Problem(a)

Naïve Bayes

Spam 1 3 0.1 ==========Final report========== Accuracy:0.605 0.000 Precision:0.764 0.000 Recall:0.385 0.000 Area under ROC 0.423Volcanoes 1 3 0.1 ==========Final report========== Accuracy:0.630 0.000 Precision:0.463 0.000 Recall:0.817 0.000 Area under ROC 0.500Voting 1 3 0.1 ==========Final report========== Accuracy:0.982 0.000 Precision:0.990 0.000 Recall:0.969 0.000 Area under ROC 0.911

Logistic Regression

Spam 1 0.1 ==========Final report========== Accuracy:0.541 0.000 Precision:0.540 0.000 Recall:0.987 0.000 Area under ROC 0.655Volcanoes 1 0.1 ==========Final report========== Accuracy:0.688 0.000 Precision:0.514 0.000 Recall:0.893 0.000 Area under ROC 0.337Voting 1 0.1 ==========Final report========== Accuracy:0.991 0.000 Precision:1.000 0.000 Recall:0.980 0.000 Area under ROC 0.996

Problem(b)

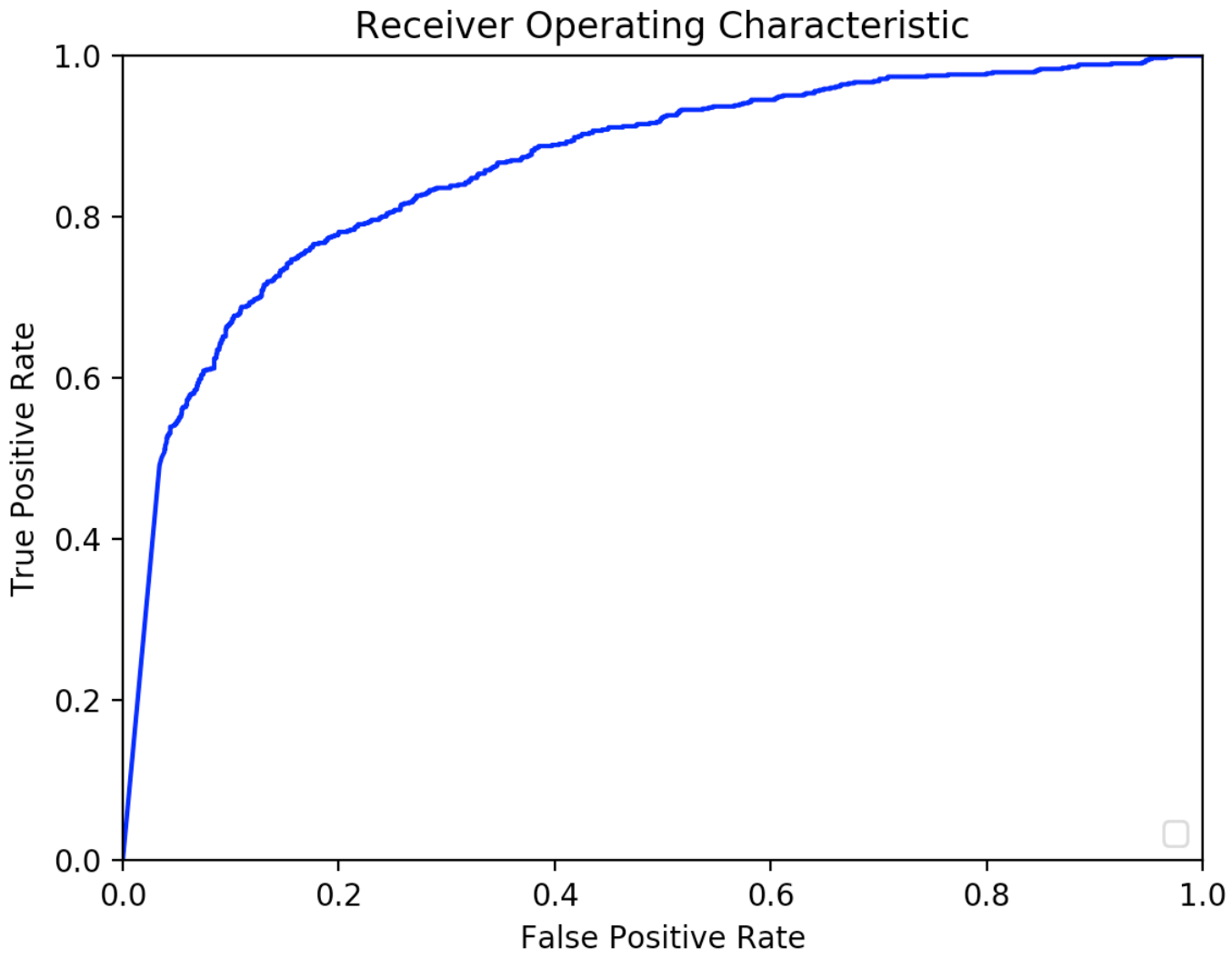
Now we'll check the effort of λ

next 3 cell will show λ = 0, 0.01, 0.1, 1, let's take the volcanoes dataset as an example

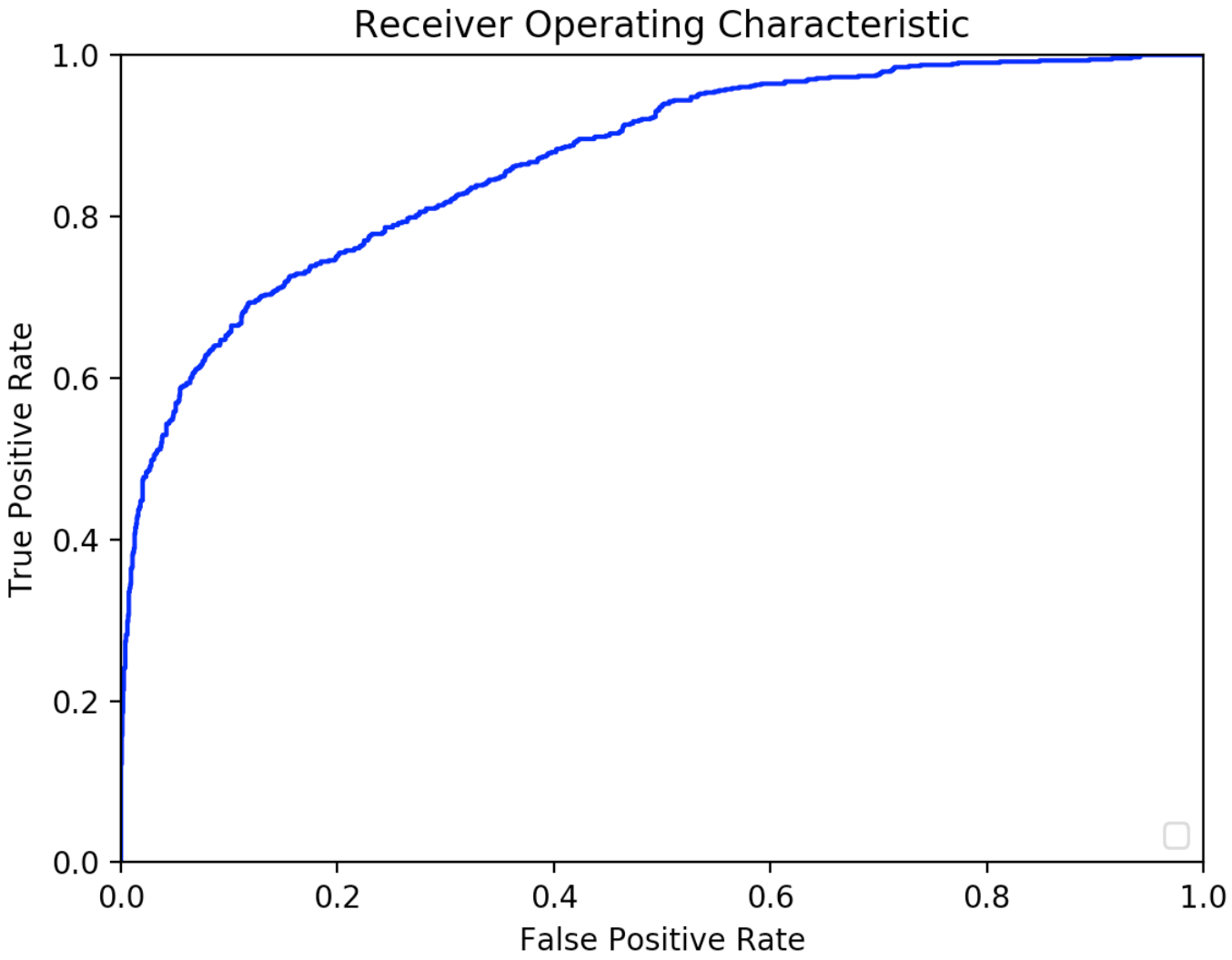
In [8]:

In [10]:
```
!python logreg.py /Users/victor/Desktop/CWRU/CSDS440\ Machine\ Learnin
g/programming1/440data/volcanoes 1 0
```

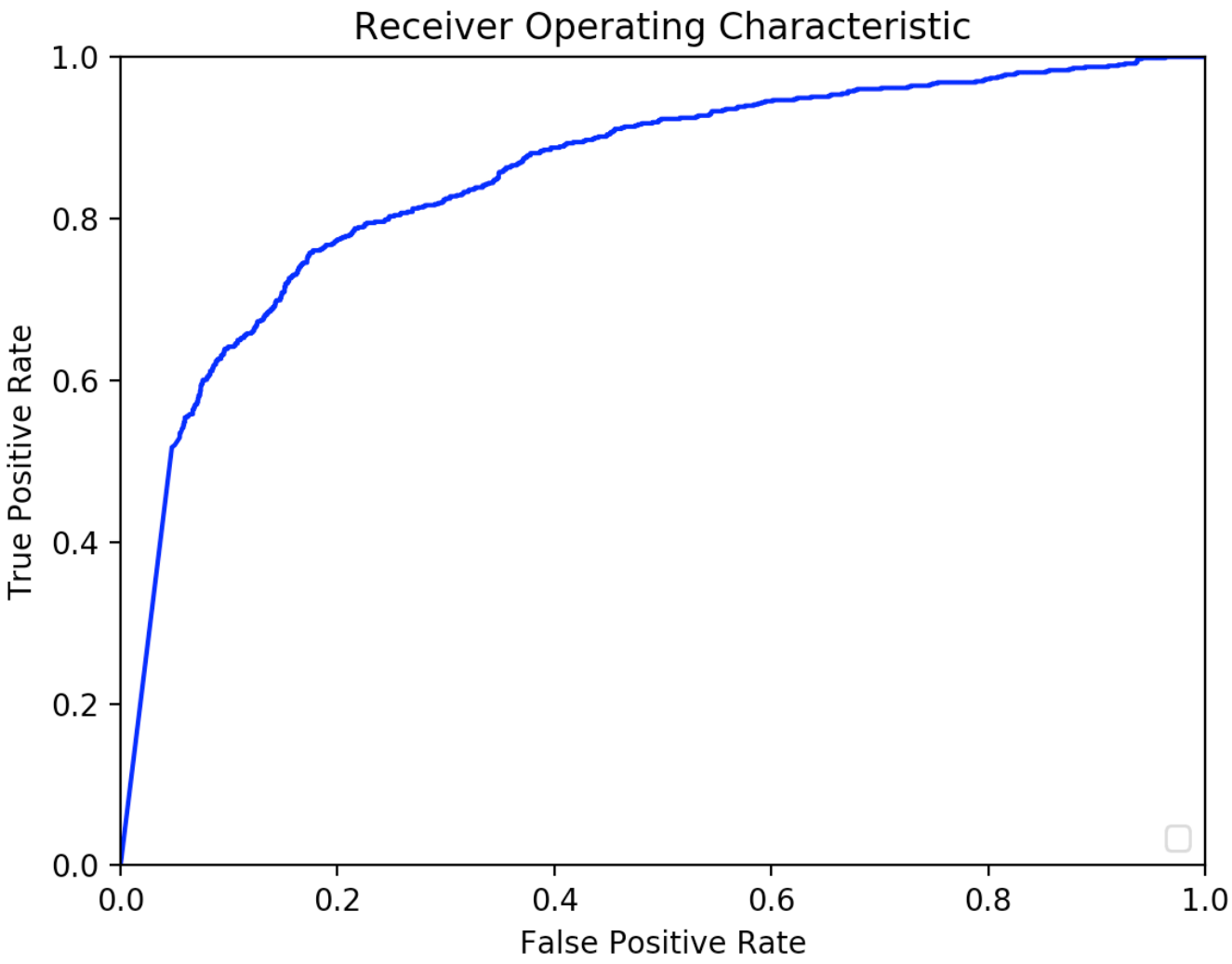
```
No handles with labels found to put in legend.
Figure(640x480)
==========Final report==========
Accuracy:0.727 0.000
Precision:0.555 0.000
Recall:0.854 0.000
Area under ROC 0.368
```



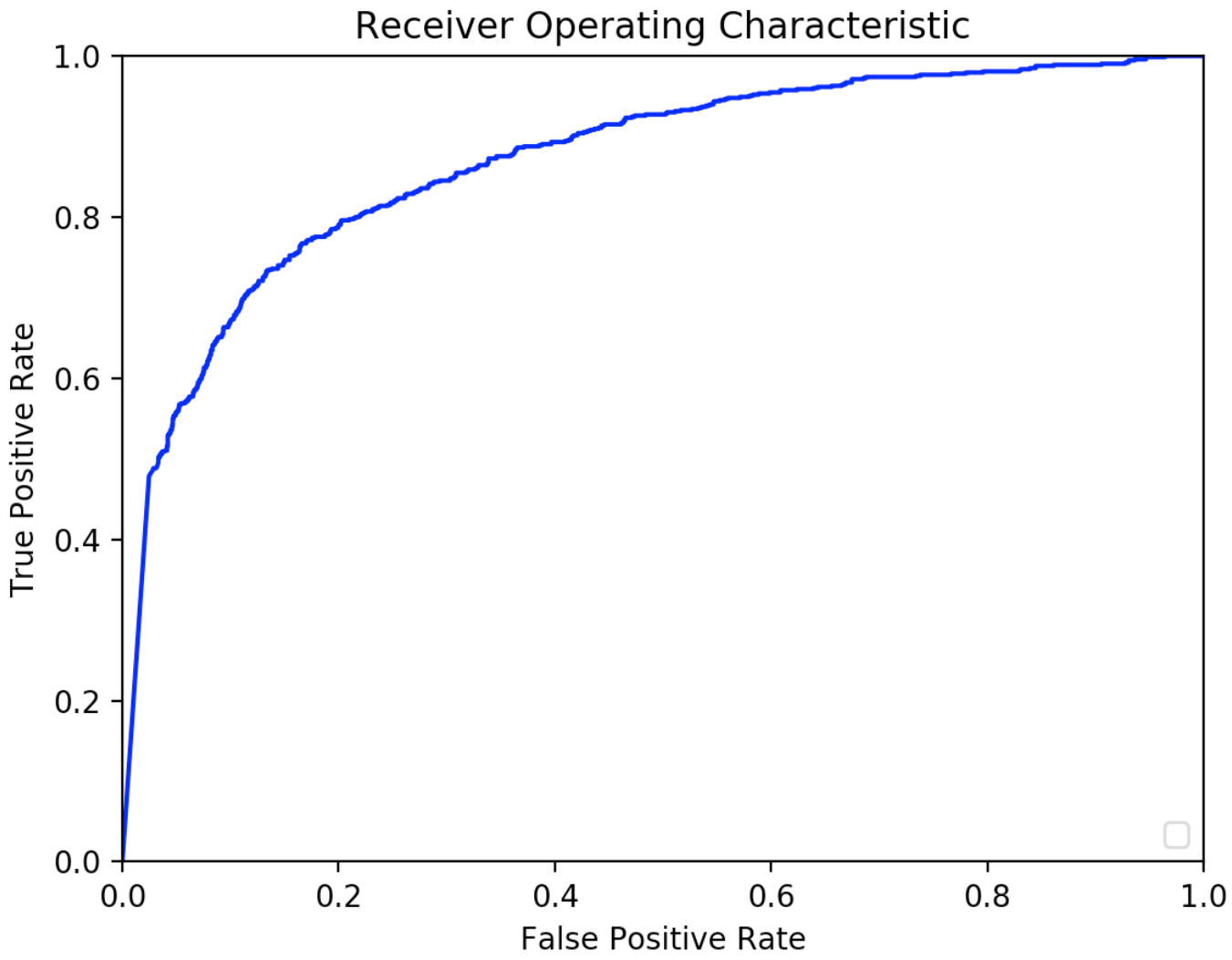
volcanoes 0 0.01 ==========Fold report========== Accuracy:0.716 Precision:0.565
Recall:0.808 ==========Fold report========== Accuracy:0.679 Precision:0.487
Recall:0.803 ==========Fold report========== Accuracy:0.646 Precision:0.484
Recall:0.899 ==========Fold report========== Accuracy:0.709 Precision:0.944
Recall:0.116 ==========Fold report========== Accuracy:0.706 Precision:0.525
Recall:0.895 ==========Final report========== Accuracy:0.691 0.026 Precision:0.601 0.174
Recall:0.704 0.297 Area under ROC 0.433



volcanoes 0 0.1 ==========Fold report========== Accuracy:0.649 Precision:0.498
Recall:0.840 ==========Fold report========== Accuracy:0.787 Precision:0.636
Recall:0.715 ==========Fold report========== Accuracy:0.798 Precision:0.915
Recall:0.436 ==========Fold report========== Accuracy:0.760 Precision:0.649
Recall:0.582 ==========Fold report========== Accuracy:0.798 Precision:0.827
Recall:0.469 ==========Final report========== Accuracy:0.758 0.057 Precision:0.705 0.148
Recall:0.608 0.152 Area under ROC 0.528



volcanoes 0 1 ==========Fold report========== Accuracy:0.796 Precision:0.715
Recall:0.692 ==========Fold report========== Accuracy:0.756 Precision:0.967
Recall:0.212 ==========Fold report========== Accuracy:0.778 Precision:0.652
Recall:0.718 ==========Fold report========== Accuracy:0.655 Precision:0.484
Recall:0.829 ==========Fold report========== Accuracy:0.798 Precision:0.835
Recall:0.462 ==========Final report========== Accuracy:0.757 0.053 Precision:0.731 0.164
Recall:0.582 0.221 Area under ROC 0.505



In [ ]:

as we can see, when λ is rising, accuracy is increasing at first and then decreasing

Problem(c)

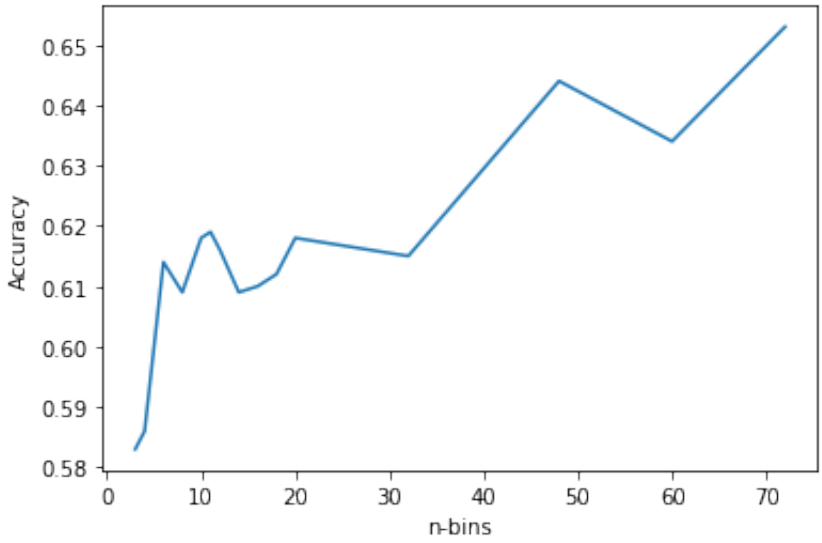
As we go through the experiment, what came to us is how parameter n-bins make differences to Accuracy.Assume that when n-bins is increasing, the classifier can keep more information and make better decisions. Or it's posible that more number of bins may cause over-fitting problem.So we try to figure out that.

spam 1 3 -1 ==========Final report========== Accuracy:0.583 0.000 Precision:0.768 0.000

Recall:0.371 0.000 Area under ROC 0.434spam 1 4 -1 ==========Final report========== Accuracy:0.586 0.000 Precision:0.828 0.000 Recall:0.333 0.000 Area under ROC 0.441spam 1 6 -1 ==========Final report========== Accuracy:0.614 0.000 Precision:0.805 0.000 Recall:0.414 0.000 Area under ROC 0.437spam 1 8 -1 ==========Final report========== Accuracy:0.609 0.000 Precision:0.803 0.000 Recall:0.405 0.000 Area under ROC 0.460spam 1 10 -1 ==========Final report========== Accuracy:0.619 0.000 Precision:0.833 0.000 Recall:0.404 0.000 Area under ROC 0.451spam 1 11 -1 ==========Final report========== Accuracy:0.616 0.000 Precision:0.831 0.000 Recall:0.399 0.000 Area under ROC 0.450spam 1 12 -1 ==========Final report========== Accuracy:0.609 0.000 Precision:0.801 0.000 Recall:0.405 0.000 Area under ROC 0.463spam 1 14 -1 ==========Final report========== Accuracy:0.618 0.000 Precision:0.828 0.000 Recall:0.406 0.000 Area under ROC 0.466spam 1 16 -1 ==========Final report========== Accuracy:0.610 0.000 Precision:0.799 0.000 Recall:0.410 0.000 Area under ROC 0.459spam 1 18 -1 ==========Final report========== Accuracy:0.612 0.000 Precision:0.816 0.000 Recall:0.401 0.000 Area under ROC 0.458spam 1 20 -1 ==========Final report========== Accuracy:0.618 0.000 Precision:0.824 0.000 Recall:0.408 0.000 Area under ROC 0.470spam 1 32 -1 ==========Final report========== Accuracy:0.615 0.000 Precision:0.817 0.000 Recall:0.408 0.000 Area under ROC 0.501spam 1 48 -1 ==========Final report========== Accuracy:0.644 0.000 Precision:0.848 0.000 Recall:0.447 0.000 Area under ROC 0.500spam 1 60 -1 ==========Final report========== Accuracy:0.634 0.000 Precision:0.869 0.000 Recall:0.413 0.000 Area under ROC 0.500spam 1 72 -1 ==========Final report========== Accuracy:0.653 0.000 Precision:0.849 0.000 Recall:0.466 0.000 Area under ROC 0.500

In [39]:
```
import numpy as np
import matplotlib.pyplot as plt
n=np.asarray([3,4,6,8,10,11,12,14,16,18,20,32,48,60,72])
accuracy=np.asarray([0.583,0.586,0.614,0.609,0.618,0.619,0.616,0.609,0
.610,0.612,0.618,0.615,0.644,0.634,0.653])
plt.xlabel("n-bins")
plt.ylabel("Accuracy")
plt.plot(n,accuracy)
```

Out[39]:
```
[<matplotlib.lines.Line2D at 0x11a279d0>]
```



Analysis

When putting all these data together, accuracy is keep increasing when n-bins go up to 72. So in spam dataset,I can make the conclusion that, instead of over-fit, naive bayes classifier can get more information as the number of bins increasing in certain limits.

In [ ]: