Original Article

Assessing model performance: The Gini statistic and its standard error

Received (in revised form): 13th January 2010

Henry J. Greene

is an assistant professor of marketing at Central Connecticut State University. His specialty is direct marketing. Before his academic life, he worked in industry as a direct marketing analyst for Donnelley Marketing, Advanced Database Marketing, Strategic Mapping, ADVO, Inc, IBM and Information Xperts. He has also taught mathematics, computer science and statistics as a faculty member at several colleges and universities.

George R. Milne

is an associate professor of marketing at the University of Massachusetts Amherst. His primary research focus is in the area of information privacy and direct marketing. He has published over 45 journal articles appearing in various marketing journals.

ABSTRACT Database marketers frequently create statistical models to assess customers in terms of loyalty, lifetime value, responsiveness or acquisition. An assortment of modeling techniques, for example RFM (Recency-Frequency-Monetary value) models, regression, logit, neural networks and genetic algorithms, have been investigated to determine the most appropriate and effective technique. For the most part, traditional statistical measures utilize R^2, the *F* statistic, the Chi Square statistic, various classification indices and so forth to assess model performance – with an emphasis on goodness of fit, and measuring how closely data points fit a statistical model. Practitioners, on the other hand, typically use summarized descriptive methods to assess model performance: decile analysis, lift charts, cumulative lift charts, gains tables and cumulative gains tables. Both of these approaches have limitations. Some academic researchers have suggested that traditional goodness of fit statistics are not appropriate for evaluating model performance when the objective is to create models that maximize differentiation between population segments in terms of response rates. The traditional statistical measures are appropriate for assessing how well individual response values fit a given model (for example minimize least square errors between response data points and predicted values); however, they are not appropriate for effectively selecting market segments or individual customers for targeting and meeting business objectives. The descriptive measures used by practitioners, although visually appealing, do not assess overall model performance with statistical certainty. This research offers a remedy for the current situation by proposing the use of the Gini statistic and the associated standard error. We explain the Gini statistic and how it is connected to commonly used assessment measures. We then describe a simple method for computing Gini and its standard error. The accuracy of the method is demonstrated with specific industry data files.

Journal of Database Marketing & Customer Strategy Management (2010) **17,** 36–48. doi:10.1057/dbm.2010.2

Keywords: Gini; standard error; model assessment; model measurement; gains charts; lift charts

Correspondence:
Henry J. Greene
Department of Marketing,
1615 Stanley Street, Central Connecticut State University,
New Britain, Connecticut
06050, USA
E-mail: GreeneHej@ccsu.edu

INTRODUCTION

Database marketing continues to grow in importance as technology developments make it possible for even the smallest businesses to track customers and prospects efficiently.[1–4] In 2006, database marketing services grew by more than 14 per cent.[5] Since 2001, digital marketing and database marketing have continued to grow in importance as advertising strategies. Between 2001 and 2009, digital marketing (database marketing is a major component) represented 33 per cent of all advertising spending. In 2008 digital marketing represented 88 per cent of all advertising expenditures.[6]

An important task in database marketing is to create effective market segments or customer segments for the purpose of identifying appropriate targets for communication and advertising campaigns. Statistical models are created to determine lifetime value of customers, assess customer loyalty and value, and identify customers and prospects for acquisition and retention marketing programs. The business objective for creating database models is to increase advertising efficiency by targeting households or individuals that are the most likely to respond to specific advertising offers, thus reducing communication expenses while maintaining or increasing customer response or sales.

In the database world of model building, two popular analytic activities are model comparison and model validation. More specifically, database marketing analysts frequently construct multiple response models and need a measurement tool to compare and assess model performance. Once a candidate model is selected, it is prudent to validate the model with an independent data file and determine whether the model's performance is reliable (consistent) and sample independent. Database marketing acquisition models typically generate low response rates. Small incremental improvements in response rates can have a very significant impact on campaign profitability. Thus, it is important to create response models that maximize performance, assess performance with meaningful metrics and evaluate the significance of performance differences between competing models.

From a business perspective, a common objective is to create a response model that generates the highest return. The performance of the selected model should not only be superior to other models, but should also be significantly superior. Currently, the assessment methods employed by practitioners[7–11] for assessing model performance or model reliability are not statistically rigorous. The database analyst cannot assess significance in measured differences. Gains and lift charts are valuable for evaluating some aspects of marketing campaigns (for example response/profit comparisons of different decile segments); however, they are cumbersome when comparing overall model performance or for validating a model with different data files. It is possible to add statistical rigor to individual decile computations with bootstrapping or jackknifing sampling methods, yet this requires a significant programming effort.

This article provides remedies for these two situations through an explication of the merits of the Gini statistic, often used in social sciences, and more recently applied in direct marketing research.[2,4,12] Our contribution is twofold. First, we support the utilization of the Gini statistic as a performance measure for assessing database marketing response models. We show how the statistic is related to many of the popular descriptive methods used by practitioners. Second, we present a formula for approximating the standard error of the Gini statistic. This provides analysts with the ability to compare statistical differences between competing models and to validate response models with different data files. Our methodology is derived from

a regression analysis of over 1000 different data conditions created from a Monte Carlo simulation in which we varied the value of the Gini coefficient, sample file size and sample response rate. It is important to note that the simulation was created completely independent of any specific type of underlying response model (regression, Chaid, Neural Networks and so on). Our assessment for the standard error of Gini depends only on file size, response rate and the actual Gini obtained, regardless of the model that created the Gini. Essentially, we assume a very liberal variation in response (the simulation utilizes a uniform distribution of responses within deciles), which should serve as a conservative estimate.

The methodology that we introduce for computing the standard error of Gini is most appropriate when response models are applied to relatively large data files ($n > 15\,000$). Files of this size or larger are commonly used in database marketing applications. In order to employ a response model assessment in a database marketing context, an analyst might follow a process such as scoring all data records with the model; sorting all records on the file by model score (predicted values); classifying each record with an appropriate n-tile (typically deciles are used – 10 equal size segments); aggregating or averaging the measure of interest (for example responses, inquiries, sales) for each decile; and displaying the results in tabular or graphical form. The Gini statistic is a single number that represents the area under the cumulative lift chart relative to the area under a uniform distribution. The value's association with the cumulative lift represents the cumulative percentage of responses.

It is not uncommon for database marketers to utilize two other statistics, Area under the Curve (AUC) and receiver operating characteristic (ROC). AUC is equivalent to the Gini differing by a scale factor. A benefit of Gini is that it is scaled from 0–1, where 0 indicates no difference between segments and 1 indicates a maximum difference. ROC curves are designed to assess two competing conditions simultaneously rather than assess one. Situations such as signal/noise for interpreting radar, illness and extraneous symptoms for diagnosing patients benefit from ROC analysis. One can use ROC to balance two types of potential errors (False Positives and True Negatives). However, when there is one dominant objective (Maximize Response Rate), a measure like Gini is simpler and more useful. The Gini statistic has been recognized in the literature by different marketers [2,3,9] and is used by practitioners.

Our approach to deriving the standard error of Gini differs from other methods discussed in the literature. When the data points represent individual data points and are sorted according to their unique values, Giles[13] describes a regression approach. There is still some controversy about the approach.[14–16] Moreover, the database marketing summaries do not possess the properties appropriate for computing the standard error of Gini using Giles method.[13]

The remaining portion of this article is organized as follows. We discuss response model evaluation, summarize the descriptive methods used by practitioners, describe how descriptive methods are related to the Gini statistic, describe a method for approximating Gini, present a procedure for estimating the standard error of Gini and, lastly, illustrate our suggested procedure using three data sets, two available from the Direct Marketing Association (DMA) and the third a subset of a proprietary data file from a large, national insurance company.

BACKGROUND

Database model evaluation

Traditionally, statistical response models used in database marketing have been

evaluated based on some form of goodness of fit. Assumptions are made regarding underlying data distributions and models are evaluated based on how well predicted data values from the response model actually fit the observed data values from a sample data set. Various statistical measures (R^2, the F statistic, the Chi Square statistic, classification indices and so on) are used to evaluate the goodness of fit. In database marketing, the goals are more focused on developing response models to meet business objectives.[17,18] Although fit is often useful, the goal is differentiating consumer units (for example people, households) based on their likelihood of responding to a specific offer. As such, marketers create metrics that measure the magnitude of separation between productive market segments and non-productive market segments.

Practitioners engaged in direct marketing efforts have been using alternative metrics for evaluating model performance for at least 25 years.[7,9] The metrics that are most commonly used are the decile chart,[7,8,11] the gains and cumulative gains table,[8] the lift chart and the cumulative lift chart. [1,6,8,11] Decile analysis is used in database marketing in order to more easily visualize data files consisting of thousands to hundreds of thousands of data records. For each decile segment, the number and percentage of responses are recorded. Gains indices are commonly created. A common form of the index is created by forming the ratio of segment to the average response rate of the total sample, multiplied by 100. This provides the analyst a method for determining which segments perform significantly better than average. A segment with a gains index of 120 indicates a segment that performs 20 per cent higher than average. The segments and the corresponding households that fall in high-performing segments are the best candidates for targeting. The graph shown in Figure 1, Panel A illustrates both a gains chart and a cumulative gains chart. The index (y axis value) represents the gains index, the ratio of a response rate of a segment to the overall response rate multiplied by 100.

Although the gains index is quite useful for identifying better performing segments, a difficulty arises when one tries to compare the performance of two models or to validate the performance of a model with two different sample files (see Figure 1, Panel B). In order to determine whether two models perform similarly, practitioners rely on 'eyeball' judgments or prior experience. There is no formal or rigorous procedure to claim that one of the models is superior to the other in terms of performance. In addition, if the two graphs represent one model applied to two different sample files, there is no statistical test to conclude that the model is reliable, consistent or sample independent. Another issue that frequently occurs when response models are built with small samples or low response rates is over-fitting. Providing a Gini statistic alone does not warn the analyst of potential model failure owing to an over-fit (sample specific) condition. By supplying the standard error for Gini, the analyst is made aware of potential risk that can be realized. This is a very important point. Research papers that describe model performance with Gini should always include the standard error[13]

Descriptive metrics used by practitioners

A brief survey among practitioners (13 members of the DMA Research Council) was conducted in 2004 to determine their preferred method of choice for evaluating response models. The gains and cumulative gains chart were the most popular metrics for assessing model performance, followed by measures of lift and cumulative lift. The applicability of these metrics is also supported in the

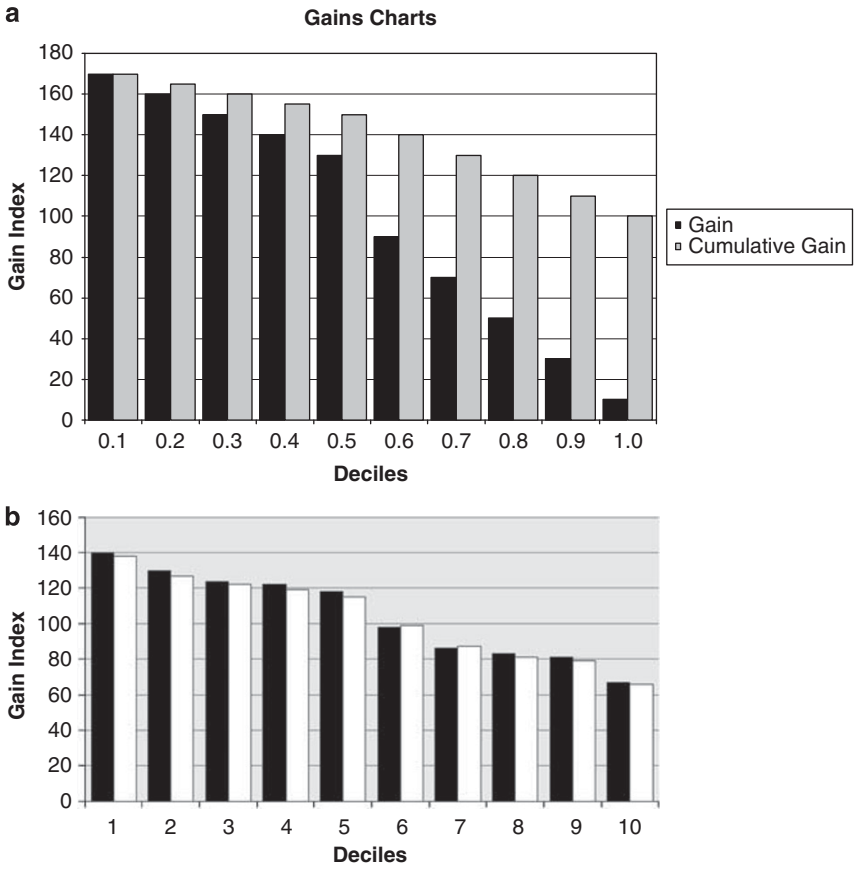

Figure 1: Gains chart, cumulative gains chart and comparison. **(a)** Gains chart and cumulative gains chart. **(b)** Comparison of two gains charts

Table 1: Descriptive metrics used by industry practitioners

(1) Decile	*(2) Responses*	*(3) Customers*	*(4) Response rate*	*(5) Gains index*	*(6) Cumulative response rate*	*(7) Cumulative gains index*	*(8) Lift*	*(9) Cumulative responses*	*(10) Cumulative lift*
1	220	10000	0.0220	168.3	0.0220	168.3	0.168	220	0.168
2	213	10000	0.0213	163.0	0.0217	165.6	0.163	433	0.331
3	197	10000	0.0197	150.7	0.0210	160.7	0.151	630	0.482
4	185	10000	0.0185	141.5	0.0204	155.9	0.142	815	0.624
5	142	10000	0.0142	108.6	0.0191	146.4	0.109	957	0.732
6	97	10000	0.0097	74.2	0.0176	134.4	0.074	1054	0.806
7	83	10000	0.0083	63.5	0.0162	124.3	0.064	1137	0.870
8	71	10000	0.0071	54.3	0.0151	115.5	0.054	1208	0.924
9	57	10000	0.0057	46.6	0.0141	107.5	0.044	1265	0.968
10	42	10000	0.0042	32.1	0.0131	100	0.032	1307	1.000
Total	1307	100000	0.01307	—	GINI=	0.281	—	—	—

literature.[1,8,19,20] The descriptive metrics used by practitioners are illustrated in Table 1.

Column 1 shows 10 segments used in decile analysis. Each decile has an equal number of customers, 10000 in this case (column 3). The responses (column 2) divided by the customers (column 3) is used to determine the response rate

(column 4). The gains index (column 5) is derived by dividing the response (column 4) by the average response rate (0.01307 in this example) and multiplying by 100. The cumulative gains index is computed by first determining the cumulated response rate (column 6), which for decile *i* is the sum of all responses from decile 1 to decile *i*, divided by all customers from decile 1 to decile *i*. Once the cumulative response rate is computed, the cumulative gains index (column 7) is the cumulative response rate (column 7) divided by the overall response rate (0.01307) multiplied by 100. Lift (column 8) is the number of responses in a decile (column 2) divided by total responses. Cumulative lift (column 10) is computed by accumulating responses from all prior deciles (column 9) inclusively and dividing the accumulation of responses by total number of responses (for example 1307).

THE RELATIONSHIP AMONG THE GAIN, LIFT AND THE MODIFIED GINI

The original Gini coefficient

Although there are many derivations of the Gini coefficient, we start with the original coefficient as articulated by Corado Gini in 1921, which is the form utilized in the direct marketing community. The Original Gini coefficient[19] is computed by sorting scores from a distribution from low to high, determining the corresponding cumulative lift (Lorenz curve), computing the area between the cumulative uniform distribution (A^U) and the Lorenz curve(A^L) on the interval [0, 1], and dividing the result by the area under the cumulative uniform distribution on the same interval.

$$A^U = 0.5.$$

$$Gini = \Gamma = (A^U - A^L)/A^U = 1 - 2A^L$$

The modified Gini coefficient

The modified Gini coefficient[21] (Γ) measures the relative difference between two areas, that is, the area under a Lorenz curve (cumulative percentage of responses) or cumulative lift curve (A^L), and the area under a cumulative uniform distribution (A^U), relative to the area under the cumulative uniform distribution. The Gini we discuss is very similar to the original Gini used by economists, except in the case of the original Gini scores are ranked from low to high and in our modified form scores are ranked from high to low. In either case, Gini represents the area between the same two curves. The area between the two curves is computed in such a way that it is always positive. Otherwise, the calculation for our modified Gini is the same as the original Gini, the relative area between two curves, divided by 0.5. This allows Gini to range from 0 to 1.

$$0 \leq Gini(\Gamma) \leq 1 \qquad \text{on the interval } [0,1]$$

$$\Gamma = \frac{A^L - A^U}{A^U} = \frac{A^L - 0.5}{0.5} = 2A^L - 1$$

Modified Gini

And

$$\Gamma = 2\int_0^1 F(x)dx - 1 \qquad \text{where } \int_0^1 F(x) = A^L$$

The Gini coefficient is minimized when the responses are spread equally across all deciles (Gini = 0) and maximized when all of the responses are in the top decile (Gini = 1). Theoretically, Gini could be negative, but no practitioner or researcher would either use or continue to investigate the results of a model that performed worse than random chance. The faster the cumulative lift reaches 1 (see Figure 2), the greater the AUC will be and the greater the Gini coefficient will be.

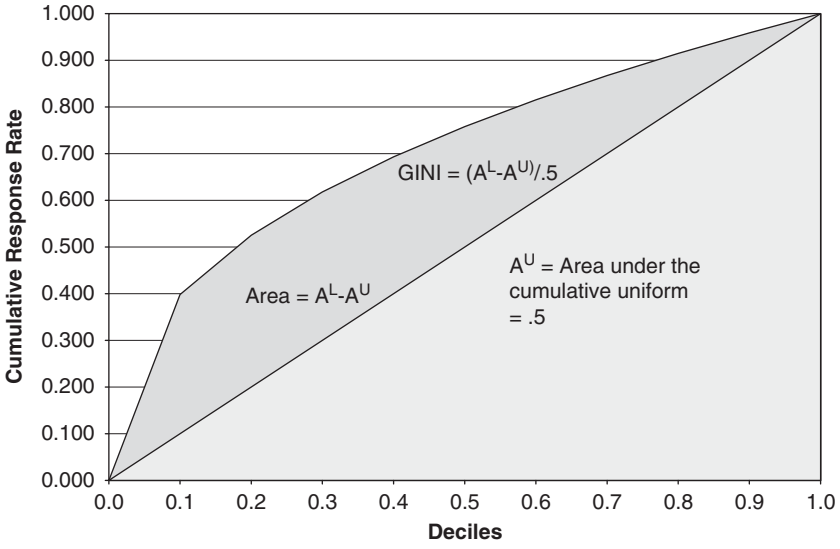

Figure 2: Gini graph.

COMPUTING THE GINI STATISTIC

In practice, population response distributions are unknown and are estimated from sampled data. The Gini statistic is a function of the area under the Lorenz curve (area under the cumulative lift).

$$\Gamma = 2A^L - 1$$

Trapezoids are frequently used to approximate the area between the empirical response curve and the cumulative uniform distribution. Using trapezoids derived from deciles ($n = 10$), and noting that $F_{0=0}$

$$\Gamma = 2(1/2\ \Delta x \sum_{i=0}^{n-1}(F_i + F_{i+1})) - 1$$

$$= 2(\Delta x[\sum_{i=1}^{n} F_i - 0.5]) - 1$$

The Gini coefficient is computed directly from the cumulative lift (A^L), which is a linear transformation of the cumulative gains index (A^G), one of the descriptive techniques frequently used in database model building. Every cumulative gains index corresponds to a unique cumulative lift index,

$$A_i^G = 1000\frac{A_i^L}{i},$$

where i represents the decile segment of interest.

Thus, with substitution, the Gini coefficient can also be expressed in terms of cumulative gains indices.

$$\Gamma = \frac{2}{1000}\int_0^1 iA_i^G - 1$$

Whereas the cumulative gains graph and cumulative lift graph provide a visual display of response model performance, the Gini statistic provides a single number, which explains the degree of separation between segments in terms of response rate.

Quick calculation for Gini

The Gini coefficient can be calculated using the trapezoid approximation for computing the area under a continuous function. Gini is calculated by summing

Table 2: Relationship between Gini coefficient and trapezoid estimation of Gini coefficient

Gini coefficient (Γ)	*Trapezoid Gini (Γ_T)*	*Ratio=Γ/Γ_T*
Panel A – Data		
0.100	0.098	1.020
0.200	0.194	1.031
0.300	0.289	1.038
0.400	0.383	1.044
0.500	0.475	1.053
0.600	0.564	1.064
0.700	0.652	1.074
0.800	0.737	1.085
0.900	0.818	1.100

Variable	*Beta*	*T-value*	*Significance*
Panel B – Regression adjustment for trapezoids			
Intercept	0.00433	2.23	0.0674
Γ_T	0.97122	101.32	<0.0001
Γ_T^2	0.1413	13.8	<0.0001
Adjusted R^2=0.99	—	—	—

To adjust trapezoid Gini coefficients:
$\Gamma=0.00433+0.97122\ \Gamma_T+0.1413\ \Gamma_T^2$.

the cumulative lift values, subtracting 0.5, multiplying by 0.2 (assuming deciles) and subtracting 1. As such, the more segments used in the summary chart, the more accurate the approximation to the true area. The Gini coefficient from the data in Table 1 is 0.281.

In order to better approximate the AUC we offer the following correction. This is based on simulated data in Table 2. To recalibrate Gini, first calculate the trapezoid Gini and then use the following formula (see Table 2).

$$\Gamma = 0.00433 + 0.97122\Gamma_T + 0.1413\Gamma_T^{\ 2}$$

where Γ_T is Gini statistic calculated with trapezoids; and Γ is Gini calculated with a continuous function. Thus, Gini is recalibrated to 0.288, based on a trapezoid Gini of 0.281.

COMPUTING THE STANDARD ERROR FOR GINI

Knowing the standard error of Gini is useful for assessing the reliability of a model and determining the superiority of one model over another. First, we describe how we determined the standard error of Gini.

Monte Carlo simulation

In order to estimate the standard error of Gini we ran a Monte Carlo simulation. We created data files of 100 000 records and each record was given a 0 or 1 to reflect response = 1 or non-response = 0. For a given Gini and response rate, we selected the appropriate proportion of 0s and 1s to match the given Gini and response rate. We then created 200 sample files, ranging between 5000 and 50 000 records to compute the standard error associated with a variety of Gini values, response rates and sample file sizes. We tested a total of 1620 conditions, varying response rate from 0.01 to 0.1 in increments of 0.01, and then from 0.1 to 0.9 in increments of 0.1; and file sizes were varied from 5000 to 50 000 records in increments of 5000 for a total of $18\times9\times10=1620$ test cases. For each test case, we extracted 200 random samples. We selected a variety of conditions that we felt would simulate actual database marketing situations as well as some extreme cases. We recognize that database analysts frequently have hundreds of thousands and even millions of records available for analysis. However, as they fine-tune their campaigns, they may test tens or hundreds of sub-file conditions, reducing their analysis to files ranging from thousands to tens of thousands of records. Details on the simulation implementation are shown in the Appendix.

For each of the 1620 test conditions, we computed the standard errors of the Gini statistic. Standard errors increase when the file structures are more unstable. As shown in Figure 3, standard errors increase as file sizes, response rates and Gini coefficients decrease. For example, if the response rate is low, there are fewer 1s in the file. Small

N=10.000

Gini	Response Rate 0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1	*	0.041	0.038	0.036	0.034	0.033	0.032	0.031	0.030
0.2	*	0.038	0.036	0.034	0.032	0.031	0.030	0.029	0.028
0.3	*	0.036	0.033	0.031	0.030	0.029	0.028	0.027	0.026
0.4	*	0.034	0.031	0.029	0.028	0.026	0.025	0.024	0.024
0.5	*	0.032	0.029	0.027	0.025	0.024	0.023	0.022	0.022
0.6	*	0.029	0.026	0.025	0.023	0.022	0.021	0.020	0.019
0.7	*	0.027	0.024	0.022	0.021	0.02	0.019	0.018	0.017
0.8	*	0.025	0.022	0.020	0.019	0.017	0.016	0.016	0.015
0.9	*	0.023	0.020	0.018	0.016	0.015	0.014	0.013	0.013

N=10.000

Gini	Response Rate 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	*	*	*	*	*	*	*	*	*
0.2	0.027	0.023	0.021	0.020	0.020	0.019	0.019	0.019	0.019
0.3	0.025	0.021	0.019	0.018	0.018	0.018	0.018	0.018	0.018
0.4	0.023	0.019	0.017	0.017	0.016	0.016	0.016	0.016	0.017
0.5	0.021	0.017	0.016	0.015	0.014	0.015	0.015	0.015	0.016
0.6	0.019	0.015	0.014	0.013	0.013	0.013	0.013	0.014	0.014
0.7	0.017	0.013	0.012	0.011	0.011	0.011	0.012	0.012	0.013
0.8	0.014	0.011	0.010	0.009	0.009	0.010	0.010	0.011	0.012
0.9	0.012	0.009	0.008	0.008	0.008	0.008	0.009	0.010	0.011

N=25.000

Gini	Response Rate 0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1	0.033	0.028	0.025	0.023	0.022	0.021	0.020	0.019	0.019
0.2	0.031	0.026	0.023	0.021	0.020	0.019	0.018	0.017	0.016
0.3	0.028	0.024	0.021	0.019	0.018	0.016	0.016	0.015	0.014
0.4	0.026	0.021	0.019	0.017	0.015	0.014	0.013	0.013	0.012
0.5	0.024	0.019	0.016	0.014	0.013	0.012	0.011	0.010	0.010
0.6	0.022	0.017	0.014	0.012	0.011	0.010	0.009	0.008	0.008
0.7	0.019	0.015	0.012	0.010	0.009	0.008	0.007	0.006	0.005
0.8	0.017	0.012	0.010	0.008	0.006	0.005	0.004	0.004	0.003
0.9	0.015	0.010	0.007	0.005	0.004	0.003	0.002	0.002	0.001

N=25.000

Gini	Response Rate 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.018	0.015	0.014	0.014	0.014	0.015	0.016	0.017	0.018
0.2	0.016	0.013	0.012	0.012	0.012	0.013	0.014	0.015	0.017
0.3	0.014	0.011	0.010	0.010	0.011	0.012	0.013	0.014	0.015
0.4	0.011	0.009	0.008	0.008	0.009	0.010	0.011	0.013	0.014
0.5	0.009	0.007	0.006	0.007	0.007	0.009	0.010	0.011	0.013
0.6	0.007	0.005	0.004	0.005	0.006	0.007	0.008	0.010	0.012
0.7	0.005	0.003	0.002	0.003	0.004	0.005	0.007	0.009	0.011
0.8	0.003	0.000	0.000	0.001	0.002	0.004	0.006	0.007	0.009
0.9	0.000	0.000	0.000	0.000	0.000	0.002	0.004	0.006	0.008

N=40.000

Gini	Response Rate 0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1	0.020	0.016	0.013	0.011	0.010	0.009	0.008	0.007	0.007
0.2	0.018	0.013	0.011	0.009	0.008	0.007	0.006	0.005	0.005
0.3	0.016	0.011	0.008	0.007	0.005	0.004	0.004	0.003	0.002
0.4	0.014	0.009	0.006	0.004	0.003	0.002	0.001	0.000	0.000
0.5	0.011	0.007	0.004	0.002	0.000	0.000	0.000	0.000	0.000
0.6	0.009	0.004	0.002	0.000	0.000	0.000	0.000	0.000	0.000
0.7	0.007	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.8	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.9	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

N=40.000

Gini	Response Rate 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.006	0.004	0.004	0.006	0.007	0.009	0.011	0.013	0.015
0.2	0.004	0.002	0.003	0.004	0.005	0.007	0.009	0.012	0.014
0.3	0.002	0.000	0.000	0.002	0.004	0.006	0.008	0.010	0.013
0.4	0.000	0.000	0.000	0.000	0.002	0.004	0.006	0.009	0.012
0.5	0.000	0.000	0.000	0.000	0.000	0.003	0.005	0.008	0.010
0.6	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.006	0.009
0.7	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.005	0.008
0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.007
0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.006

* outside the boundary condition

Figure 3: Boundary conditions for Gini coefficient standard errror.

changes in the number of 1s in any segment cause a high degree of variability in the sample calculation of Gini. The figure shows six panels representing various conditions. The three panels on the left reflect response rates from 0.01–0.09 and the three columns on the right reflect response rates from 0.1–0.9. The first panel row represents an n of 10 000, the second row an n of 25 000 and the third row an n of 40 000. The cells with the darker coloring reflect larger standard errors. The figure shows that higher standard errors are found when (1) Gini is small, (2) response rate is low (<0.02), (3) Gini is low or (4) the file size is small. When models are constructed with file sizes (n) less than 30 000, the standard errors of Gini should be included in any report. One should also consider including the standard error when response rates are lower than 0.04 and file sizes are $<50\,000$.

Normality distribution of sample Ginis was checked in 95 of the test cases representing the range of combinations of n – file size, r – response rate and Gini. Of the 95 cases tested, 78 could not be rejected for normality, $P=0.05$, based on the Anderson–Darling and Kolmogorv–Smirnov tests.

Table 3: Gini standard error formula estimate

Variables	*Betas*	*T-value*	*Significance*	*VIF*
Intercept	0.02258000	30.73	<0.0001	0
[n+log(n)]	−0.00000085	−54.86	<0.0001	1.829
Γ	−0.02299000	−26.63	<0.0001	1.807
Log (r)	−0.01692000	−56.5	<0.0001	5.314
[r×n]	0.00000075	22.53	<0.0001	5.517
[r×Γ]	0.01227000	6.59	<0.0001	5.634
Adjusted R^2=0.90	—	—	—	—

Regression formula for Gini coefficient standard error (GSE): SE_Γ=0.02258−0.00000085×[n+log(n)]−0.02299 ×Γ−0.01692×[Log (r)]+0.00000075×[r×n]+0.01227×[r×Γ] or 0 if SE_Γ<0, where n=number observation to calculate Gini coefficient; Γ=Gini coefficient; r=response rate of data file.

Those cases in which normality could not be assumed fell on the boundaries of the test cases. In these cases Gini was extremely small (<0.2), the file size was small (<10 000), the response rate was small (<=0.01) or a combination of all three factors. We observed that if Gini is computed with a file size of at least 20 000 records, a response rate of at least 1 per cent and a Gini >0.2, it will be reasonable to assume that the sample Gini is normally distributed. This will provide an opportunity to make statistical inferences about Gini.

Estimating the standard error of Gini with Ordinary Least Squares

A regression equation was created to predict standard errors for the Gini statistic as a function of sample size, response rate and the Gini coefficient. The data set for building the regression equation was based on 1136 different parametric conditions. We eliminated the boundary conditions that are unrealistic for customer acquisition.

The regression formula that was used to estimate the standard error of the Gini coefficient is shown in Table 3. To validate the formula, we compared the predicted values to actual sample values. In absolute terms, 98.6 per cent of the standard error estimates are within 0.01 of the sample Gini. Confidence intervals for the Gini coefficient can be calculated as $\Gamma \pm Z\sigma_\Gamma$, where σ_Γ = standard error of the Gini coefficient and will be approximated with the regression equation shown in Table 3.

APPLICATION OF GINI AND ITS STANDARD ERROR

In this section we demonstrate the accuracy of the Gini standard error using a model calculated on an entire data set. First, we calculate Gini for the entire data set. We took 30 random samples from the file and computed an estimate of the standard error of Gini. We calculated one standard error estimate from our formula and the other standard error from the sample of 30 Ginis. We compare the differences of the standard errors. We did both of these calculations on two different data sets. We selected data sets provided by the DMA (catalog customer file and non-profit contributor file). We then show how including a standard error can assist analysts with model assessment, reliability assessment and model selection. These data files were selected because they are similar to the types of files that database marketers routinely use, and they have also been used in various academic research studies.

Validation

We test our methodology with three different data files, two from the DMA and one representing a national insurance company. Regression models are built to predict response, and we then compute

Table 4: Standard error validation results

Data set	*Sample creation parameters*	*Gini from model*	*Gini sample mean across 30 samples*	*Gini sample standard error*	*Gini regression estimated standard error*
DMA catalog	*N*=15 000 *r*=0.15	0.400	0.405	0.016	0.017
	N=30 000 *r*=0.15	0.400	0.403	0.009	0.006
DMA non-profit	*N*=5400 *r*=0.17	0.384	0.391	0.018	0.024
	N=15 200 *r*=0.17	0.384	0.382	0.007	0.017

Table 5: Gini, standard errors and confidence intervals

Data sets	*Gini, standard error*
Auto insurance *N*=8611, *r*=0.059	
Analytic (training)	0.332 (0.0290)
Validation (testing)	0.302 (0.0297)
Confidence interval	0.332 +/−0.056
DMA catalog *N*=25 398, *r*=0.11	
Analytic (training)	0.497 (0.0085)
Validation (testing)	0.537 (0.0076)
Confidence interval	0.497 +/−0.0166
DMA non profit *N*=39 938 *r*=0.27	
Analytic (training)	0.508 (0.0000)
Validation (testing)	0.517 (0.0000)
Confidence interval	0.508 +/−0.000

the Gini statistic as outlined in this article and the standard error of the Gini statistic with the regression equation that we described. Finally, we compare our computed value with the sample standard error generated from 30 random samples (each sample was randomly selected from the original data file). The results are presented in Table 4.

Using the DMA data sets, Table 4 shows the validation results that compare standard errors generated from the sample Gini statistics to the standard error estimated by the regression formula. Comparisons are made for the catalog and non-profit data sets. The results show that the estimated standard error is very close to the standard error computed from sampling 30 Ginis. These sample files were relatively small. For larger files, the standard error becomes very small, that is, Gini becomes quite stable. Table 5 demonstrates confidence intervals for the two DMA data files and the insurance data file. The size of the data files varies from 8600 to 39 000. As the sample sizes increase the confidence intervals shrink dramatically.

CONCLUSIONS

In this article we discussed why the Gini index is a useful measure for assessing model performance. Gini can also be used to measure the consistency (reliability) of a response model. In order to add statistical rigor to model assessment, we include a method for approximating the statistical error for the Gini statistic. Our methodology is most relevant when data files range in size from 20 000 to 60 000 records. When files are small ($n < 20\,000$), Gini becomes unreliable; when files are larger than 60 000 records, Gini becomes very stable and the standard errors will be very small and insignificant.

We recommend that if a split half validation is to be performed for assessing model reliability or if there is a need to compare two different models with the same data file, the Gini statistic be used as the performance measure of choice. It is easy to compute and easy to apply. The range of file sizes is easy to understand, and one can construct confidence intervals and hypotheses tests. This gives the Gini statistic a decided advantage over current descriptive techniques employed by both researchers and practitioners. Inclusion of the Gini standard error allows analysts the opportunity to determine whether model performance is statistically significant.

Limitations

Applying the Gini statistic and the appropriate standard error is limited to data files that have reasonable response rates, file sizes and Gini values. When Gini is less than 0.2, one must question whether utilizing a model will provide any significant benefit. When file sizes are small ($<20\,000$), model results will be unstable if the response rate is also small. Marketing efforts designed to acquire new customers frequently experience very low response rates. Under these circumstances, one would need to greatly increase the training file size to gain any confidence in the accuracy of the response model. The results of this article are based on a large simulation. For future research it would be interesting to develop a closed-form solution for the standard error of Gini.

REFERENCES AND NOTES

1 Magliozzi, T.L. and Berger, P. (1993) List segmentation strategies in direct marketing. *OMEGA, The International Journal of Management Science* 21: 71.
2 Greene, H. and Milne, G.R. (2005) Alternative data sources in targeted marketing: The value of exographics. *Journal of Targeting* 14(1): 33–46.
3 Levin, N. and Zahavi, J. (2005) Data Mining for Target Marketing. Working Paper, Chapter 1, Target Marketing.
4 Malthouse, E. (2002) Performance based variable selection for scoring models. *Journal of Interactive Marketing* 6(4): 37–50.
5 Channel Web http://www.crn.com/software/199202647.
6 SFN Blocg.com http://www.sfnblog.com/financials/2009/01/digital_takes_up_most_revenue_growth_in.php.
7 Hughs, A. (1996) *The Complete Database Marketer*. New York: McGraw Hill.
8 David Shepard Associates. (1999) *The New Direct Marketing*, 3rd edn. New York: McGraw Hill.
9 Nash, E. (2000) *Direct Marketing*. New York: McGraw Hill.
10 Nash, E. (1995) *Database Marketing*. New York: McGraw Hill.
11 Roberts, M.L. and Berger, P. (1999) *Direct Marketing Management*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.
12 Mulhern, F. (1999) Customer profitability analysis: Measurement, concentration, and research directions. *Journal of Interactive Marketing* 13(10): 25–40.
13 Giles, D. (2002) Calculating a Standard Error for the Gini Coefficient: Some Further Results. Department of Economics, University of Victoria. Working Paper EW PO202.
14 Ogwang, T. (2000) A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics* 26: 123–129.
15 Shlomo, Y. (1991) Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics* 9(2): 235–239.
16 Ogwang, T. (2004) Calculating a standard error for the Gini coefficient: Some further results: Reply. *Oxford Bulletin of Economics and Statistics* 66(3): 435.
17 Ratner, B. (2000) *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*. Boca Raton, FL: Chapman and Hall CRC Press.
18 Kyoungnam, H., Cho, S. and MacLachjan, D. (2005) Response models based on bagging neural networks. *Journal of Interactive Marketing* 19(1): 17–30.
19 Gini, C. (1921) Measurement of inequality and incomes. *The Economic Journal* 31: 124–126.
20 Malthouse, E. (2001) Assessing the performance of direct marketing scoring models. *Journal of Interactive Marketing* 15(1): 49–62.

21 This is a slight modification of the original definition of the Gini coefficient, which typically ranks observations from low to high and the Gini coefficient is computed as $\Gamma=1-2\int_0^1 F(x)dx$.

APPENDIX

Monte Carlo simulation for generating gains charts

In order to generate random responses a power function was utilized in the simulation. First, note that power functions of the form $Y=X^p$ have the same properties as Lorenz curves.

(a) when $X=0$, $Y=0$
(b) when $X=1$, $Y=1$
(c) $Y'=pX^{p-1}$ is always >0 when $P>0$

implying that Y is a monotonically increasing function

Therefore, we used power curves to simulate Lorenz curves.

1. The Gini for a Lorenz curve can be expressed as

$$\Gamma=\frac{\left(\int_0^1 x^{p-1}\right)/1}{2}=\frac{p-1}{p+1}$$

$$p=\frac{\Gamma-1}{\Gamma+1}$$

2. Step 1 in the simulation was to create a master file of 100 000 records; each record had two variables: decile number (an integer from 1 to 10 representing the decile) and response value (either a 0 or a 1).
3. 10 per cent of the records were given decile code 1, 10 per cent decile code 2 and so on.
4. For a given Gini value (ranging from 0.1 to 0.9), the corresponding p was computed, and then a set of Ys was computed, where each Y is computed from $Y = X^p$ and X was varied from 0.1 to 1.0 in increments of 0.1. The Xs represent the decile breaks 1–10. The Ys represent cumulative percentage of responses, the difference between two successive Ys (use ys) – represent percentage of responses for each decile.

$$\text{For example} \quad y_3 = Y_4 - Y_3$$

5. Once the percentage of responses is known for a decile, the number of records in a decile that are given $r = 1$ for response is = response rate $\times y_i \times 10\,000$. The responses in the decile are randomly assigned according to a uniform distribution.
6. The 100 000 record master file represents a specific Gini, a specific response rate and randomly assigned records by decile, where each decile is populated with responses and non-responses according to random assignment from a uniform distribution.
7. For the simulation, a sample size n is chosen. If $n = 20\,000$ records, then 20 000 records are randomly selected (uniform distribution) from the master file and the Gini of the sample is computed. For each set of N (file size), G (Gini value) and r (response rate) 200 samples were selected. From the 200 samples, the Gini and the standard error were computed; the results are displayed in Figure 3.