Two-by-Two Classification and Decile Tables - A Comparison
Bruce Ratner, Ph.D.

The purpose of this article is to compare and contrast two approaches of assessing the predictive power of an estimated binary dependent-variable classification model, regardless of the modeling method used. One approach is the traditional two-by-two classification table, appropriate for small data settings like clinical experiments. The second approach - the decile table - has become for most modelers a generalized measure of model performance, for a either binary or continuous dependent variable. The decile table is widely used for today's big data. I outline how to construct both tables, and pose questions to raise awareness that each approach has its own weakness.

Two-by-Two Classification Table

The everyday workhorse classification method is the traditional statistical logistic regression model. The traditional statistical paradigm for building a binary dependent-variable classification model: The data analyst fits the data to the logistic regression model, whose equation is the sum of weighted predictor variables, which are declared statistically significant. The weights (better known as regression coefficients) are the main appeal of the statistical paradigm, as they provide the key to interpreting what the equation means. The two-by-two classification table is most appropriate for small data settings like clinical experiments. (Really?) Lest one forgets, every classification technique can yield the two-by-two table. The information needed to assess the *goodness* of a classification model exists within the 2x2 table, aka the *confusion matrix*, whose construction:

1. Construction of the 2x2 table of actual versus predicted outcomes - the confusion matrix. See Table 1, below.
2. Calculation of five standard terms based on the confusion matrix entries.
3. Little intelligence is exercised for the understanding, and inseeing interpretation of the quintuplet terms. The latter terms present the modifier *confusion* to the term matrix.
4. The criterion of a *good* model: The (1, 1) cell value is *large*. The larger the cell value, the more predictiveness of the classification model.

Table 1: Confusion Matrix Response Classification Model Results				
		Predicted		Total
		0	1	
Actual	1	a	**b**	a + b
	0	c	d	c + d
Total		a + c	b + d	a + b + c + d

Regarding the first step of construction, the matrix entries are imbued directly from the freshly built (estimated) classification model based a validation dataset. Note: The treatment of the matrix in this fashion can easily be extended to a polychotomous (aka multinomial) dependent variable. Also, of course, a + b + c + d = N (sample size).

The matrix entries based on the classification model:

- *a* is the number of **incorrect** predictions - actual Response = 1 and predicted Response = 0
- *b* is the number of **correct** predictions -   actual Response = 1 and predicted Response = 1
- *c* is the number of **correct** predictions -   actual Response = 0 and predicted Response = 0
- *d* is the number of **incorrect** predictions - actual Response = 0 and predicted Response = 1
- The classification model predicts - the Probability of Response (=1) equals **b** / **N**

The five standard terms, which are drawn from the confusion matrix, render the matrix *confusing*. I identify three essential terms and their meanings that are important for evaluation of the predictive power of a classification model. The two remaining terms are found when the confusion matrix is dressed up with the inferential statistics in Table 2, *Confusion Matrix with Signifiance Test Terms*, below.

- *Accuracy* (A) is the proportion of the total number of <u>correct</u> predictions:
$$A = (b + c) / (a + b + c + d)$$
- *Recall* (R) is the proportion of 1's that are correctly predicted <u>among</u> the actual 1s:
$$R = b / (a + b)$$
- *Precision* (P) is the proportion of the predicted 1s that are correct <u>among</u> the predicted 1s:
$$P = b / (b + d)$$

Table 2: Confusion Matrix with Signifiance Test Terms Response Classification Model Results				
		Predicted		
		Negative	Positive	
Actual	Positive	False Negative (Type II error)	True Positive	Sensitivity
	Negative	True Negative	False Positive (Type I error. P-value)	Specificity
		Negative predictive value	Positive predictive value	

Sensitivity = number of True Positives ÷ (number of True Positives + number of False Negative)

Specificity = number of True Negatives ÷ (number of True Negatives + number of False Positives)

Question #1

How are the Actual rows defined to yield the confusion matrix?
Hint #1: The goal is to maximize **b** in Table 1, or the **positive-positive** cell in the Table 2, below.
Hint #2: The *theoretical* assignment of the Actual rows virtually never yields satisfactory results.
Hint #3: The *practical* assignment of the Actual rows mostly yields satisfactory, yet *bias* results.

Answer #1 - Click here.

Decile Table

Historians trace the first use of the decile table, originally called a *gains chart* with roots in the *direct mail* business, circa wee 1950s. [1] The gains chart is hallmarked by solicitations found inside the covers of matchbooks. More recently, the decile table has transcended the origin of the gains chart toward a generalized measure of model performance. The term *decile* was first used by Galton in 1882. [2]

The decile table is a tabular display of model performance. It has become for most modelers a generalized measure of model performance, for a either binary or continuous dependent variable. The decile table is widely used for today's big data. I illustrate the construction and interpretation of the binary response (yes=1, no =0) decile table found in slides#4 and #5, below. The response model, on which the decile table is based, is not shown. Be mindful of the decile table indicates only the accuracy of a response model. Keep in mind: The eight-step construction detailed below for the binary dependent variable is identical (except for Prob_est (Probability_estimate) is replaced by Y_pred(iction)) for a continuous dependent variable Y, such as profit, sales, or write-offs.

Construction of the Response Decile Table

1. Score the validation sample or file using the response model under consideration. Every individual receives a model score, Prob_est, the model's estimated probability of response.

2. Rank the scored file, in descending order by Prob_est.

3. Divide the ranked and scored file into ten *equal* groups. The Decile variable is created, which takes on ten ordered 'values': top (1), 2, 3, 4, 5, 6, 7, 8, 9, and bottom (10). The 'top' decile consists of the best 10% of individuals most likely to respond; decile 2 consists of the next 10% of individuals most likely to respond. And so on, for the remaining deciles. Accordingly, Decile separates and orders the individuals on an ordinal scale ranging from most to least likely to respond.

4. Number of Individuals is the number of individuals in each decile; 10% of the total size of the file.

5. Number of Responses (actual) is the actual - not predicted - number of responses in each decile. The model identifies 865 actual responders in the top decile. In decile 2, the model identifies 382 actual responders. And so on, for the remaining deciles.

6. Decile Response Rate is the actual response rate for each decile group. It is Number of Responses divided by Number of Individuals for each decile group. For the top decile, the response rate is 18.7% (=865/4,617). For the second decile, the response rate is 8.3% (=382/4,617). And so forth, for the remaining deciles.

7. Cumulative Response Rate for a given depth-of-file (the aggregated or cumulative deciles) is the response rate among the individuals in the cumulative deciles. For example, the cumulative response rate for the top decile (10% depth-of-file) is 18.7% (=865/4,617). For the top two deciles (20% depth-of-file), the cumulative response rate is 13.5% = ([865+382]/[4,617+4,617]).
Et cetera, for the remaining deciles.

8. Cum Lift - for a given depth-of-file - is the Cumulative Response Rate divided by the overall response rate of the file (4.6%), multiplied by 100. It measures how much better one can expect to do with the model than without a model. For example, a Cum Lift of 411 for the top decile means that when soliciting to the top 10% of the file based on the model, one can expect 4.11 times the total number of responders found by randomly soliciting 10%-of-file. The Cum Lift of 296 for top two deciles means that when soliciting to 20% of the file based on the model, one can expect 2.96 times the total number of responders found by soliciting 20%-of-file without a model. And so and so, for the remaining deciles.

Rule: The larger the Cum Lift value the better the accuracy, for a given depth-of-file.

Reference

1 - The decile table has ten rows of equal number of individuals, irrespectively of model score. There can be individuals with the same model score in adjacent deciles. In a gains chart, there are as many rows as there are distinct model scores. Thus, there are no individuals with the same model score across gains-chart rows.

2 - Galton, F., "Report of the Anthropometric Committee," in *Report of the 51st Meeting of the British Association for the Advancement of Science*, 1882, pp. 245-260.

***Slide-show* Construction of a Response Decile Table**

Slide 1: What is the Decile Table?

What is the Decile Table?

RESPONSE Model Criterion

- How well the model **correctly classifies** response in the **upper** deciles.

- Perfect Response Model
  - among 100 individuals
    - 40 responders
    - 60 nonresponders.

Decile	Number of Individuals	Number of Responses
1	10	10
2	10	10
3	10	10
4	10	10
5	10	0
6	10	0
7	10	0
8	10	0
9	10	0
bottom	10	0
Total	100	40

Slide 2: Response Model Criterion

What is the Decile Table?

RESPONSE Model Goal

- We seek a model that identifies the **maximum** responses in the **upper** deciles.

Decile	Total Response
top	max
2	max
3	max
4	
5	
6	
7	
8	
9	
bottom	

Slide 3: Response Model Goal

Slide 4: Response Decile Analysis Top/(1) Decile Cum Lift

Slide 5: Response Decile Analysis Top-two/(1+2) Deciles Cum Lift

From step#3 in the construction of the Decile Table, the decile table is formed by "[dividing] the ranked and scored file into ten *equal* groups." This division of equal groups requiring no intelligence and effort, and therefore losing insight into the model performance is perhaps a *dumb* division. How can one turn the dumb decile table into a *smart* decile table?

Answer #2 - Click here.