feature:

1. submission\_list( former submissions of a image)

2. title\_len\_list (the title length)

3. raw\_time\_list ( four time periods during a day, and the timezone(-7,-8, a vector containing 6 binary elements)

4. date\_list(month1-11, day(Mon-Sat))

5. subreddit\_list (subreddit submissions)

6. community effects(whether the submission was large before in popular subreddits)

dataset:

1. train set : test set = 3 : 1 (99230 :33077)

2. 10 fold cross validation on train set

3. Normalize: StandardScalar

predict label : whether or not the submission is good(large)

A. binary classification:

**a) whether or not the submission is good**

1 : data[i]['number\_of\_comments']\*0.2 + data[i]['total\_votes']\*0.8 > 36 (large)

0. otherwise (not large)

**b) model used**

# 

# 3. LR

**validation**

Accuracy: 0.6676 (+/- 0.0117)

Recall: 0.6675 (+/- 0.0117)

Precision: 0.6677 (+/- 0.0118)

**test**

precision: 66.51%,

recall: 68.90%

accuracy: 67.32%

Theta :

[-0.32147647, -0.02664795, 0.04501319, 0.00388396, 0.0233998 ,

0.01831231, -0.01864645, 0.37314239, 0.08381971, 0.1221402 ,

0.1425188 , 0.09996822, 0.09172504, 0.05362183, -0.08433367,

-0.13157605, -0.18554324, -0.14712468, 0.00440766, -0.03267718,

-0.05198689, -0.04046564, -0.05997207, -0.03795328, 0.02877001,

-0.29633535, -0.1396907 ]

# 5. DT

**validation**

Accuracy: 0.6742 (+/- 0.0101)

Recall: 0.6740 (+/- 0.0101)

Precision: 0.6745 (+/- 0.0102)

**test**

precision: 66.38%,

recall: 70.41%

accuracy: 67.58%

# 6. ADA

**validation**

Accuracy: 0.6742 (+/- 0.0101)

Recall: 0.7107 (+/- 0.0098)

Precision: 0.7109 (+/- 0.0098)

**test**

precision: 71.82%,

recall: 70.60%

accuracy: 71.64%(improved 0.13%)