```
 1 /Users/xingkaiwu/PycharmProjects/Zillow/venv/bin/python /
   Users/xingkaiwu/PycharmProjects/Zillow/model_catboost.py
 2 Number of features for CatBoost: 69
 3    cooling_id  bathroom_cnt  bedroom_cnt  quality_id
   floor1_sqft  \
 4 0           0          2.0          3.0          4.0
                NaN
 5 1          -1          3.5          4.0          NaN
                NaN
 6 2           0          3.0          2.0          4.0
                NaN
 7 3           0          2.0          2.0          4.0
                NaN
 8 4          -1          2.5          4.0          NaN
                NaN
 9
10    finished_area_sqft_calc  floor1_sqft_unk
   base_total_area  fireplace_cnt  \
11 0                   1684.0              NaN
   NaN           NaN
12 1                   2263.0              NaN
   NaN           NaN
13 2                   2217.0              NaN
   NaN           NaN
14 3                    839.0              NaN
   NaN           NaN
15 4                   2283.0              NaN
   NaN           NaN
16
17    bathroom_full_cnt  garage_cnt  garage_sqft  spa_flag
   heating_id  \
18 0                2.0         NaN          NaN       NaN
              1
19 1                3.0         2.0        468.0       NaN
             -1
20 2                3.0         NaN          NaN       NaN
              1
21 3                2.0         NaN          NaN       NaN
              1
22 4                2.0         2.0        598.0       NaN
             -1
23
24     latitude     longitude   lot_sqft  pool_cnt
   pool_total_size  pool_unk_2  \
25 0  34280992.0 -118488536.0    7528.0       NaN
            NaN          NaN
26 1  33668120.0 -117677552.0    3643.0       NaN
            NaN          NaN
27 2  34136312.0 -118175032.0   11423.0       NaN
            NaN          NaN
```

```
28  3  33755800.0 -118309000.0   70859.0         NaN
                  NaN         NaN
29  4  33485644.0 -117700232.0    6000.0         1.0
                  NaN         NaN
30
31     pool_unk_3 landuse_type_id     census_1   city_id
   neighborhood_id  \
32  0         NaN             230  60371068.0  12447.0
                  31817.0
33  1         NaN             230  60590524.0  32380.0
                  NaN
34  2         NaN             230  60374640.0  47019.0
   275411.0
35  3         NaN             235  60372964.0  12447.0
                  54300.0
36  4         1.0             230  60590424.0  17686.0
                  NaN
37
38     region_zip  room_cnt  bathroom_small_cnt  unit_cnt
   patio_sqft  year_built  \
39  0     96370.0       0.0                 NaN       1.0
                  NaN      1959.0
40  1     96962.0       0.0                 1.0       NaN
                  NaN      2014.0
41  2     96293.0       0.0                 NaN       1.0
                  NaN      1940.0
42  3     96222.0       0.0                 NaN       1.0
                  NaN      1987.0
43  4     96961.0       8.0                 1.0       NaN
                  NaN      1981.0
44
45     story_cnt  tax_structure  tax_parcel  tax_year   tax_land
   tax_property  \
46  0        NaN       122754.0    360170.0    2015.0   237416.0
      6735.879883
47  1        NaN       346458.0    585529.0    2015.0   239071.0
     10153.019531
48  2        NaN        61994.0    119906.0    2015.0    57912.0
     11484.480469
49  3        NaN       171518.0    244880.0    2015.0    73362.0
      3048.739990
50  4        2.0       169574.0    434551.0    2015.0   264977.0
      5488.959961
51
52     tax_overdue_flag  tax_overdue_year      census_2
   avg_garage_size  \
53  0               NaN               NaN  6.037107e+13
                  NaN
54  1               NaN               NaN           NaN
                  234.0
```

```
55  2                 NaN             NaN  6.037464e+13
                      NaN
56  3                 NaN             NaN  6.037296e+13
                      NaN
57  4                 NaN             NaN  6.059042e+13
                      299.0
58
59     property_tax_per_sqft  location_1   location_2
    location_3   location_4  \
60  0                 3.999929 -84207544.0  152769536.0 -
    24963276.0  93525260.0
61  1                 4.486531 -84009432.0  151345664.0 -
    25170656.0  92506896.0
62  2                 5.180190 -84038720.0  152311344.0 -
    24951204.0  93223828.0
63  3                 3.633778 -84553200.0  152064800.0 -
    25398700.0  92910300.0
64  4                 2.404275 -84214592.0  151185872.0 -
    25364472.0  92335760.0
65
66     missing_finished_area  missing_total_area
    missing_bathroom_cnt_calc  \
67  0                     0.0                 1.0
                          0.0
68  1                     0.0                 1.0
                          0.0
69  2                     0.0                 1.0
                          0.0
70  3                     0.0                 1.0
                          0.0
71  4                     0.0                 1.0
                          0.0
72
73     derived_room_cnt  avg_area_per_room
    derived_avg_area_per_room  \
74  0                 5.0                NaN                  336
    .799988
75  1                 7.5                NaN                  301
    .733337
76  2                 5.0                NaN                  443
    .399994
77  3                 4.0                NaN                  209
    .750000
78  4                 6.5            285.375                  351
    .230774
79
80     region_zip-groupcnt  region_zip-lot_sqft-diff
    region_zip-lot_sqft-percent  \
81  0                 14719.0             -13398.968750
                      -0.640273
```

```
82  1               17682.0                    -2715.032715
                            -0.427024
83  2                4422.0                   -14927.021484
                            -0.566490
84  3                7293.0                   -43346.804688
                            -0.379550
85  4                9875.0                    -1155.377441
                            -0.161470
86
87      region_zip-year_built-diff   region_zip-
    finished_area_sqft_calc-diff  \
88  0               -3.998413
                                        -247.725464
89  1               35.535156
                                         526.538208
90  2              -12.917847
                                        -173.867432
91  3               21.690186
                                        -782.150757
92  4                0.695679
                                         244.801147
93
94      region_zip-finished_area_sqft_calc-percent   region_zip-
    tax_structure-diff  \
95  0                                   -0.128241
                        -50475.015625
96  1                                    0.303225
                        213678.171875
97  2                                   -0.072721
                       -236757.281250
98  3                                   -0.482466
                         30903.765625
99  4                                    0.120107
                        -50359.125000
100
101     region_zip-tax_structure-percent   region_zip-tax_land-
    diff  \
102 0                         -0.291377                    51026.
    421875
103 1                          1.609267                    16302.
    671875
104 2                         -0.792490                   -427605.
    093750
105 3                          0.219777                   -129440.
    812500
106 4                         -0.228975                   -195977.
    250000
107
108     region_zip-tax_land-percent   region_zip-tax_property-
    diff  \
```

```
109 0                          0.273762                      2047.
    035645
110 1                          0.073182                      6339.
    847656
111 2                         -0.880721                      1845.
    573242
112 3                         -0.638259                     -1337.
    844971
113 4                         -0.425156                     -2742.
    872070
114
115     region_zip-tax_property-percent  region_zip-
    property_tax_per_sqft-diff  \
116 0                          0.436576
                                        1.521634
117 1                          1.662618
                                        2.160548
118 2                          0.191471
                                        1.178391
119 3                         -0.304986
                                        0.830251
120 4                         -0.333203
                                       -1.339566
121
122     region_zip-property_tax_per_sqft-percent  year  month
    quarter
123 0                                    0.613984     0      1
              1
124 1                                    0.928875     0      1
              1
125 2                                    0.294465     0      1
              1
126 3                                    0.296145     0      1
              1
127 4                                   -0.357805     0      1
              1
128 0     0.0276
129 1    -0.1684
130 2    -0.0040
131 3     0.0218
132 4    -0.0050
133 Name: logerror, dtype: float32
134 X_train shape: (131426, 69)
135 y_train shape: (131426,)
136 X_val shape: (33578, 69)
137 y_val shape: (33578,)
138 [0, 13, 21, 66, 67, 68]
139 Train score: 5.123913700391014
140 Val score: 6.867147035970136
141 finished_area_sqft_calc: 4.031489261537851
```

```
142 month: 3.8433361292251567
143 lot_sqft: 3.516320396874277
144 year_built: 3.298611677986749
145 region_zip-finished_area_sqft_calc-percent: 3.
    181865452672254
146 region_zip-tax_property-percent: 3.1107117086903453
147 region_zip-tax_land-percent: 3.0811252446142863
148 pool_cnt: 3.080630914055197
149 tax_parcel: 2.943742150615708
150 region_zip-property_tax_per_sqft-percent: 2.
    9092672800344515
151 catboost_X: (164299, 69)
152 catboost_y: (164299,)
153 {'loss_function': 'MAE', 'eval_metric': 'MAE', 'nan_mode
    ': 'Min', 'random_seed': 0, 'iterations': 1000, '
    learning_rate': 0.015, 'border_count': 254, 'max_depth': 6
    , 'random_strength': 1, 'l2_leaf_reg': 5, '
    bagging_temperature': 1}
154 sanity check score: 6.827995571922846
155 Start model 0 (2016)
156 Start model 0 (2017)
157 Length of submission DataFrame: 2985217
158 Submission header:
159     ParcelId   201610   201611   201612   201710   201711
    201712
160 0  10754147  -0.0004  -0.0004  -0.0004   0.0004   0.0004  0.
    0004
161 1  10759547   0.0083   0.0083   0.0083   0.0124   0.0124  0.
    0124
162 2  10843547   0.0217   0.0217   0.0217   0.0412   0.0412  0.
    0412
163 3  10859147   0.0223   0.0223   0.0223   0.0223   0.0223  0.
    0223
164 4  10879947   0.0052   0.0052   0.0052   0.0077   0.0077  0.
    0077
165 catboost_X: (164299, 69)
166 catboost_y: (164299,)
167 Start training model 0
168 Start training model 1
169 Start training model 2
170 Start training model 3
171 Start training model 4
172 Start training model 5
173 Start training model 6
174 Start training model 7
175 model 0: 6.827995571922846
176 model 1: 6.828339569601775
177 model 2: 6.82575520899792
178 model 3: 6.82699496377183
179 model 4: 6.827676240485794
```

```
180 model 5: 6.828339213117111
181 model 6: 6.826599958096197
182 model 7: 6.82705851883068
183 Saved 8 CatBoost models to files.
184 Start model 0 (2016)
185 Start model 0 (2017)
186 Start model 1 (2016)
187 Start model 1 (2017)
188 Start model 2 (2016)
189 Start model 2 (2017)
190 Start model 3 (2016)
191 Start model 3 (2017)
192 Start model 4 (2016)
193 Start model 4 (2017)
194 Start model 5 (2016)
195 Start model 5 (2017)
196 Start model 6 (2016)
197 Start model 6 (2017)
198 Start model 7 (2016)
199 Start model 7 (2017)
200 Length of submission DataFrame: 2985217
201 Submission header:
202     ParcelId  201610  201611  201612  201710  201711
    201712
203 0  10754147  0.0060  0.0060  0.0060  0.0058  0.0058  0.
    0058
204 1  10759547  0.0086  0.0086  0.0086  0.0124  0.0124  0.
    0124
205 2  10843547  0.0219  0.0219  0.0219  0.0289  0.0289  0.
    0289
206 3  10859147  0.0208  0.0208  0.0208  0.0212  0.0212  0.
    0212
207 4  10879947  0.0076  0.0076  0.0076  0.0081  0.0081  0.
    0081
208
209 Process finished with exit code 0
210
```

Feature Importance

Feature	Importance
year_built	5462
lot_sqft	5297
region_zip-year_built-diff	4944
ed_avg_area_per_room	4557
region_zip-lot_sqft-diff	4254
month	4211
on_zip-tax_land-percent	4191
tax_structure	4111
region_zip-lot_sqft-percent	4075
finished_area_sqft_calc	4022
location_1	3945
_area_sqft_calc-percent	3938
property_tax_per_sqft	3818
region_zip	3681
ip-tax_structure-percent	3630
ty_tax_per_sqft-percent	3613
region_zip-groupcnt	3612
neighborhood_id	3561
region_zip-tax_land-diff	3473
tax_land	3421
tax_parcel	3280
property_tax_per_sqft-diff	3254
location_3	3186
latitude	3186
shed_area_sqft_calc-diff	3181
location_2	3159
on_zip-tax_structure-diff	3030
census_2	2991
zip-tax_property-percent	2833
longitude	2804
location_4	2784
tax_property	2685
city_id	2671
on_zip-tax_property-diff	2491
census_1	2209
landuse_type_id	1865
quality_id	1373
derived_room_cnt	1201
pool_cnt	1071
avg_area_per_room	967
quarter	757
tax_year	731
garage_sqft	691
unit_cnt	690
bedroom_cnt	689
tax_overdue_year	678
bathroom_cnt	651
avg_garage_size	622
missing_finished_area	535
pool_unk_3	533
heating_id	491
bathroom_full_cnt	430
year	397
tax_overdue_flag	379
floor1_sqft_unk	358
patio_sqft	345
floor1_sqft	325
pool_unk_2	242
missing_total_area	206
base_total_area	201
cooling_id	185
spa_flag	169
story_cnt	127
room_cnt	97
garage_cnt	90
fireplace_cnt	66
bathroom_small_cnt	14
sing_bathroom_cnt_calc	13
pool_total_size	9

Feature importance