Total number of data points: 2598

Number of data that should trigger buy: 470

Number of data that should not trigger buy: 2128

So the data is quite unbalanced.

**Evaluation metric**

For this case, we are more concerned about false negatives:

i.e. declaring a losing position to be ‘should buy’, as this results in capital loss.

Precision: percentage of true positives among all positives declared:

i.e. correct buying calls identified against all buying calls.

Recall: percentage of true positives among all true positive cases.

i.e. correct buying calls identified against all buying opportunities.

Consider putting more weight on precision: e.g. F0.5 score

**CV**

StratifiedShuffleSplit

Dropping volume

Looking at volume statistics:

count 2.598000e+03

mean 1.107663e+08

std 1.366913e+08

min 0.000000e+00

25% 0.000000e+00

50% 0.000000e+00

75% 2.157447e+08

max 1.019911e+09

More than 50% of them are missing. Therefore the volume information is dropped.