house_prices_modeling

May 26, 2021

[1]:
```
import numpy as np
import pandas as pd
from scipy import stats

from sklearn.metrics import mean_absolute_error, make_scorer
from sklearn.model_selection import train_test_split, cross_val_score

from tqdm import tqdm
tqdm.pandas()

import seaborn as sns
sns.set(font_scale=1.5)

import matplotlib.pyplot as plt
import matplotlib.style as style
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
```

```
C:\Users\HaKKe\anaconda3\lib\site-packages\tqdm\std.py:658: FutureWarning: The
Panel class is removed from pandas. Accessing it from the top-level namespace
will also be removed in the next version
  from pandas import Panel
```

[2]:
```
X_train = pd.read_csv('dataset/X_train.csv')
X_val = pd.read_csv('dataset/X_test.csv')
```

[3]:
```
y_train = X_train.iloc[:, -1]
X_train = X_train.iloc[:, :-1]
```

[4]:
```
y_val = X_val.iloc[:, -1]
X_val = X_val.iloc[:, :-1]
```

[5]:
```
X_train.head(5)
```

[5]:

	id	date	street_id	build_tech	floor	area	rooms	balcon	metro_dist	\
0	0	2011-1	616	0.0	4	43	2	0	0.0	

```
1   1  2011-1        112         0.0      3    33      1       0        15.0
2   2  2011-1        230         0.0      9    34      1       0         5.0
3   3  2011-1        302         1.0      4    60      3       0        15.0
4   4  2011-1        578         0.0      3    49      2       0         0.0

   g_lift  …  kw8  kw9  kw10  kw11  kw12  kw13  mean_square_root_price  \
0     1.0  …    0    0     0     0     0     0            60749.113126
1     1.0  …    0    0     0     0     0     0            70386.034256
2     1.0  …    0    0     0     0     0     0           121089.980060
3     0.0  …    0    0     0     0     0     0            67595.600000
4     0.0  …    0    0     0     0     0     0            46023.349880

   avg_room_area  area_and_balcon  mean_street_floor_square_price
0           21.5             43.0                     64713.953488
1           33.0             33.0                     61434.343434
2           34.0             34.0                    129696.078431
3           20.0             60.0                     68161.458333
4           24.5             49.0                     46278.571429

[5 rows x 28 columns]
```

[6]:
```
from catboost import CatBoostRegressor
```

[7]:
```
cat_features = ['street_id', 'build_tech', 'balcon', 'date']
X_train.drop(columns=['id'], axis=0, inplace=True)
X_val.drop(columns=['id'], axis=0, inplace=True)
```

[8]:
```
X_train.build_tech = X_train.build_tech.astype(int)
X_train.metro_dist = X_train.metro_dist.astype(int)
```

[9]:
```
X_val.build_tech = X_val.build_tech.astype(int)
X_val.metro_dist = X_val.metro_dist.astype(int)
```

[10]:
```
Catboost = CatBoostRegressor(
    depth=6,
    n_estimators=1000,
    learning_rate=0.03,
    max_ctr_complexity=4,
    leaf_estimation_iterations=5,
    l2_leaf_reg=3,
    bagging_temperature=1,
    leaf_estimation_method='Newton',
    cat_features=cat_features,
    eval_metric='MAE',
)
```

[11]:
```
Catboost.fit(
    X_train,
    y_train,
    eval_set=[(X_val, y_val)],
    verbose=100,
)
```

```
0:      learn: 2580644.3627537  test: 2837651.9151693   best: 2837651.9151693
(0)       total: 250ms    remaining: 4m 9s
100:    learn: 847054.3101292   test: 1131886.6170456   best: 1131886.6170456
(100)     total: 13.2s    remaining: 1m 57s
200:    learn: 705645.4597846   test: 1001802.8407799   best: 1001802.8407799
(200)     total: 28.7s    remaining: 1m 54s
300:    learn: 625714.8790884   test: 925036.9036554    best: 925036.9036554
(300)     total: 44.7s    remaining: 1m 43s
400:    learn: 574361.0414064   test: 877452.8315847    best: 877452.8315847
(400)     total: 1m       remaining: 1m 30s
500:    learn: 535182.8713221   test: 845035.5195872    best: 845035.5195872
(500)     total: 1m 14s   remaining: 1m 13s
600:    learn: 511169.8262400   test: 830136.3486306    best: 830136.3486306
(600)     total: 1m 26s   remaining: 57.3s
700:    learn: 491708.1718820   test: 819293.0085228    best: 819293.0085228
(700)     total: 1m 40s   remaining: 42.7s
800:    learn: 475653.3093635   test: 810947.8293067    best: 810358.4185833
(798)     total: 1m 55s   remaining: 28.6s
900:    learn: 462997.1745493   test: 794850.8086941    best: 794850.8086941
(900)     total: 2m 9s    remaining: 14.3s
999:    learn: 453156.8033173   test: 789565.0834406    best: 789327.8890819
(960)     total: 2m 25s   remaining: 0us

bestTest = 789327.8891
bestIteration = 960

Shrink model to first 961 iterations.
```

[11]:
```
<catboost.core.CatBoostRegressor at 0x22724da5048>
```

[12]:
```
y_pred = Catboost.predict(X_val)
coef = np.mean(y_val/y_pred)
```

[13]:
```
mean_absolute_error(y_val, y_pred*coef)
```

[13]:
```
555218.3177386109
```