加州房價預測模型 : 使用R

資料集介紹

此次分析的資料為加州房價的資料集，每一筆資料分別紀錄一個街區中所有房屋的相關資訊。資料集中總共有10個欄位，分別是經度、緯度、平均屋齡、總房間數、總臥室數、總人口數、總家庭數、收入中位數、房價中位數和離海遠近。在預測模型中，房價中位數是依變數，其餘則為自變數。

資料集來源 ： https://www.kaggle.com/camnugent/california-housing-prices

資料預處理

首先輸入資料集並確認相關資訊

```
> # import dataset
> house.df = read.csv('D:/housing.csv')
> # check details
> str(house.df)
'data.frame': 20640 obs. of  10 variables:
 $ longitude         : num  -122 -122 -122 -122 -122 ...
 $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
 $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
 $ total_rooms       : num  880 7099 1467 1274 1627 ...
 $ total_bedrooms    : num  129 1106 190 235 280 ...
 $ population        : num  322 2401 496 558 565 ...
 $ households        : num  126 1138 177 219 259 ...
 $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
 $ median_house_value: num  452600 358500 352100 341300 342200 ...
 $ ocean_proximity   : Factor w/ 5 levels "<1H OCEAN","INLAND",..: 4 4 4 4 4 4 4 ...
```

接著確認資料中是否有缺失值，以及缺失值可能造成的影響

```
> dim(house.df[!complete.cases(house.df),])
[1] 207  10
```

在兩萬多筆資料中只有207筆資料有缺失值，總共只佔了資料集不到1%，於是決定直接將這些資料刪除。

特徵工程

選擇資料集當中的四項變數進行特徵工程，分別是總房間數、總臥室數、總人口數、總家庭數 ：

- 總房間數除以總家庭數，得到每個家庭的平均房間數量。
- 總臥室數除以總家庭數，得到每個家庭的平均臥室數量。
- 總人口數除以總家庭數，得到每個家庭的平均人數。

最後再把得到的三個新的特徵併入原先的資料中

```
avg_rooms = house.df$total_rooms/house.df$households
avg_bedrooms = house.df$total_bedrooms/house.df$households
persons_per_house = house.df$population/house.df$households
house.df = cbind(house.df, avg_rooms, avg_bedrooms, persons_per_house)
```

資料視覺化

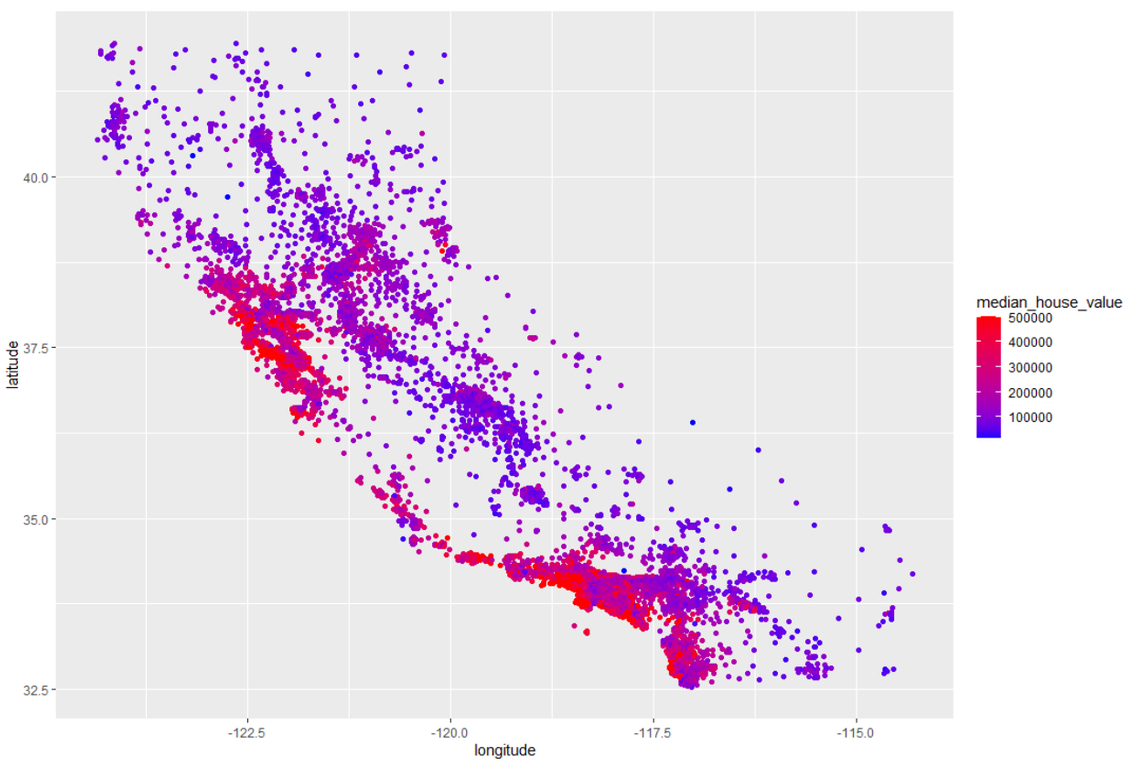

圖1 ： 依照各筆資料的經緯度所繪製的散佈圖，顏色由藍到紅代表房價由低到高。圖中可以看見在靠近灣區的地方房價中位數較高，離海越遠則房價中位數越低。

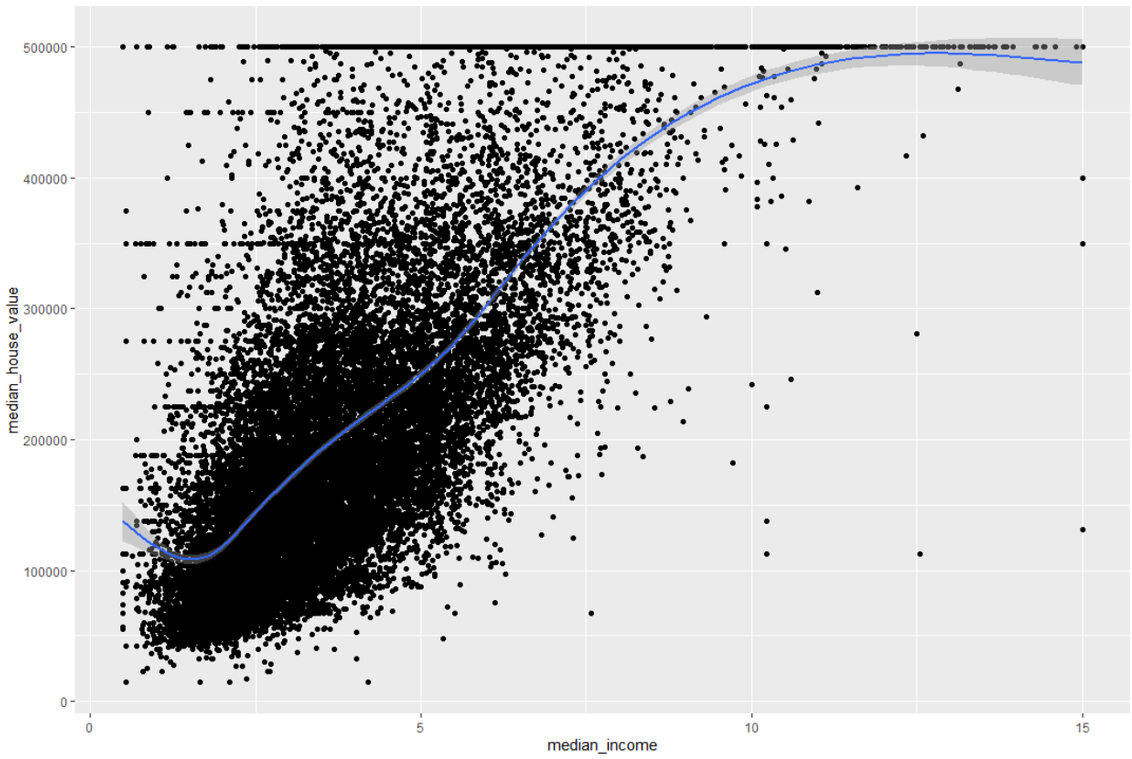

圖2 ： 房價中位數對收入中位數的散佈圖，可以發現明顯呈正相關。

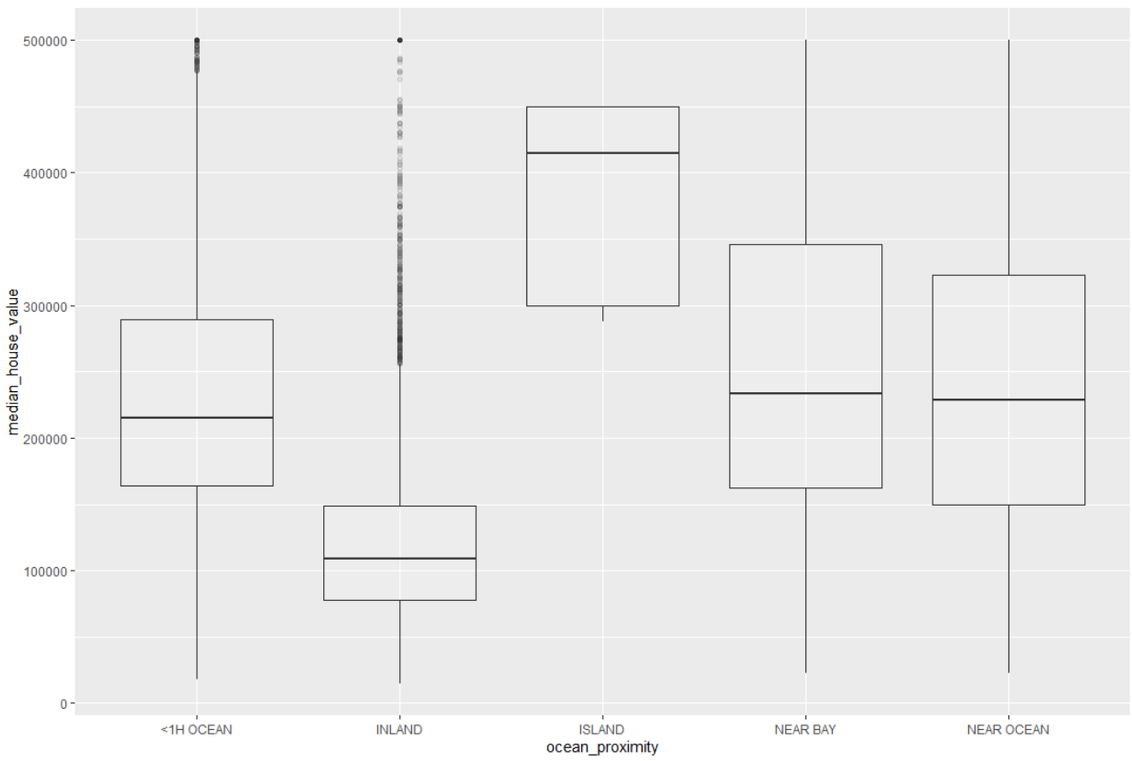

圖3 ： ocean proximity對房價的盒狀圖。離海較遠(inland)的房價中位數較低，而島嶼上的房屋(island)雖然看起來房價分布偏高，但實際上資料筆數並不多，不見得是好的評斷標準。

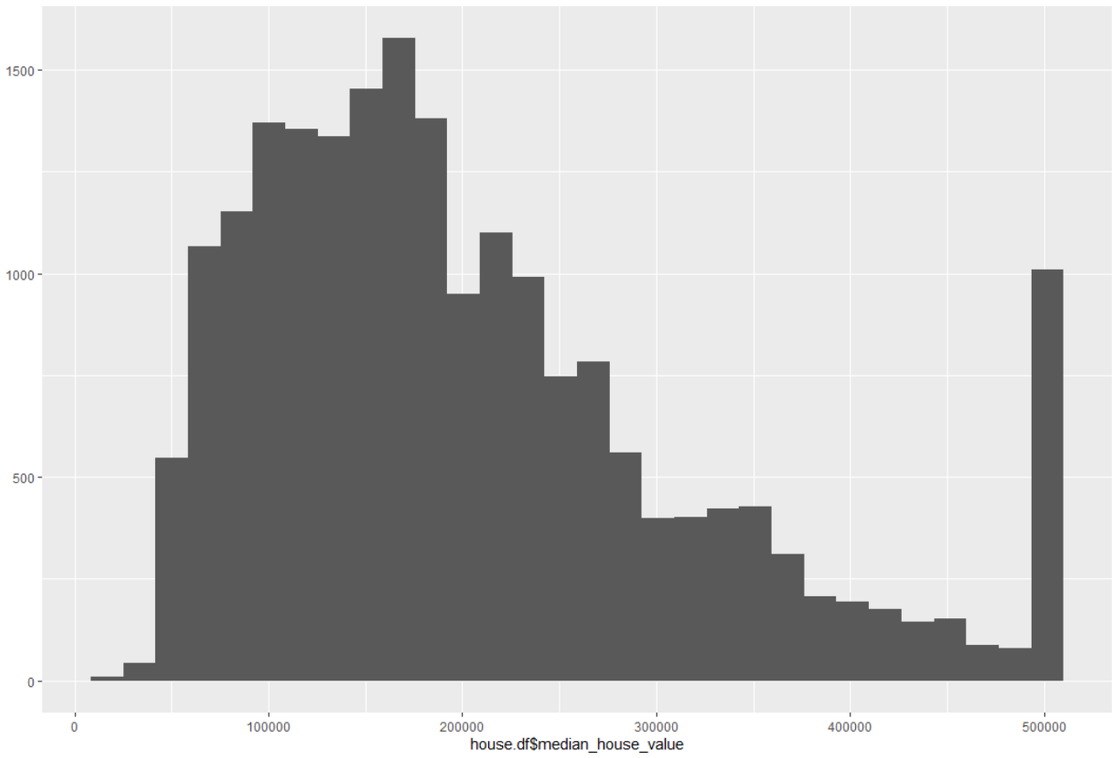

圖4 ： 房價分布直條圖。整體來說資料集的房價中位數為正偏態，但五十萬以上的房屋也非常多。

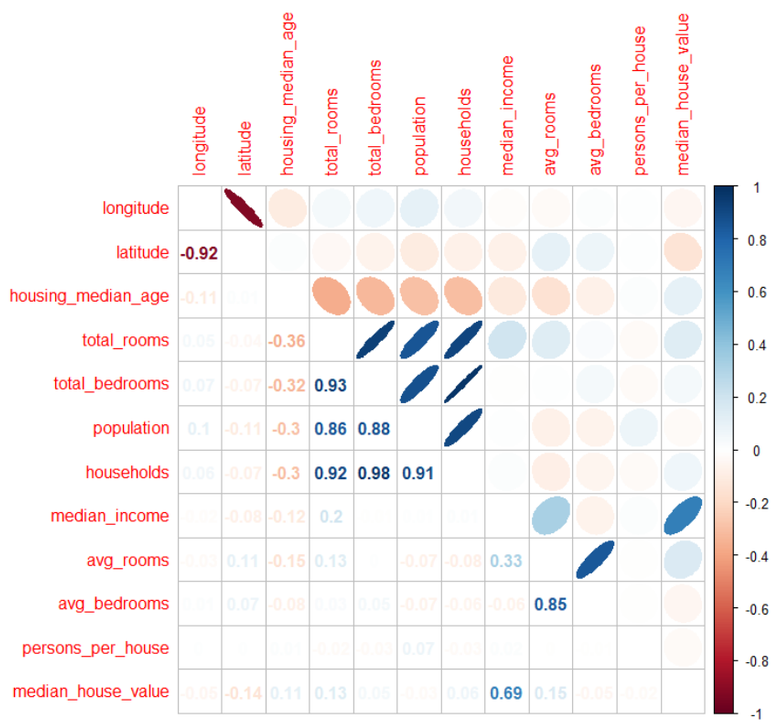

圖5 ： 所有欄位(包含新欄位)的相關矩陣

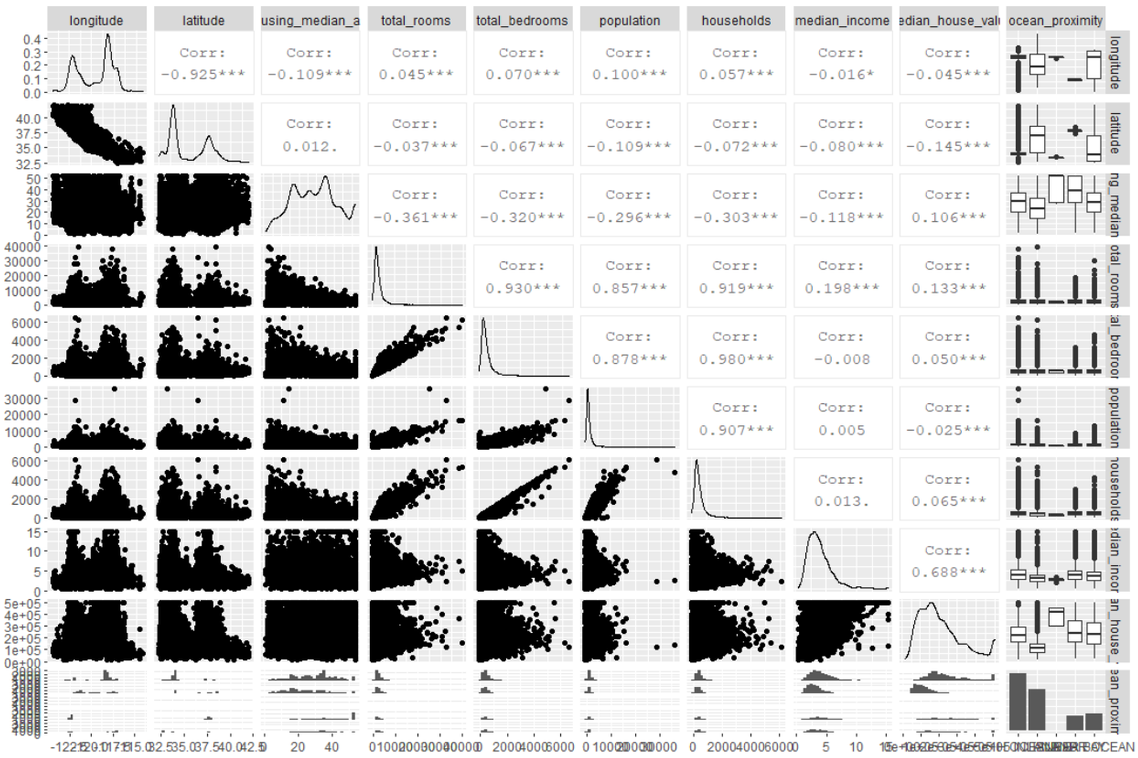

圖6 ： 原始資料中每兩個欄位之間的相關性

建立模型

首先將原先的資料集全數放進第一個迴歸模型中訓練，觀察模型的解釋力；第二個模型則是使用特徵工程後產生的新欄位進行迴歸

```
> # in sample explanation---------
> house.m1 = lm(median_house_value ~ longitude + latitude + housing_median_age +
+               total_rooms + total_bedrooms + population + households +
+               median_income + ocean_proximity, data = house.df)
> house.m2 = lm(median_house_value ~ longitude + latitude + housing_median_age +
+               median_income + avg_rooms + avg_bedrooms +  persons_per_house +
+               ocean_proximity, data = house.df)
```

接著分別對兩個模型進行5-fold cross validation，得到以下結果：

```
> cv.lm(data=house.df, house.m1, m=5)
> cv.lm(data=house.df, house.m2, m=5)
```

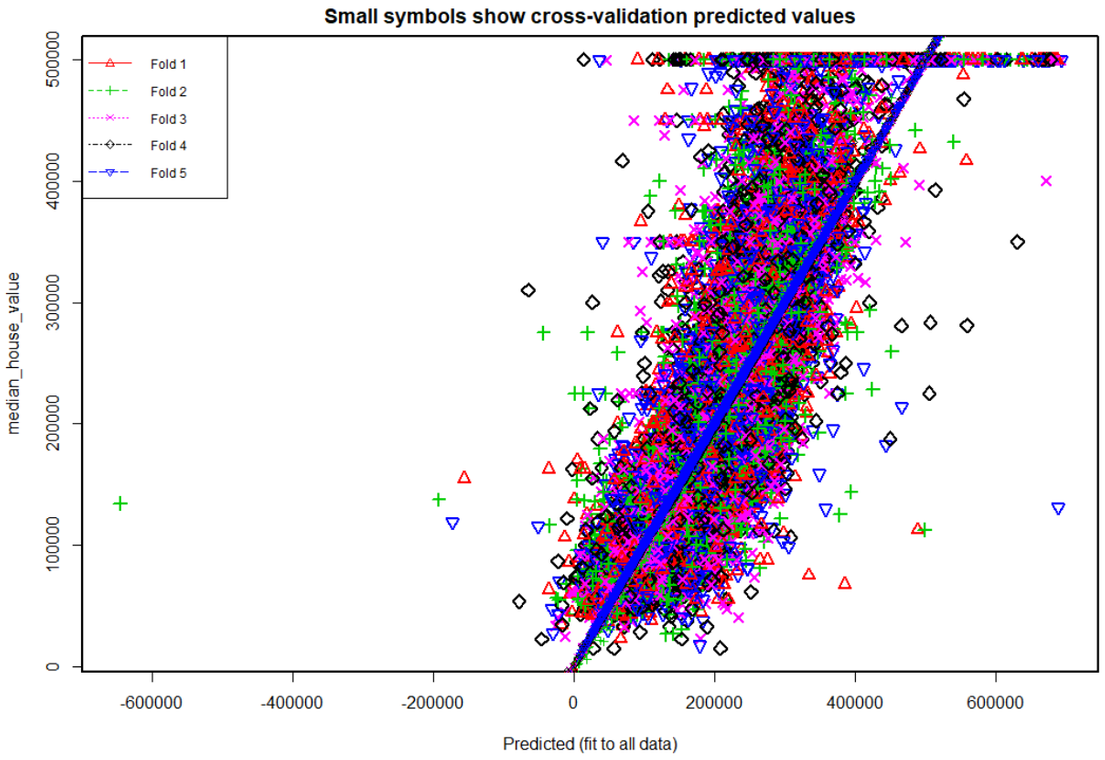

圖7 ： 第一個模型的5-fold cross validation

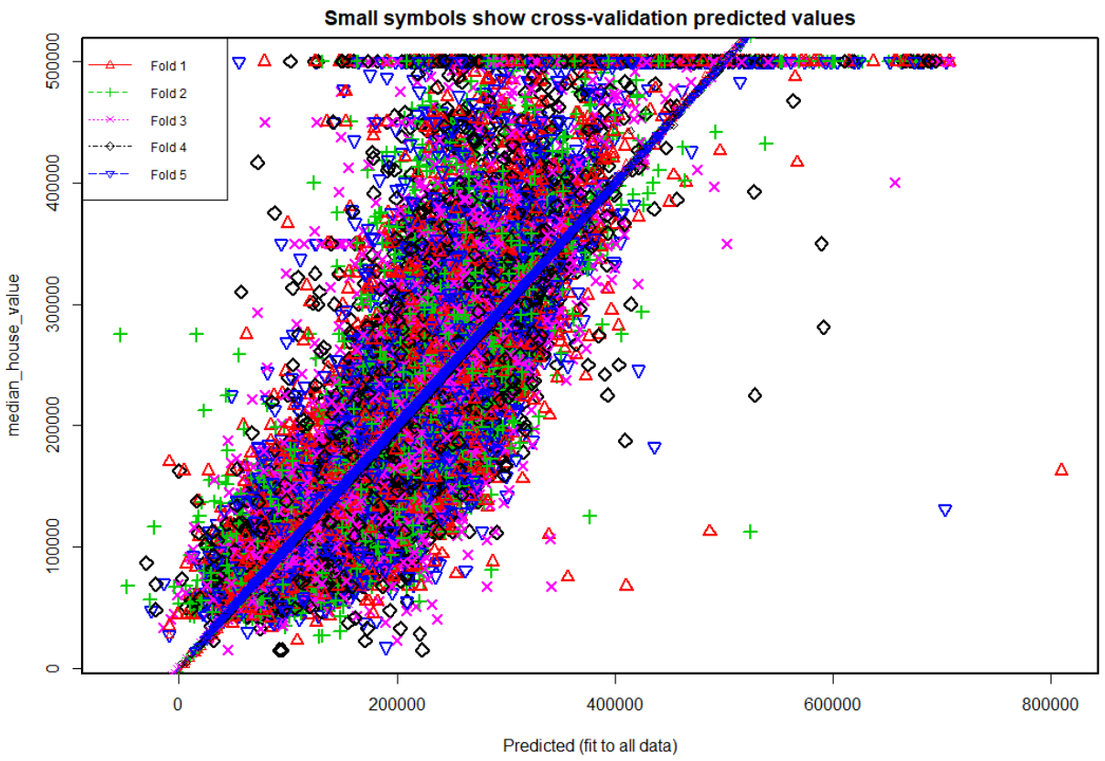

圖8 ： 第二個模型的5-fold cross validation

從cross-validation的結果可以看出第一個使用原始資料的模型的交叉驗證略優於第二個模型，也因為每個自變數的p-value足夠小，因此接下來會採用第一個模型進行預測。

預測

首先切分原始資料集為訓練和測試兩個部分

```
train.index = sample(c(1:20433), 16000, replace=FALSE)
house.train = house.df[train.index,]
house.test = house.df[-train.index,]
```

接著使用訓練集來訓練待會要使用的預測模型，以下是模型的相關資訊

```
> predict.house = lm(median_house_value ~ longitude + latitude + housing_median_age +
+                    total_rooms + total_bedrooms + population + households +
+                    median_income + ocean_proximity, data = house.train)
> summary(predict.house)

Call:
lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + ocean_proximity, data = house.train)

Residuals:
    Min      1Q  Median      3Q     Max
-554798  -42415  -10162   28596  785990

Coefficients:
                            Estimate   Std. Error t value             Pr(>|t|)
(Intercept)               -2183906.109    98784.534  -22.11 < 0.0000000000000002 ***
longitude                   -25784.397     1144.142  -22.54 < 0.0000000000000002 ***
latitude                    -24381.972     1128.327  -21.61 < 0.0000000000000002 ***
housing_median_age            1047.741       49.783   21.05 < 0.0000000000000002 ***
total_rooms                     -5.754        0.894   -6.44      0.0000000001257 ***
total_bedrooms                  91.587        7.554   12.12 < 0.0000000000000002 ***
population                     -38.276        1.227  -31.19 < 0.0000000000000002 ***
households                      56.970        8.172    6.97      0.0000000000033 ***
median_income                39074.511      384.169  101.71 < 0.0000000000000002 ***
ocean_proximityINLAND       -40756.855     1982.992  -20.55 < 0.0000000000000002 ***
ocean_proximityISLAND       142699.381    39817.157    3.58              0.00034 ***
ocean_proximityNEAR BAY      -4869.595     2154.558   -2.26              0.02383 *
ocean_proximityNEAR OCEAN     5578.836     1778.512    3.14              0.00171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68900 on 15987 degrees of freedom
Multiple R-squared:  0.643, Adjusted R-squared:  0.643
F-statistic: 2.4e+03 on 12 and 15987 DF,  p-value: <0.0000000000000002
```

預測模型的各項係數如下

```
> coef(predict.house)
              (Intercept)                 longitude                  latitude
              -2183906.11                -25784.40                -24381.97
       housing_median_age              total_rooms           total_bedrooms
                  1047.74                   -5.75                    91.59
               population               households            median_income
                   -38.28                   56.97                 39074.51
    ocean_proximityINLAND     ocean_proximityISLAND   ocean_proximityNEAR BAY
                -40756.86                142699.38                -4869.59
ocean_proximityNEAR OCEAN
                  5578.84
```

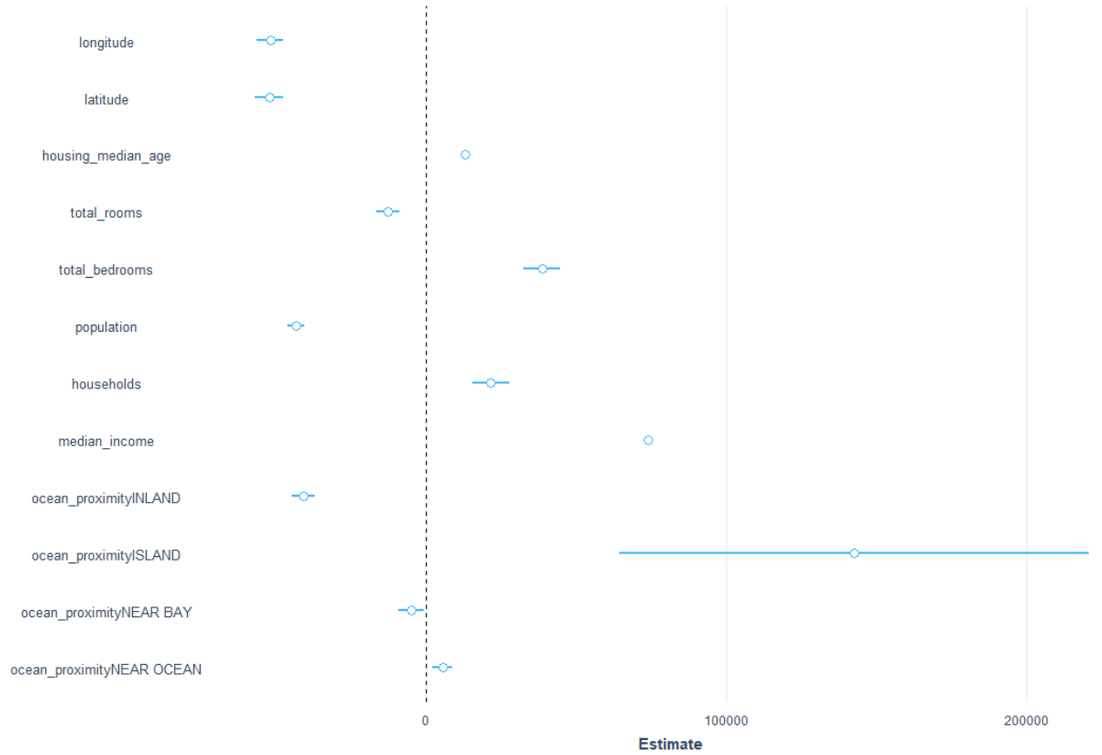

圖9 ： 預測模型中各項變數的信賴區間。其中ocean proximity的island因為資料筆數太少，因此信賴區間範圍很大。

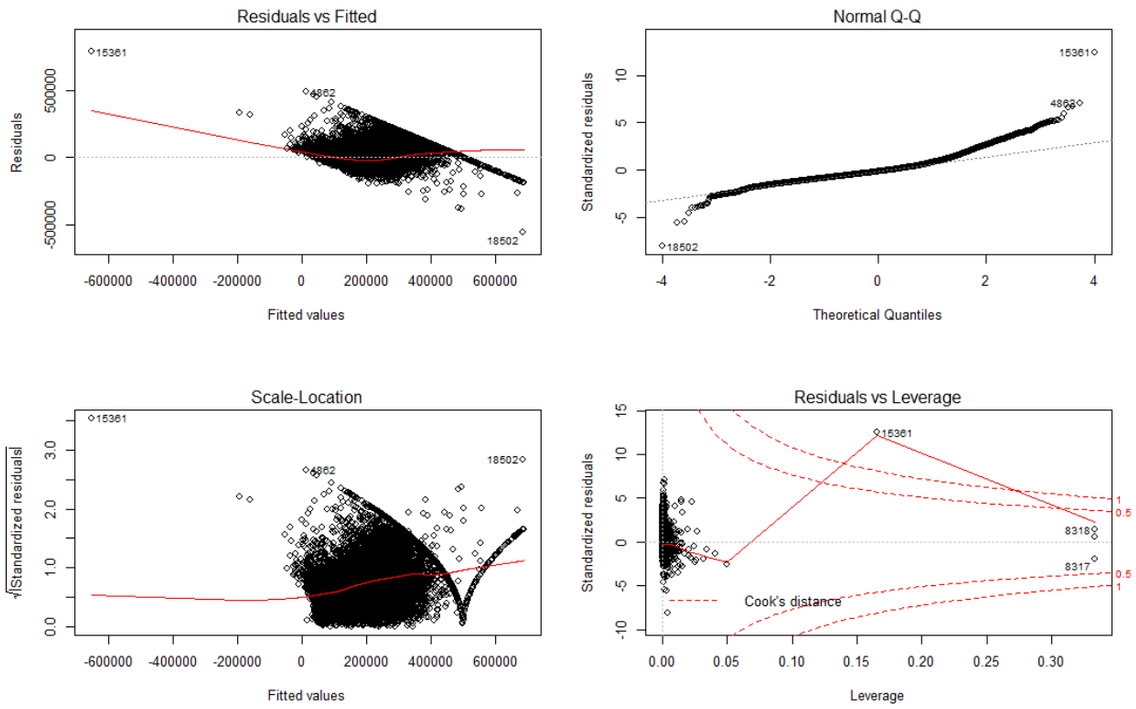

圖10 ： 預測模型的相關圖表

將測試資料集放入訓練好的迴歸模型中進行預測，得到以下結果 ：

```
> test.pred = predict(predict.house, house.test)
> accuracy(test.pred, house.test$median_house_value)
           ME   RMSE    MAE    MPE MAPE
Test set 2.87 67865 49694 -10.5 28.2
```

最後列出部分詳細的預測結果和殘差

```
> # residuals
> res.index = sample(c(1:4433), 20, replace=FALSE)
> test.residuals = house.test$median_house_value[res.index] - test.pred[res.index]
> data.frame("Actual" = house.test$median_house_value[res.index],
+            "Predicted" = test.pred[res.index],
+            "Residual" = test.residuals)
       Actual Predicted Residual
15511 157700    243625   -85925
13285 135000     74440    60560
18194 232700    287681   -54981
5103  131700    193053   -61353
13548  70900     52695    18205
8066  277700    286217    -8517
10739 500001    306485   193516
18410 279100    300685   -21585
11871 145200    106519    38681
1910   97600     71646    25954
12432  58600     36488    22112
13121 151600    156392    -4792
1629  399700    426372   -26672
8549  233300    216221    17079
16635 196100    226060   -29960
20493 204600    252942   -48342
297    90600    162787   -72187
6438  233700    213996    19704
3807  195600    163969    31631
14701 181000    274261   -93261
```

結語

在此次建立的預測模型有不少問題需要克服，例如feature engineering、資料分布不均或極端值等等。而最後的預測模型結果尚可，雖然特徵工程的結果不盡理想，資料集當中的欄位可能也沒辦法完整推測得知每個地區的不動產細節特性，但得到的預測結果算是在接受範圍內。