In [1]:

```
import numpy as np
import pandas as pd
import os
from sklearn.metrics import roc_auc_score
```

executed in 11.9s, finished 13:16:32 2019-03-07

In [2]:

```
key = 'SmartScreen'
#val = 'ExistsNotSet'
val = 'existsnotset'
```

executed in 4ms, finished 13:16:52 2019-03-07

In [4]:

```
tr = pd.read_feather('../data/train.f')[[key]]
y_train = pd.read_feather('../data/target.f')['HasDetections']

te = pd.read_feather('../data/test.f')[[key]]
```

executed in 1m 52.0s, finished 13:23:29 2019-03-07

```
/home/Kazuki/anaconda3/lib/python3.6/site-packages/pandas/i
o/feather_format.py:124: FutureWarning: `nthreads` argument
is deprecated, pass `use_threads` instead
  nthreads=int_use_threads)
/home/Kazuki/anaconda3/lib/python3.6/site-packages/pyarrow/
pandas_compat.py:751: FutureWarning: .labels was deprecated
in version 0.24.0. Use .codes instead.
  labels, = index.labels
```

train ratio

In [9]:

```
((tr[key]==val)*1).value_counts(normalize=True)
```

executed in 566ms, finished 13:35:12 2019-03-07

Out[9]:

```
0    0.882734
1    0.117266
Name: SmartScreen, dtype: float64
```

train auc

In [10]:

```
roc_auc_score(y_train, ((tr[key]==val)*1))
```

executed in 3.80s, finished 13:35:16 2019-03-07

Out[10]:

```
0.5721349663572715
```

test ratio

In [11]:

```
((te[key]==val)*1).value_counts(normalize=True)
```

executed in 439ms, finished 13:35:29 2019-03-07

Out[11]:

```
0    0.923542
1    0.076458
Name: SmartScreen, dtype: float64
```

Adjust train to test

In [12]:

```
x = 394099
(1046183 - x) / (8921483 - x)
```

executed in 5ms, finished 13:35:54 2019-03-07

Out[12]:

```
0.07646940726487748
```

In [13]:

```
ix = tr[tr[key]==val].head(x).index
tr_ = tr.drop(ix)
y_train_ = y_train.drop(ix)
```

executed in 5.02s, finished 13:36:08 2019-03-07

In [14]:

```
print( ((tr_[key]==val)*1).value_counts(normalize=True) )
print( roc_auc_score(y_train_, ((tr_[key]==val)*1)) )
```

executed in 4.11s, finished 13:36:19 2019-03-07

```
0    0.923531
1    0.076469
Name: SmartScreen, dtype: float64
0.549255360699972
```

In [ ]:

```

```