Data Mining, CSCI 347, Fall 2019
Evaluation – Counting Costs, Nov. 13

1. A confusion matrix for a model is given. Tell the accuracy of the model.

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

Accuracy = 140/200 = 70%

2. The same confusion matrix is given below on the left. On the right is a confusion matrix for a random predictor. Calculate the Kappa statistic to measure the relative improvement of the model over a random predictor.

Model Predictions

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

Random Predictor

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	60	30	10	*100*
	b	36	18	6	*60*
	c	24	12	4	*40*
	total	*120*	*60*	*20*	

The correct predictions of our model are shown in the diagonal of the left matrix: 88+40+12 = 140

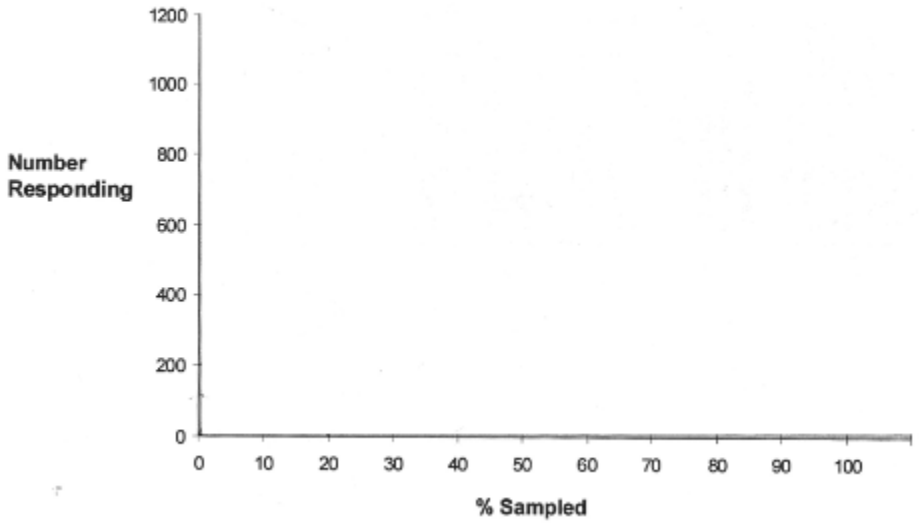
The random predictions are shown in the diagonal of the right matrix: 60+18+4 = 82

Kappa statistic $\frac{140 - 82}{200 - 82} = \frac{58}{118} = 0.49$

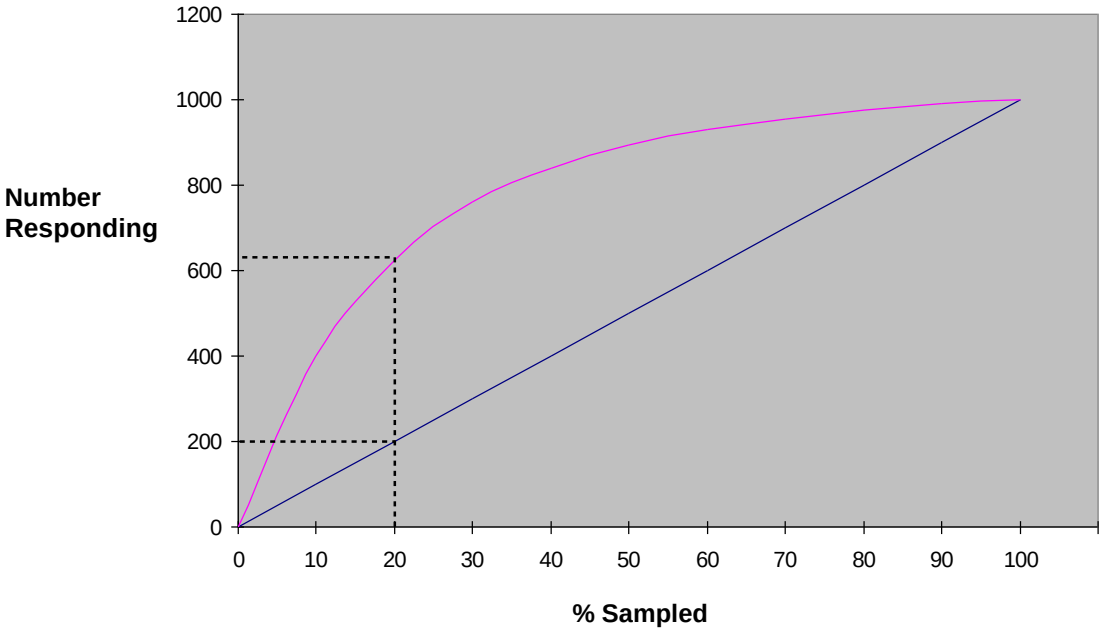
Thus the accuracy of the model is 70%, however, it is only 49% over a random predictor.

3. Example: Say that we are going to do a direct mailing and we have a million addresses in our database. In general we know that 0.1% (0.001) household will respond to our mailing. Say that using data mining, we have an algorithm that identifies a subset of 100,000 of the most promising addresses, where these households are likely to responds at a rate of 0.4% Another model identifies a subset of 400,000 household where 0.2% are likely to respond.

   Create a lift chart to visualize this information.
   x axis is the percent sampled
   y axis is number of true positives



Answer:



4. Say when predicting the recurrence of breast cancer, 70% of the women that survived breast cancer had not had recurrence within 5 years. The other 30% had. Call the recurrence of breast cancer a positive. For parts a and b, say that there are 100 women.

a. Calculate the F1 score of a model that always predicts that there will be no recurrence.

If a model always predicted 'no' it would be correct 70% of the time.

		Predicted	
		Recurrence	**No Recurrence**
Actual	**Recurrence**	0	30
	No Recurrence	0	70

Precision: 0/(0+0) which isn't a number
Recall: 0/(0+30) or 0
F1: 2*(0*0)/(0+0) which isn't a number

b. Calculate the F1 score if a model always predicts that there will be a recurrence.

Alternatively, since an algorithm could always predict 'yes' there will be a recurrence. This will be right 30% of the time.

		Predicted	
		Recurrence	**No Recurrence**
Actual	**Recurrence**	30	0
	No Recurrence	70	0

Precision: 30/(30+70) = 0.3
Recall: 30/(30+0) =1.0, it is very sensitive and catches all recurrences ☺
F1: 2*(0.3*1.0)/(0.3+1.0) = 0.6/1.3 = 0.46

c. CART (Classification And Regression Trees) for 300 women, has the confusion matrix:

		Predicted	
		Recurrence	**No Recurrence**
Actual	**Recurrence**	77	13
	No Recurrence	39	171

Calculate the F1 score.

Precision: 77/(77+39) = 0.66
Recall: 77/(77+13) = 0.86
F1: 2*(0.66*0.86)/(0.66+0.86) = 1.14/1.52 = 0.75