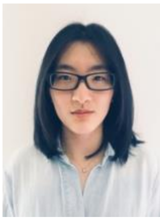
Real Estate Valuation

Math 5671 Group 1

Shengyang Ni – graduate student major in biostatistics



Presentation link:

https://kaltura.uconn.edu/media/Real+Estate+Valuation+Group1+Math+5671/1_d3qhnp6v

Project Background

The project aims to predict the house price of unit area in New Taipei City, Taiwan.

The raw data given comprises several attributes:

The inputs are as follows:

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (unit: degree)

X6=the geographic coordinate, longitude. (unit: degree)

The output is as follow:

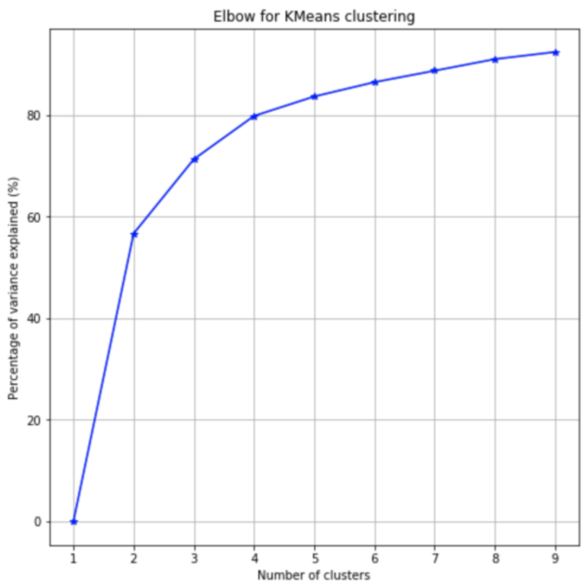
Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Methodology

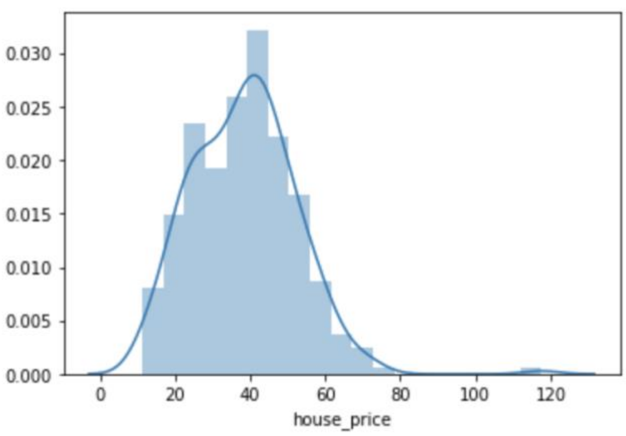
We use pandas, seaborn and matplotlib to deal with the data exploration and processing, k-means clustering to classify the geographic coordinate variables, and pyspark to predict the model.

Data exploration and processing

Since the dataset is collected only in New Taipei City, little decimal changes in geographic coordinate can differ in district so we decide to use k-means clustering to classify the geographic location and make it into a categorical variable. To find the proper number of clusters, we use Elbow method, which looks at the percentage of variance explained as a function of the number of clusters and we choose the number that didn't much better-off after adding another cluster. According what we have here, the cluster should be 4.



Then, we start analysis by visualizing the distribution of house price and to determine if there were possible outliers.



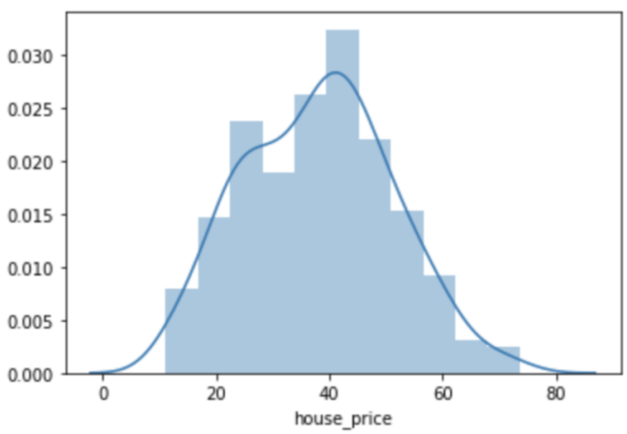
After plotting a histogram, we see there is an outlier. Since the dataset is small, we search online to see if the high price is reasonable.

中央六街實價登錄一覽表
揭露完整門牌
永慶 實價登錄3.0
列印
路段成交行情 定位點周遭500公尺成交行情
瀏覽模式：列表 地圖
排序：預設排序
單價 地址

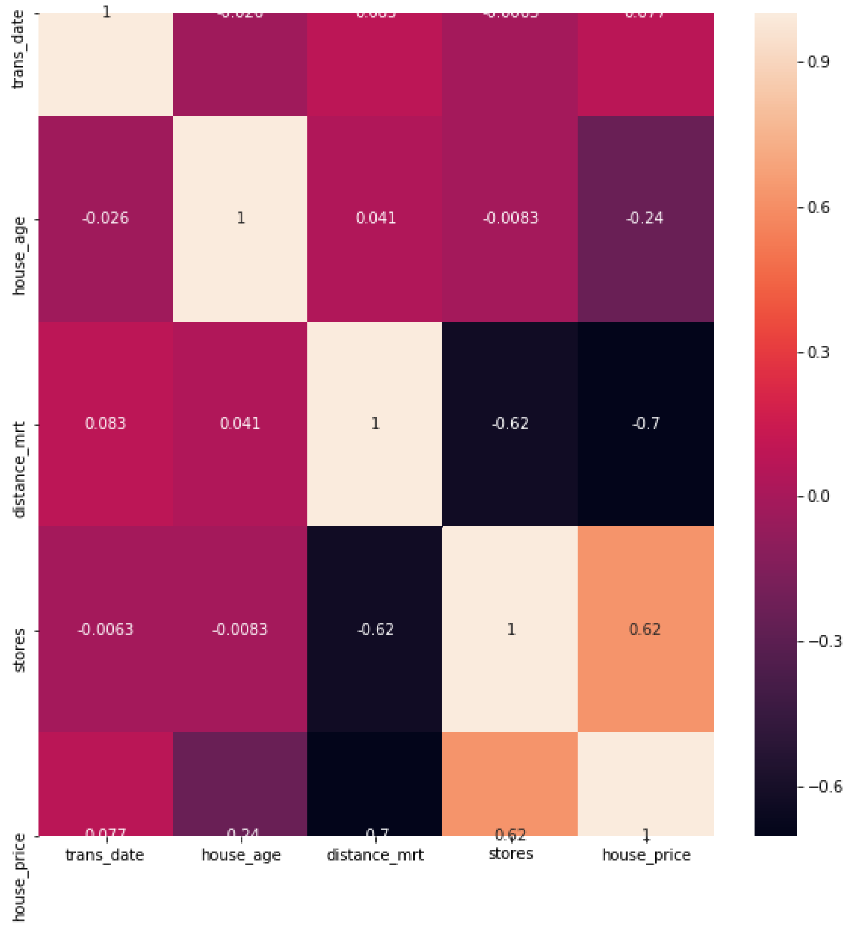
成交	型態	地址	成交價	建坪	單價	地坪	樓別	屋齡	車位
10808	華廈	新店區中央六街31~60號 格局：3房(室)2廳2衛	2,800萬 (含車位)	46.06坪 (含車位坪數)	60.8萬	11.41坪	5~5/6	0.7年	有
10805	電梯大樓	新店區中央六街1~30號 格局：4房(室)2廳2衛 (加蓋格局：0房(室)0廳1衛)	3,150萬 (含車位180萬)	71.52坪 (含車位3.69坪)	43.8萬	9.52坪	2~2/14	4.0年	有

house price

We see the sale price at that location in 5 years varies between 40-60, which is similar to other data we get, so we decide to remove this outlier. Our new histogram looks better and perform nearly normal distribution.



We check the correlation matrix to see the correlation among variables.



The correlation between transaction date and house price is only 0.077, so we further look at the transaction data.

```
count      288.000000
mean      2013.149016
std          0.282665
min       2012.666667
25%       2012.916667
50%       2013.166667
75%       2013.416667
max       2013.583333
Name: trans_date, dtype: float64
```

The data only collect the transaction in one year from 2012 to 2013, which have little influence on the house price, so we also remove this variable. After cleaning the data, we have 4 variables to predict the house price, with dist means the classification of longitude and latitude.

```
Data columns (total 5 columns):
house_age       288 non-null float64
distance_mrt    288 non-null float64
stores          288 non-null float64
house_price     288 non-null float64
dist            288 non-null int32
```

Model Predictions

We use linear regression, gradient boosting, and random forest model to predict our training dataset, with a random split of ratio 7:3 to have train and validation samples.

The prediction using linear regression and the fitness are as follows:

```
RMSE: 8.156030
r2: 0.599819
R Squared (R2) on test data = 0.614114
Root Mean Squared Error (RMSE) on test data = 8.19857

+------------------+-----------+--------------------+
|        prediction|house_price|            features|
+------------------+-----------+--------------------+
|41.795942002304294|       52.2|[0.0,185.4296,0.0...|
| 47.30183398342608|       50.7|[1.0,193.5845,6.0...|
|47.280236947940715|       49.0|[1.1,193.5845,6.0...|
| 47.09175372301922|       50.4|[1.7,329.9747,5.0...|
| 52.92946898168616|       61.9|[3.6,373.8389,10....|
|35.488085237206505|       31.7|[3.9,2147.376,3.0...|
|35.275473090536394|       28.6|[4.0,2180.245,3.0...|
| 45.98372287570154|       52.2|[5.2,390.5684,5.0...|
| 52.71420453572941|       58.0|[6.2,90.45606,9.0...|
| 45.22076980426108|       57.1|[7.1,451.2438,5.0...|
| 44.98375424112128|       51.6|[8.1,104.8101,5.0...|
| 44.89736609917985|       56.8|[8.5,104.8101,5.0...|
|32.782131031328404|       38.5|[9.0,1402.016,0.0...|
|  44.6447066756093|       46.8|[11.4,390.5684,5....|
| 21.87362042141281|       46.6|[11.9,3171.329,0....|
| 31.64386638002354|       28.9|[12.0,1360.139,1....|
|38.725514098814536|       34.1|[12.5,1144.436000...|
| 35.85022286840146|       40.6|[12.8,732.8528,0....|
| 41.69563456473903|       42.5|[12.9,492.2313,5....|
|  41.6092464227976|       31.3|[13.3,492.2313,5....|
+------------------+-----------+--------------------+
only showing top 20 rows
```

The prediction using gradient boosting and the fitness are as follows:

```
Root Mean Squared Error (RMSE) on train data = 3.57468
Root Mean Squared Error (RMSE) on test data = 5.78077
```

```
+------------------+-----------+--------------------+
|        prediction|house_price|            features|
+------------------+-----------+--------------------+
| 47.27735896466555|       52.2|[0.0,185.4296,0.0...|
|45.477969888250485|       50.7|[1.0,193.5845,6.0...|
|45.477969888250485|       49.0|[1.1,193.5845,6.0...|
| 44.25257833876711|       50.4|[1.7,329.9747,5.0...|
| 50.81629766954042|       61.9|[3.6,373.8389,10....|
| 26.90225618398063|       31.7|[3.9,2147.376,3.0...|
| 26.90225618398063|       28.6|[4.0,2180.245,3.0...|
|  53.3096698932702|       52.2|[5.2,390.5684,5.0...|
| 57.44245200528963|       58.0|[6.2,90.45606,9.0...|
| 52.18840178549108|       57.1|[7.1,451.2438,5.0...|
| 48.15916592446566|       51.6|[8.1,104.8101,5.0...|
| 48.15916592446566|       56.8|[8.5,104.8101,5.0...|
| 43.98954343746342|       38.5|[9.0,1402.016,0.0...|
|46.581519090564576|       46.8|[11.4,390.5684,5....|
| 43.98954343746342|       46.6|[11.9,3171.329,0....|
|25.906210104130086|       28.9|[12.0,1360.139,1....|
| 26.17691535286829|       34.1|[12.5,1144.436000...|
| 38.88768998136467|       40.6|[12.8,732.8528,0....|
|  40.4557987788148|       42.5|[12.9,492.2313,5....|
|  40.4557987788148|       31.3|[13.3,492.2313,5....|
+------------------+-----------+--------------------+
only showing top 20 rows
```

The prediction using random forest and the fitness are as follows:

```
Root Mean Squared Error (RMSE) on train data = 5.13822
Root Mean Squared Error (RMSE) on test data = 5.93767
+------------------+-----------+--------------------+
|        prediction|house_price|            features|
+------------------+-----------+--------------------+
|  46.4088973475355|       52.2|[0.0,185.4296,0.0...|
|48.529367764312795|       50.7|[1.0,193.5845,6.0...|
|48.529367764312795|       49.0|[1.1,193.5845,6.0...|
| 51.79514221460514|       50.4|[1.7,329.9747,5.0...|
| 49.00474613617378|       61.9|[3.6,373.8389,10....|
| 26.85221031011472|       31.7|[3.9,2147.376,3.0...|
|  26.5849886983931|       28.6|[4.0,2180.245,3.0...|
| 51.40194399942163|       52.2|[5.2,390.5684,5.0...|
| 55.00006547987347|       58.0|[6.2,90.45606,9.0...|
| 48.32058067075683|       57.1|[7.1,451.2438,5.0...|
| 47.82785990411743|       51.6|[8.1,104.8101,5.0...|
| 47.82785990411743|       56.8|[8.5,104.8101,5.0...|
|   35.858718837535|       38.5|[9.0,1402.016,0.0...|
| 45.75558383267992|       46.8|[11.4,390.5684,5....|
| 34.05048074229692|       46.6|[11.9,3171.329,0....|
|26.869970467849146|       28.9|[12.0,1360.139,1....|
| 33.88185756563698|       34.1|[12.5,1144.436000...|
| 38.51240739710569|       40.6|[12.8,732.8528,0....|
|40.672106331600865|       42.5|[12.9,492.2313,5....|
|40.672106331600865|       31.3|[13.3,492.2313,5....|
+------------------+-----------+--------------------+
only showing top 20 rows
```

Conclusion

From comparing the RMSE in each model, we choose the model with lowest RMSE and predict the model with test dataset. We got our final score 8.14937.

Submission	Description	Score
rf.csv	111	8.14937