資料科學 - 第三組期末報告

Sberbank Russian Housing Market

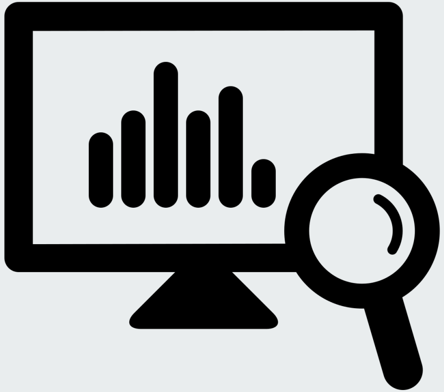
柯敦瀚 統計四 107304050

陳羿丞 資科碩一 110753138

蘇俊憲 資科碩一 110753158

朱進益 資科碩一 110753144

洪丞桀 資科三 108703045



Outline

- 專案簡介
- 資料預處理
- 特徵觀察
- 建造模型
- 輸出結果

Project introduction

Sberbank Russian Housing Market - 房地產價格預測

本次題目是Kaggle上2017年由

Sberbank(俄羅斯聯邦儲蓄銀行)發表的競賽

內容旨在藉由巨量資料去分析特定地方的

房地產價格受影響的因素

並且**預測出房地產價格**

Input/Goal

提供資料有 train.csv/ test.csv/ macro.csv

train dataset裡面有 **292**筆特徵資料

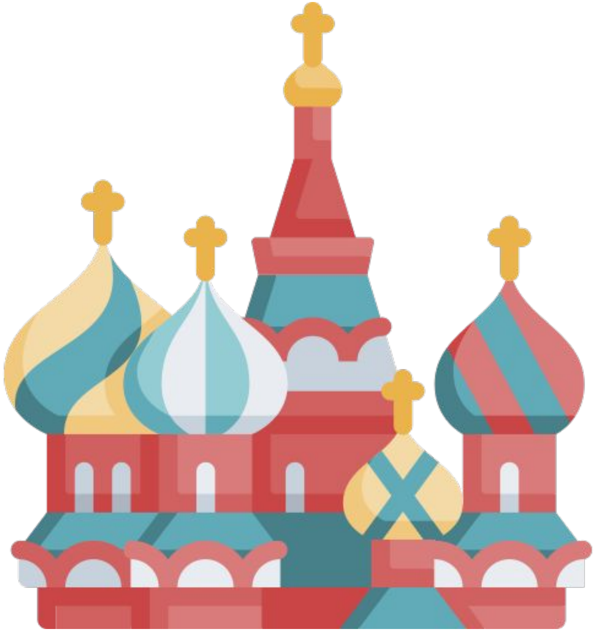
macro dataset裡面有 **100**筆特徵資料

train data從居住條件到居住環境以及附近的條件

macro data是俄羅斯的經濟與房市大觀環境條件

我們的目標在於 **price_doc** 參數

以現有的所有資料去預測出價格



Data Dictionary

30471 rows and 292 columns

```
> str(traindata)
'data.frame':	30471 obs. of  292 variables:
 $ id                                   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ timestamp                            : Factor w/ 1161 levels "2011-08-20","2011-08-23",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ full_sq                              : int  43 34 43 89 77 67 25 44 42 36 ...
 $ life_sq                              : int  27 19 29 50 77 46 14 44 27 21 ...
 $ floor                                : int  4 3 2 9 4 14 10 5 5 9 ...
 $ max_floor                            : int  NA NA NA NA NA NA NA NA NA NA ...
 $ material                             : int  NA NA NA NA NA NA NA NA NA NA ...
 $ build_year                           : int  NA NA NA NA NA NA NA NA NA NA ...
 $ num_room                             : int  NA NA NA NA NA NA NA NA NA NA ...
 $ kitch_sq                             : int  NA NA NA NA NA NA NA NA NA NA ...
 $ state                                : int  NA NA NA NA NA NA NA NA NA NA ...
 $ product_type                         : Factor w/ 2 levels "Investment","OwnerOccupier": 1 1 1 1 1 1 1 1 1 1 ...
 $ sub_area                             : Factor w/ 146 levels "Ajeroport","Akademicheskoe",..: 10 71 130 66 7 74 123 10 45 51 ...
```

And 200+ variables more…

Data pre-processing

參考了Kaggle上面的很多文章

Correlation coefficient 去一一查看每個資料的關係

發現有非常多的參數都是沒有用的

對價格的預測幾乎是沒有正面影響

像是附近五百公尺的咖啡店價格根本沒有影響房價

Variables with almost zero variance

```
##  [1] "culture_objects_top_25_raion" "oil_chemistry_raion"
##  [3] "railroad_terminal_raion"      "nuclear_reactor_raion"
##  [5] "build_count_foam"             "big_road1_1line"
##  [7] "railroad_1line"               "office_sqm_500"
##  [9] "trc_sqm_500"                  "cafe_count_500_price_4000"
## [11] "cafe_count_500_price_high"    "mosque_count_500"
## [13] "leisure_count_500"            "office_sqm_1000"
## [15] "trc_sqm_1000"                 "cafe_count_1000_price_high"
## [17] "mosque_count_1000"            "cafe_count_1500_price_high"
## [19] "mosque_count_1500"            "cafe_count_2000_price_high"
```

Data pre-processing

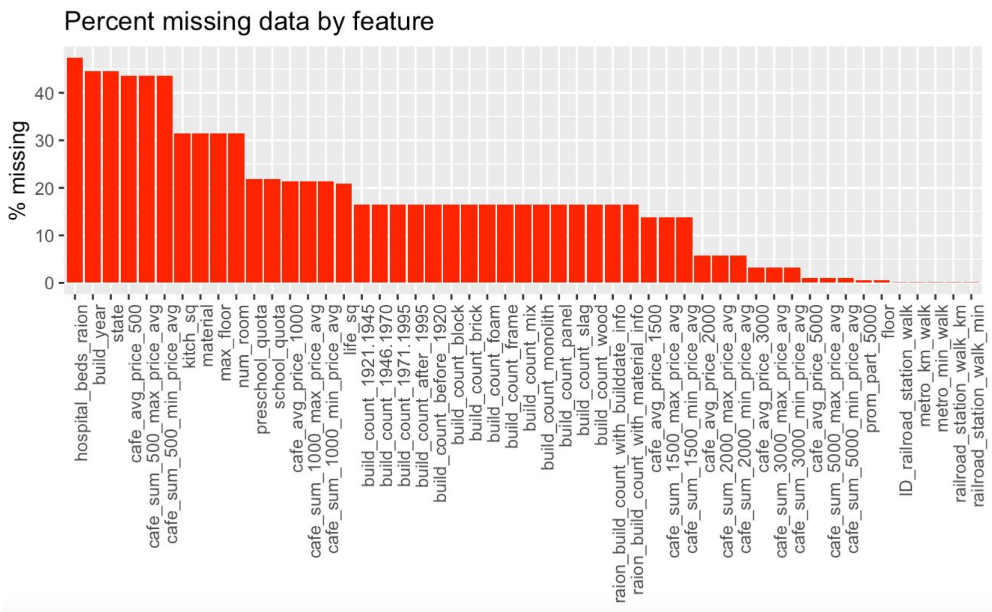
對於**NA值過多**的特徵

假設他對房價毫無影響

我們採用的方式是**直接刪除該特徵**

結果也證明了沒有某些特徵

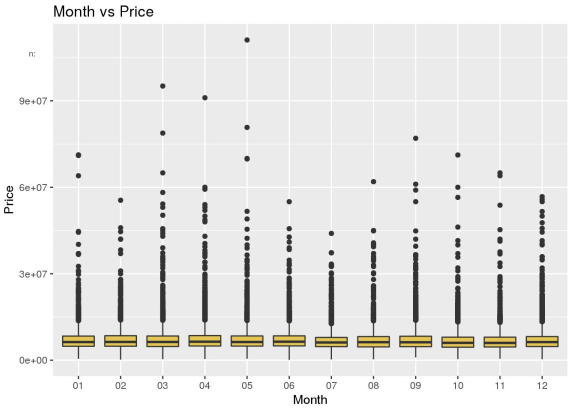
模型的預測會更加準確

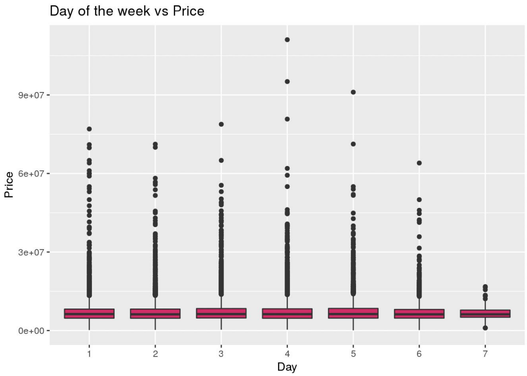


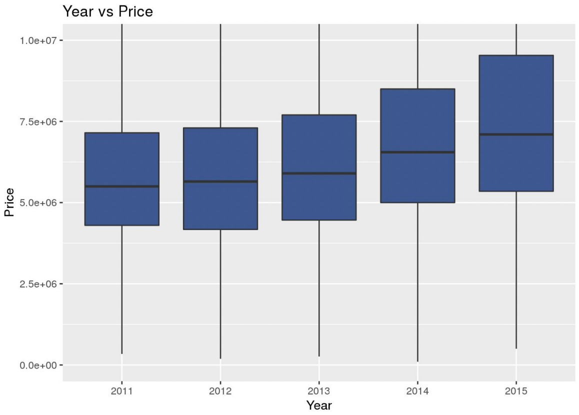
Explore Data Analysis - month、year、weekend

Extracting month, year, weekday from timestamp

An upward trend with each passing year. But no trend with months or days.

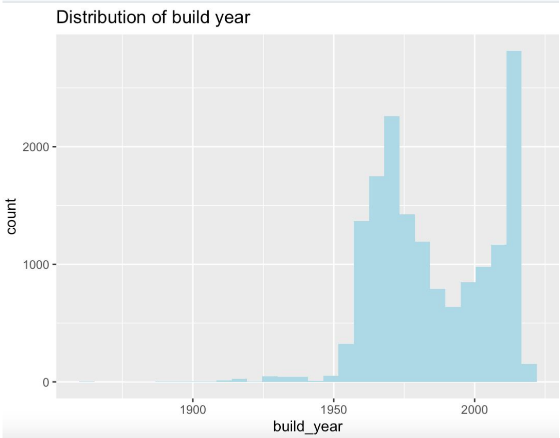


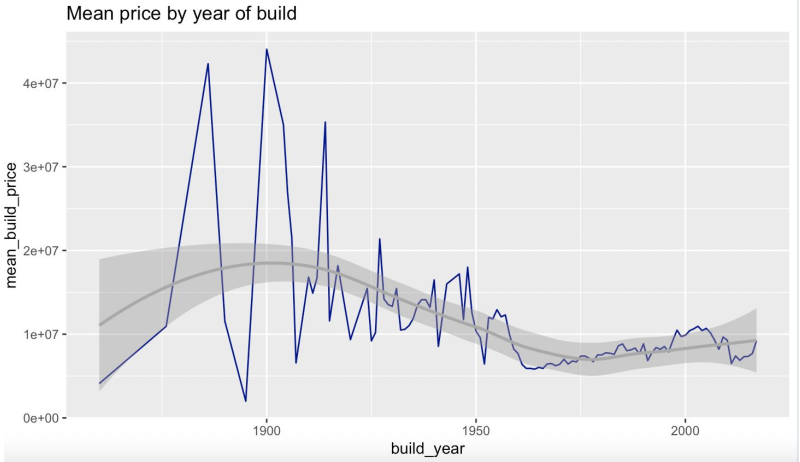




Explore Data Analysis - build year

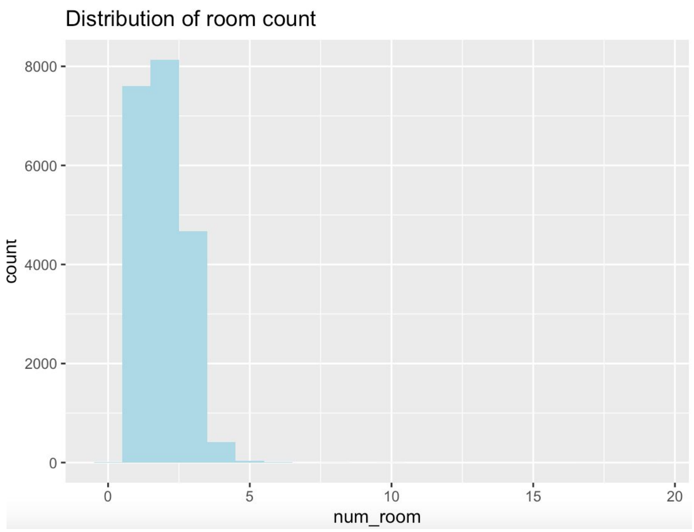
- 時間資料分布不均，大致上是右偏
- 位於 1970 到 2000 年為一個分佈，2000 年後為另一個
- 觀察發現 1950 年前的波動性極大，1950後逐漸穩定



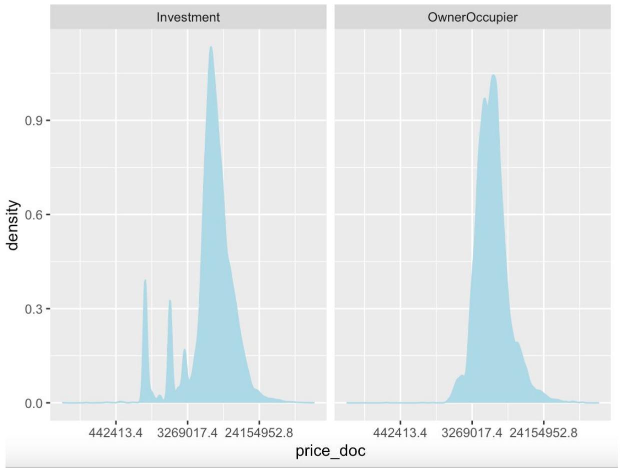


Explore Data Analysis: 觀察幾個重要的變數

- 多數的資料有有三間或小於三間房間



- 觀察房子價格跟買的用途的關係
- 投資或家用是否有影響 >> 投資用的房子比家用的房子賣得多



Model-RandomForest

```
           PC1         PC2        PC3           PC4          PC5         PC6
1 -0.38815520  0.02928363 0.03480449  0.0004469820 -0.001938400 0.002159991
2 -0.37680555  0.02893570 0.03806354  0.0041676640 -0.001252289 0.001787335
3  0.06388236  0.07285331 0.02380022  0.0005461602  0.003023358 0.001664850
4 -0.29596931  0.05284529 0.04719196  0.0133202463 -0.005152651 0.001264901
5 -0.37711888  0.03059530 0.03748995  0.0045679913 -0.002233853 0.001744311
6  0.10711896 -0.17541397 0.06156278 -0.0016041496 -0.038909206 0.002730725
           PC7          PC8           PC9          PC10
1 -0.0002294501 0.0018895811 -6.914068e-04 -0.0002013482
2 -0.0041346477 0.0008826625  6.093823e-05 -0.0004203463
3 -0.0067615807 0.0142664004  4.149853e-03  0.0051117971
4 -0.0045096568 0.0007381771 -9.553007e-04 -0.0004373865
5 -0.0038685833 0.0008610004 -1.312755e-05 -0.0002639678
6  0.0090397655 0.0277662456 -2.245335e-02  0.0137234006
```

```
  method           from
  as.zoo.data.frame zoo
    full_sq life_sq floor max_floor material build_year num_room kitch_sq state
1        43      27   4.0        12        1       2014        2        6     2
2        34      19   3.0        12        1       2014        2        6     2
3        43      29   2.0        12        1       2014        2        6     2
4        89      50   9.0        12        1       2014        2        6     2
5        77      77   4.0        12        1       2014        2        6     2
6        67      46  14.0        12        1       2014        2        6     2
7        25      14  10.0        12        1       2014        2        6     2
8        44      44   5.0        12        1       2014        2        6     2
9        42      27   5.0        12        1       2014        2        6     2
10       36      21   9.0        12        1       2014        2        6     2
11       36      19  12.0        12        1       2014        2        6     2
12       38      19  11.0        12        1       2014        2        6     2
13       43      28   4.0        12        1       2014        2        6     2
14       31      31   4.0        12        1       2014        2        6     2
```

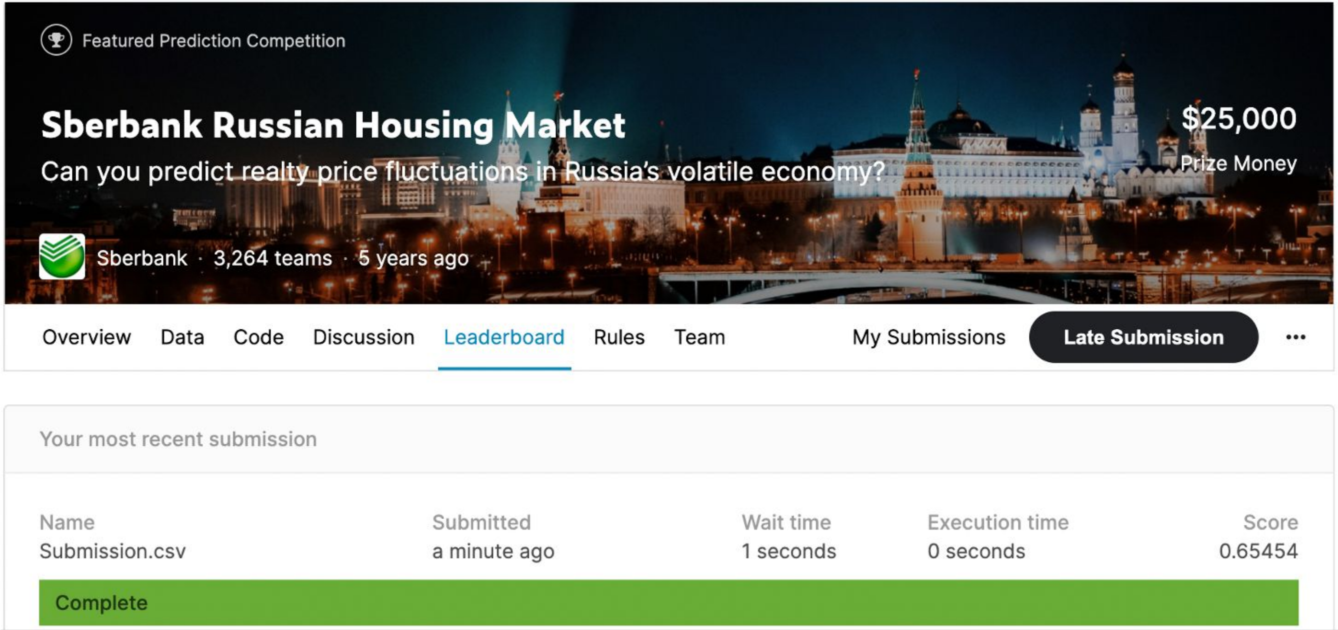
```
build_count_monolith build_count_panel build_count_foam  build_count_slag
Min.   :  0.00       Min.   :  0.0     Min.   : 0.0000   Min.   : 0.000
1st Qu.:  3.00       1st Qu.: 46.0     1st Qu.: 0.0000   1st Qu.: 0.000
Median :  6.00       Median : 92.0     Median : 0.0000   Median : 0.000
Mean   : 11.05       Mean   :104.7     Mean   : 0.1385   Mean   : 3.756
3rd Qu.: 11.00       3rd Qu.:134.0     3rd Qu.: 0.0000   3rd Qu.: 1.000
Max.   :127.00       Max.   :431.0     Max.   :11.0000   Max.   :84.000

build_count_mix  raion_build_count_with_builddate_info build_count_before_1920
Min.   :0.0000   Min.   :   1.0                        Min.   :  0.0
1st Qu.:0.0000   1st Qu.: 196.0                        1st Qu.:  0.0
Median :0.0000   Median : 271.0                        Median :  0.0
Mean   :0.4793   Mean   : 318.9                        Mean   : 15.8
3rd Qu.:0.0000   3rd Qu.: 371.0                        3rd Qu.:  2.0
Max.   :9.0000   Max.   :1680.0                        Max.   :371.0

build_count_1921.1945 build_count_1946.1970 build_count_1971.1995
Min.   :  0.00        Min.   :  0.0         Min.   :  0.00
1st Qu.:  0.00        1st Qu.: 40.0         1st Qu.: 42.00
Median :  2.00        Median :135.0         Median : 71.00
Mean   : 22.67        Mean   :140.4         Mean   : 78.63
3rd Qu.: 14.00        3rd Qu.:193.0         3rd Qu.:103.00
Max.   :382.00        Max.   :845.0         Max.   :246.00

build_count_after_1995    ID_metro       metro_min_avto   metro_km_avto
Min.   :  0.00          Min.   :  1.00   Min.   : 0.000   Min.   : 0.000
1st Qu.: 16.00          1st Qu.: 27.00   1st Qu.: 1.721   1st Qu.: 1.037
Median : 24.00          Median : 53.00   Median : 2.803   Median : 1.784
Mean   : 55.07          Mean   : 72.48   Mean   : 4.961   Mean   : 3.701
3rd Qu.: 53.00          3rd Qu.:108.00   3rd Qu.: 4.832   3rd Qu.: 3.777
Max.   :799.00          Max.   :223.00   Max.   :61.438   Max.   :74.906
```

Model-RandomForest



RMSLE(Root Mean Squared Log Error)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(\hat{y}_i+1)-log(y_i+1))^2}$$

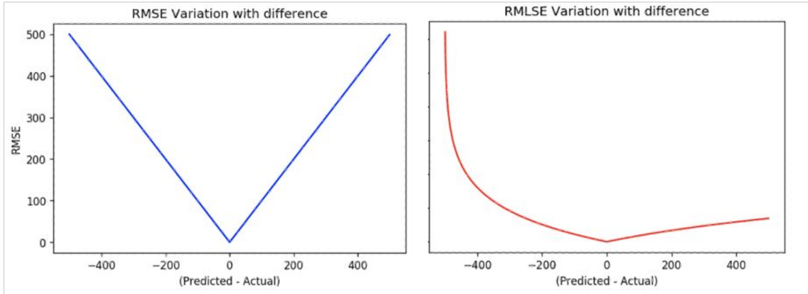
- 異常值被大幅縮小，因此消減了它們的影響。
- 由於對數的性質，RMLSE 可以廣義地看作是預測值與實際值之間的相對誤差誤差。

$$\log(x_i+1)-\log(y_i+1)=\log\frac{(x_i+1)}{(y_i+1)}$$

RMSE(Root Mean Square Error)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

- 異常值的存在會使誤差項爆炸到非常高的值。



RMSE

```
[1]     train-rmse:8217503.700000+25576.297255  test-rmse:8221296.300000+140217.893859
Multiple eval metrics are present. Will use test_rmse for early stopping.
Will train until test_rmse hasn't improved in 150 rounds.

[51]    train-rmse:2522840.800000+25862.893131  test-rmse:2885752.150000+160606.036570
[101]   train-rmse:2193154.550000+22335.683373  test-rmse:2708084.950000+165185.636743
[151]   train-rmse:2071055.725000+16831.106152  test-rmse:2663814.800000+165684.592686
[201]   train-rmse:1981774.075000+15726.751355  test-rmse:2644994.750000+162207.731152
[251]   train-rmse:1908577.575000+12947.968328  test-rmse:2631935.200000+160754.028252
[300]   train-rmse:1849687.700000+11107.380766  test-rmse:2624417.950000+157948.910663
[1] 300
```

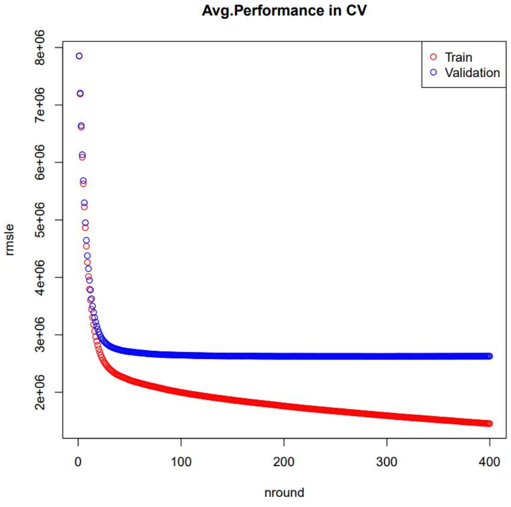
RMSLE

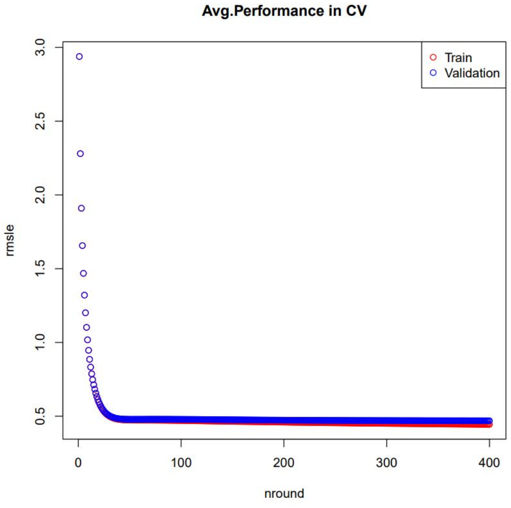
```
[51]    train-rmsle:0.464680+0.001385   test-rmsle:0.474202+0.009407
[101]   train-rmsle:0.460318+0.001926   test-rmsle:0.474541+0.010645
[151]   train-rmsle:0.452715+0.001941   test-rmsle:0.471574+0.010924
[201]   train-rmsle:0.445581+0.001847   test-rmsle:0.469593+0.011004
[251]   train-rmsle:0.439506+0.002106   test-rmsle:0.468427+0.010675
[301]   train-rmsle:0.433681+0.001985   test-rmsle:0.467494+0.010734
[351]   train-rmsle:0.428090+0.001756   test-rmsle:0.466769+0.010897
[400]   train-rmsle:0.422472+0.001909   test-rmsle:0.466277+0.010953
```

Result

Submission and Description	Private Score	Public Score	Use for Final Score
prop_price_xgb_fix.csv 6 minutes to go by 1101DS@NCCU_110753158 add submission details	0.34189	0.34214	☐

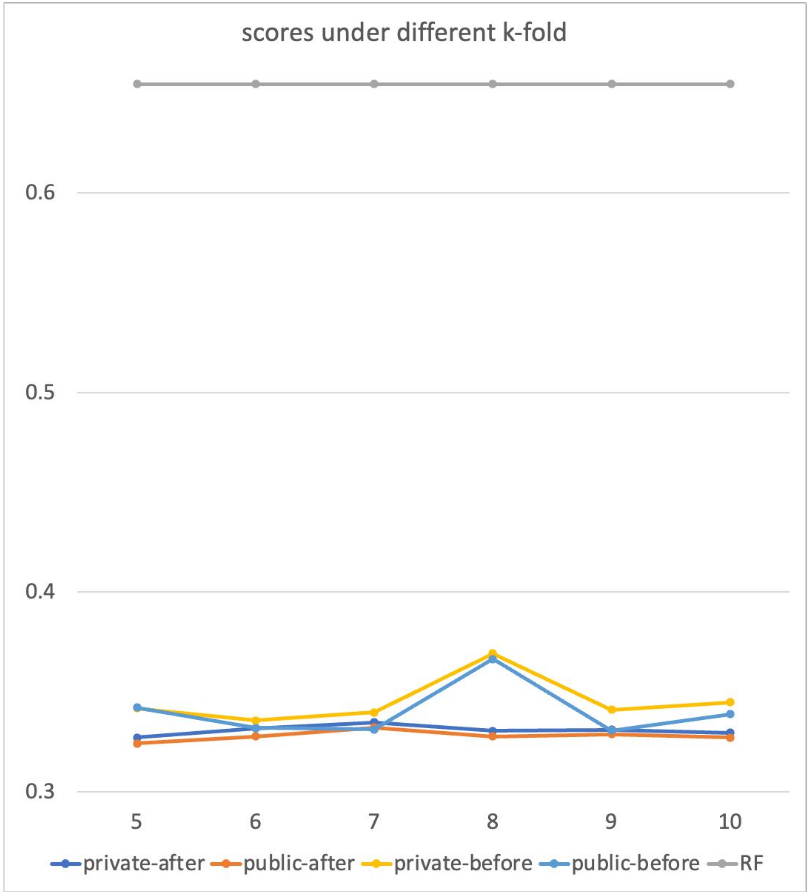
Submission and Description	Private Score	Public Score	Use for Final Score
prop_price_xgb_fix_dep4N5_fin.csv 6 minutes to go by 1101DS@NCCU_110753158 add submission details	0.32707	0.32419	☐





Result


scores under different k-fold
0.4
0.35
0.3
5
6
7
8
9
10
private-after
public-after
private-before
public-before


scores under different k-fold
0.6
0.5
0.4
0.3
5
6
7
8
9
10
private-after
public-after
private-before
public-before
RF

Result

Public Leaderboard   Private Leaderboard

The private leaderboard is calculated with approximately 65% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	alijs & Evgeny			0.30087	311	5Y
2	▲ 9	data_mining2			0.30925	151	5Y
3	▲ 27	Computer says no			0.31032	110	5Y

Reference

- https://www.kaggle.com/creatrol/basic-time-series-analysis-feature-selection
- https://www.kaggle.com/arathee2/creating-some-useful-additional-features/report
- https://www.kaggle.com/captcalculator/a-very-extensive-sberbank-exploratory-analysis
- https://rpubs.com/skydome20/R-Note16-Ensemble_Learning
- https://www.kaggle.com/keerthip/random-forest
- https://www.kaggle.com/abhishekkant/another-xgb-model
- https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a
-