each neighborhood. ([FourSquare API]
- Coordinates
- GeoJson

Data collection process:

- The average price will be scrapped from the realestate agent's website.
- Use Geocoder Python to get its coordinate.
- Use FourSquare API to get the surrounding venues.

The output of the data collecting process will be a 2 dimensions dataframe:

- Each row represents a neighborhood.
- Each column will be the count of one type of venue in that neighborhood.

Using data to solve the question:

- First, correlation between price and surrounding venues will be checked.
- Second, if correlated, machine learning techniques (PCA, Regression, PCR) will be used to analyze the data. The output will be a list of venues types that effect the most on price, along with their weight on the result.

III. Methodology:

The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices.

At the end, a regression model will be obtained. Along with a coefficients list which describes how each venue type may be related to the increase or decrease of a neighborhood's real estate average price around the mean.

Python data science tools will be used to help analyze the data. Completed code can be found here: https://github.com/lethien/coursera-ibm-ds-capstone/blob/master/Capstone_Analyze.ipynb

1. First insight using visualization:

In order to have a insight of Shanghai real estate average price between neighborhoods via visualization.

2. Linear Regression:

Linear Regression was chosen because it is a simple technique. And by using Sklearn library, implementing the model is quick and easy. Which is perfect to start the analyzing process.

The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

The result (Figure 3) doesn't seem very promising. R2 score is small, which means the model may not be suitable for the data.

```
R2-score: 0.273792308888
Mean Squared Error: 0.254179706388
Max positive coefs: [ 0.26348338  0.26213799  0.26213799  0.26213799  0.25818747  0.25818747
  0.25135936  0.24564842  0.23349638  0.22658134]
Venue types with most postive effect: ['Design Studio' 'Train Station' 'Jewish Restaurant' 'Resort' 'Buffet'
 'Cafeteria' 'Colombian Restaurant' 'Dumpling Restaurant' 'Other Nightlife'
 'Botanical Garden']
Max negative coefs: [-0.20813947 -0.20763403 -0.1798399  -0.1798399  -0.1798399  -0.17776278
 -0.17776278 -0.17776278 -0.17776278 -0.17776278]
Venue types with most negative effect: ['Board Shop' 'Gay Bar' 'Supplement Shop' 'Rest Area' 'Lighthouse' 'Office'
 'Flea Market' 'Golf Driving Range' 'Recreation Center'
 'General Entertainment']
Min coefs: [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
Venue types with least effect: ['TV Station' 'Gas Station' 'Pakistani Restaurant' 'Volleyball Court'
 'Hookah Bar' 'Indoor Play Area' 'Laser Tag' 'Christmas Market' 'Cemetery'
 'Mini Golf']
```

Looking back further to the dataset, its dimensions sizes is clearly unbalanced, only 50 samples, and more than 300 features. Logical steps to take are either collecting more samples or trying to reduce the number of features.

But since there are no other public source available, increasing sample size is not possible at the moment. So, deceasing features is the only option for now.

And that's why Principal Component Regression is chosen to analyze the dataset in the next part.

3. Principal Component Regression (PCR):

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression.

PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.

Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

Again, R2 score and MSE are used to see how well the model fit the dataset.

The result is promising as it shows improvement over the simple Linear Regression.

As for the coefficient list, the size has been reduced after performing PCA. So, a dot product with eigenvectors is needed to get it back to the original features size.

The insight is still consistent compared to the Linear Regression's.

IV. Results:

Even though the scores seem to be improved after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict a neighborhood average price.

Explanations for the poor model can be:

- The real estate price is hard to predict.
- The data is incomplete (small sample size, missing deciding factors).
- The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricy neighborhoods.

V. Discussion:

The real challenge is constructing the dataset:

- Usually the needed data isn't publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

VI. Conclusion:

Doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program.