GAIN STATISTIC

The Gain Statistic is used in Logistic Regression to assess a model's ability to rank order the data. It is more likely to be seen in banking and credit scoring than in other industries.

> *NOTE: The Cumulative Gain or "Gain" is often confused with the "Kolmogorov-Smirnov (KS)" Statistic. The KS statistic is similar to the Gain. The difference is that the KS value is derived from a ROC curve instead of a Cumulative Lift Chart. In the opinion of the instructor, the Gain and Cumulative Lift Chart are easier to interpret than KS and ROC curves.*

This is how the Cumulative Lift Curve is created:

1. Sort the predicted probabilities from highest to lowest.

2. Put the predicted values into "N" buckets (let's say N=10). So we put the data in 10 buckets with the highest scores in bucket 1 and the lowest scores in bucket 10.

3. Now we count the number of positive values in each bucket and calculate the cumulative differences. Here's how to do that:
    1. Let's say that bucket 1 has 30% of all the positive responses (but if the model was random, it would only have 10% of the positive responses). So the difference is 20% because 30% minus 10% is 20%.
    2. Then we move to bucket 2. In bucket 2, we have 20% of the positive values. So that means that in the bucket 1 and bucket 2 combined we have 30%+20%=50% of all the positive responses. However, if our model was random, we would have 10%+10%=20% of all positive responses. So now 50% cumulative responses minus 20% random would mean that the difference is 30%.
    3. Now we move to bucket 3. In bucket 3 we might see 15% of all the positives. So we add up bucket 1+ bucket 2 + bucket 3 = 30%+20%+15%=65%. But if the model was random it would be 10%+10%+10%=30%. So the difference is 35%.
    4. We keep doing this for all 10 buckets. Incidentally, when you get to bucket 10, you will have 100% of all the positives and the theoretical value will be 100% so the difference in the last bucket will always be 0.

4. OK, so now you list the Differences you calculated for all 10 buckets. Let's assume that the values we calculated were:
   1. Bucket 1: 20%
   2. Bucket 2: 30%
   3. Bucket 3: 35%
   4. Bucket 4: 39%
   5. Bucket 5: 35%
   6. Bucket 6: 30%
   7. Bucket 7: 24%
   8. Bucket 8: 17%
   9. Bucket 9: 9%
   10. Bucket 10: 0%
5. Now we just calculate the MAXIMUM Difference. In our example, that would be 39%

Therefore, the Gain is 39%

I put this example in an EXCEL spread sheet.

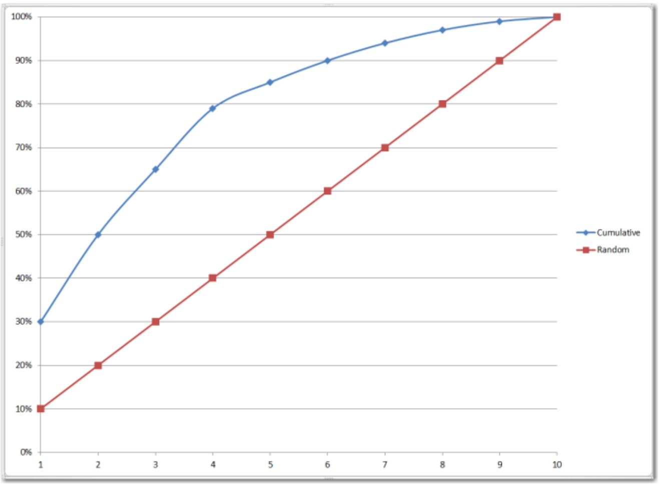
INTERPRET CUMULATIVE LIFT “GAINS” CURVES

The Cumulative Lift curve is often incorrectly called a “ROC” curve. It is not a ROC curve but it does look similar which is why the confusion occurs. The difference is that Cumulative Lift curves measure the effectiveness of a model at predicting the target. The ROC curve measures the tradeoff between a selecting as many True Positives as possible while avoiding False Positives.

In the opinion of the instructor, the Cumulative Lift curve is much easier to generate in EXCEL so we are going to use them. The real question is how do you interpret them? Well, it's pretty straight forward. I'll give you four scenarios:

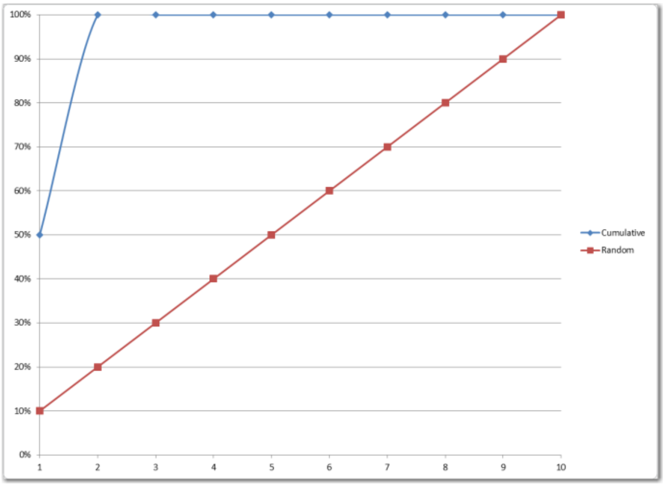
GOOD MODEL:

This model is good because there is a nice bow and all the lift is the beginning. That means that your model is good at rank ordering and when your model says that there is a high chance that a value is positive, then it is in fact positive.



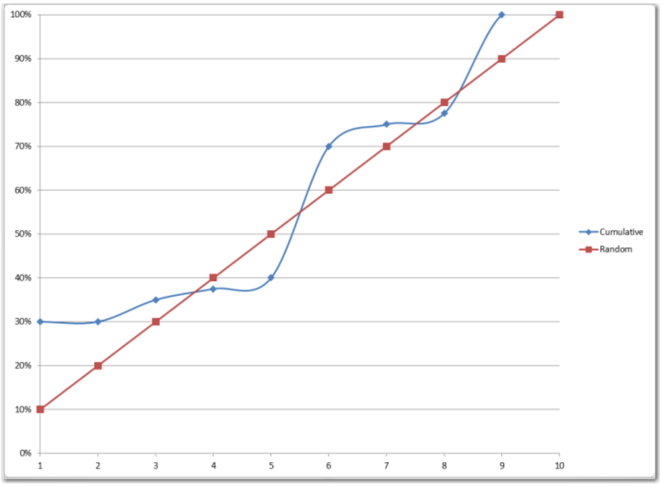
TOO PERFECT:

When your model is too perfect then you need to suspect trouble. For example, if you are predicting who will file an insurance claim, you might accidentally have a value of claim amount in your model. Then you might say, "if the person's claim amount is greater than 0 then they have a high chance of filing a claim .... no kidding!". So when your model is too perfect, suspect trouble!



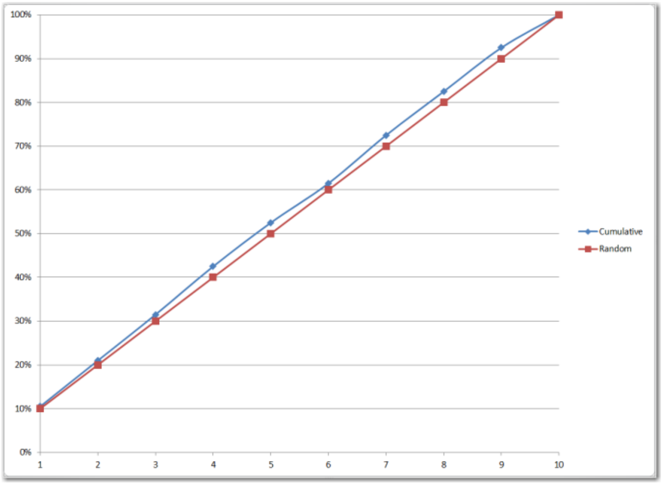
OVERFIT:

When you over fit your model, then your Cumulative Lift curve cuts through the random line. That should never happen. Your model should not ever be worse than random. Even if your lift and cumulative lift values are good, once you cut through the red line, then your model is no good. Count on it!



WORTHLESS:

If your model doesn't look too different from the random diagonal line, then it's a bad model. Your model should always be a lot better than pulling names out of a hat. That's what they are paying you to do!


Cumulative
Random