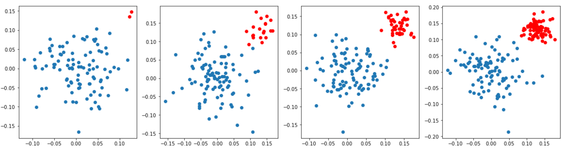
In [6]:
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import SGDClassifier
from sklearn.linear_model import LogisticRegression
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler, Normalizer
import matplotlib.pyplot as plt
from sklearn.svm import SVC
import warnings
warnings.filterwarnings("ignore")
```

In [7]:
```
def draw_line(coef,intercept, mi, ma):
    # for the separating hyper plane ax+by+c=0, the weights are [a, b] and the intercept is c
    # to draw the hyper plane we are creating two points
    # 1. ((b*min-c)/a, min) i.e ax+by+c=0 ==> ax = (-by-c) ==> x = (-by-c)/a here in place of y we are keeping the minimum value of y
    # 2. ((b*max-c)/a, max) i.e ax+by+c=0 ==> ax = (-by-c) ==> x = (-by-c)/a here in place of y we are keeping the maximum value of y
    points=np.array([[((-coef[1]*mi - intercept)/coef[0]), mi],[((-coef[1]*ma - intercept)/coef[0]), ma]])
    plt.plot(points[:,0], points[:,1])
```

What if Data is imabalanced

1. As a part of this task you will observe how linear models work in case of data imbalanced
2. observe how hyper plane is changs according to change in your learning rate.
3. below we have created 4 random datasets which are linearly separable and having class imbalance
4. in the first dataset the ratio between positive and negative is 100 : 2, in the 2nd data its 100:20, in the 3rd data its 100:40 and in 4th one its 100:80

In [8]:
```
# here we are creating 2d imbalanced data points
ratios = [(100,2), (100, 20), (100, 40), (100, 80)]
plt.figure(figsize=(20,5))
for j,i in enumerate(ratios):
    plt.subplot(1, 4, j+1)
    X_p=np.random.normal(0,0.05,size=(i[0],2))
    X_n=np.random.normal(0.13,0.02,size=(i[1],2))
    y_p=np.array([1]*i[0]).reshape(-1,1)
    y_n=np.array([0]*i[1]).reshape(-1,1)
    X=np.vstack((X_p,X_n))
    y=np.vstack((y_p,y_n))
    plt.scatter(X_p[:,0],X_p[:,1])
    plt.scatter(X_n[:,0],X_n[:,1],color='red')
plt.show()
```



> your task is to apply SVM (sklearn.svm.SVC) and LR (sklearn.linear_model.LogisticRegression) with different regularization strength [0.001, 1, 100]

Task 1: Applying SVM

1. you need to create a grid of plots like this this

in each of the cell[i][j] you will be drawing the hyper plane that you get after applying SVM on ith dataset and jth learnig rate

i.e

Plane(SVM().fit(D1, C=0.001))	Plane(SVM().fit(D1, C=1))	Plane(SVM().fit(D1, C=100))	
Plane(SVM().fit(D2, C=0.001))	Plane(SVM().fit(D2, C=1))	Plane(SVM().fit(D2, C=100))	
Plane(SVM().fit(D3, C=0.001))	Plane(SVM().fit(D3, C=1))	Plane(SVM().fit(D3, C=100))	
Plane(SVM().fit(D4, C=0.001))	Plane(SVM().fit(D4, C=1))	Plane(SVM().fit(D4, C=100))	

if you can do, you can represent the support vectors in different colors, which will help us understand the position of hyper plane

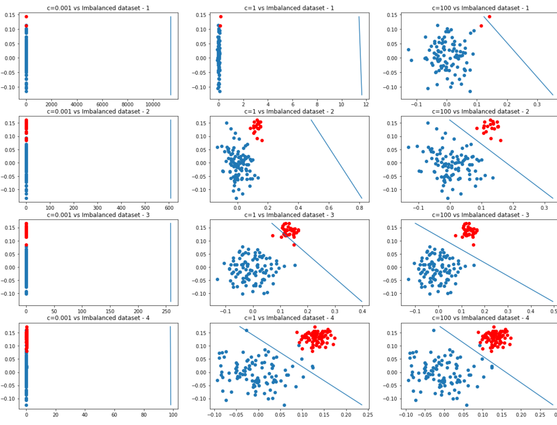
Write in your own words, the observations from the above plots, and what do you think about the position of the hyper plane

check the optimization problem here https://scikit-learn.org/stable/modules/svm.html#mathematical-formulation

if you can describe your understanding by writing it on a paper and attach the picture, or record a video upload it in assignment.

In [9]:
```
dataset=[]#Creating different datasets and storing the values
ratios = [(100,2), (100, 20), (100, 40), (100, 80)]
for j,i in enumerate(ratios):
    X_p=np.random.normal(0,0.05,size=(i[0],2))
    X_n=np.random.normal(0.13,0.02,size=(i[1],2))
    y_p=np.array([1]*i[0]).reshape(-1,1)
    y_n=np.array([0]*i[1]).reshape(-1,1)
    X=np.vstack((X_p,X_n))
    y=np.vstack((y_p,y_n))
    dataset.append([X,y,X_p,X_n,y_p,y_n])
```

In [10]:
```
#Applying the SVM on various imbalanced dataset taking
plt.figure(figsize=(20,15))
i=1
d=0
for X,y,X_p,X_n,y_p,y_n in dataset:
    d=d+1
    for C in [0.001,1,100]:
        plt.subplot(4,3,i)
        i=i+1
        clf=SVC(C=C,kernel='linear',random_state=1)
        clf.fit(X,y)
        draw_line(clf.coef_.ravel(),clf.intercept_,np.min(X),np.max(X))
        plt.scatter(X_p[:,0],X_p[:,1])
        plt.scatter(X_n[:,0],X_n[:,1],color='red')
        plt.title(f'C={c} vs Imbalanced dataset - {d}' )
plt.show()
```



Summary

1.When we have a high imbalanced dataset and a small C value then the SVM is not working well. The Planes are not in the correct area, they are not even classifying the model which is underfitting the model

2.When we started to tune the C value i.e by increasing the C value the model started to classify the most of the points correctly

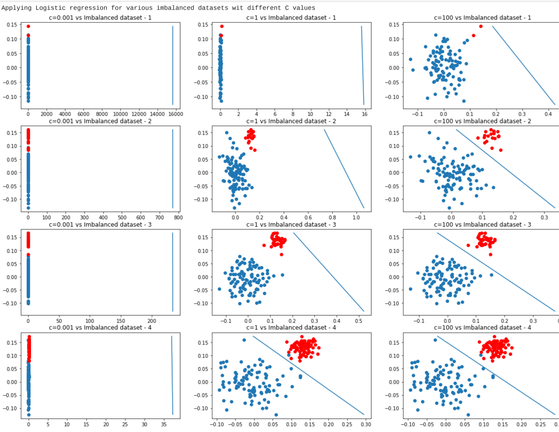
Task 2: Applying LR

you will do the same thing what you have done in task 1.1, except instead of SVM you apply logistic regression

these are results we got when we are experimenting with one of the model

In [11]:
```
print('Applying Logistic regression for various imbalanced datasets wit different C values')
plt.figure(figsize=(20,15))
i=1
d=0
for X,y,X_p,X_n,y_p,y_n in dataset:
    d=d+1
    for C in [0.001,1,100]:
        plt.subplot(4,3,i)
        i=i+1
        clf=LogisticRegression(C=C,random_state=1)
        clf.fit(X,y)
        draw_line(clf.coef_.ravel(),clf.intercept_,np.min(X),np.max(X))
        plt.scatter(X_p[:,0],X_p[:,1])
        plt.scatter(X_n[:,0],X_n[:,1],color='red')
        plt.title(f'C={c} vs Imbalanced dataset - {d}' )
plt.show()
```

```
Applying Logistic regression for various imbalanced datasets wit different C values
```



Summary

1. When the data is highly imbalanced and if the C value is very low then the hyperplane is very far away from the dataset and not doing any classification

2.Once we start to tune the C value and balance the dataset then the model will perform well , in the above figure also we can observe the same

In [ ]: