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.

**Feature**

----------------------

AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None,

learning\_rate=1.0, n\_estimators=50, random\_state=None)

precision: 0.5

recall: 0.0567375886525

----------------------

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1,

penalty='l2', random\_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm\_start=False)

precision: 0.25

recall: 0.00709219858156

----------------------

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='gini',

max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

precision: 0.25

recall: 0.0354609929078

=============================================

----------------------

AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None,

learning\_rate=1.0, n\_estimators=50, random\_state=None)

precision: 0.454545454545

recall: 0.0354609929078

----------------------

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1,

penalty='l2', random\_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm\_start=False)

precision: 0.0

recall: 0.0

----------------------

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='gini',

max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

precision: 0.421052631579

recall: 0.0567375886525

=============================================

----------------------

AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None,

learning\_rate=1.0, n\_estimators=50, random\_state=None)

precision: 0.25

recall: 0.0354609929078

----------------------

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1,

penalty='l2', random\_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm\_start=False)

precision: 0.0833333333333

recall: 0.00709219858156

----------------------

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='gini',

max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

precision: 0.28

recall: 0.0496453900709

=============================================

----------------------

AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None,

learning\_rate=1.0, n\_estimators=50, random\_state=None)

precision: 0.28

recall: 0.0496453900709

----------------------

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1,

penalty='l2', random\_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm\_start=False)

precision: 0.5

recall: 0.00709219858156

----------------------

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='gini',

max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

precision: 0.333333333333

recall: 0.0992907801418

=============================================

----------------------

AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None,

learning\_rate=1.0, n\_estimators=50, random\_state=None)

precision: 0.407407407407

recall: 0.0780141843972

----------------------

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1,

penalty='l2', random\_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm\_start=False)

precision: 0.0

recall: 0.0

----------------------

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='gini',

max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

precision: 0.458333333333

recall: 0.0780141843972

Adding RSI:

precision: 0.583333333333

recall: 0.0496453900709

F-score: 0.185185185185

=============================================

precision: 0.6

recall: 0.0425531914894

F-score: 0.165745856354

=============================================

precision: 0.181818181818

recall: 0.0283687943262

F-score: 0.0873362445415

=============================================

precision: 0.238095238095

recall: 0.0354609929078

F-score: 0.111111111111

=============================================

precision: 0.357142857143

recall: 0.0709219858156

F-score: 0.197628458498

=============================================

Done!

will use AdaBoost algorithm.