Measuring Lift Quality in Database Marketing

Gregory Piatetsky-Shapiro
Xchange Inc.
One Lincoln Plaza, 89 South Street
Boston, MA 02111
gps@xchange.com

Sam Steingold
Xchange Inc.
One Lincoln Plaza, 89 South Street
Boston, MA 02111
sds@xchange.com

ABSTRACT

Database marketers often select predictive models based on the lift in the top 5%, 10%, or 20%. However, different models may be better at different thresholds. Absent a good cost function, or when multiple cost functions are possible, we want a measure that helps to compare models by looking at the entire lift curve. In this paper, we propose such a measure, called L-quality, which indicates how close the lift curve is to the random and perfect models. We also examine the practical issues of computing L-quality from discrete quantiles available in a typical lift table.

Keywords

Database marketing, lift curves, ROC, measurement

1. MOTIVATION

Frequently, there is a need to compare the quality of different models in the context of database marketing. The traditional machine learning measure is an average model error rate applied to all cases. This measure is not directly applicable in the context of database marketing because we may not need to contact all prospects. Marketers usually contact only the top 10% or 20% of the prospects with the highest score, so the accuracy of prediction on the remaining 80% is not important.

Furthermore, the costs of errors are definitely not equal in database marketing. While a false positive may only cost a stamp or a phone call, a false negative may cost losing a customer (real cost) or losing a sale (lost profits).

Ultimately, the choice of the right cutoff should be made with economic considerations in mind, as we argued in KDD-99 paper [4]. However, there are cases when economic values are not (yet) available or are changing and we may want to compare just the model lifts.

Lift is a typical measure used in database marketing. Let $\mathrm{TargetsPercent}(M,p)$ be the percent of targets in the first $p\%$ of the list sorted by decreasing score of model M. Then we define

$$\mathrm{Lift}(M,p) = \frac{\mathrm{TargetsPercent}(M,p)}{p} \tag{1}$$

Typically, lift is examined at 5%, or 10% or 20% of the list. However, the choice of the right threshold is usually done ad-hoc. Instead of choosing an arbitrary threshold, we would like a comprehensive measure that would apply to the entire lift curve. Furthermore, we can show that if model A lift curve dominates (is strictly above) model B lift curve, then for any fixed cost/benefit matrix and for any p, the benefit of selecting the top p observations using the model A will be higher than the benefit of selecting the top p observations using B.

We are looking for a lift quality measure that will capture this idea of overall domination of one model over another.

2. MEASURES OF LIFT QUALITY

We propose to use a measure of overall lift quality, which we call L-quality. This measure is closely related to the ideas of ROC curve [3] and a similar measure, called *gains*, was used by Ismail Parsa in 1997 in evaluating KDD-Cup winners [2]. Some software tools like KXEN [1] also use this measure.

This paper proposes to establish a common terminology and formally defines and analyzes this measure and how to estimate it from the quantized lift tables.

The desired measure properties for the ideal random model R and the optimal (best) model B are

$$\begin{aligned} \text{L-quality}(B) &= 100\% \\ \text{L-quality}(R) &= 0\% \end{aligned}$$

If lift curve of model M_1 is strictly above the lift curve of model M_2, then

$$\text{L-quality}(M_1) > \text{L-quality}(M_2)$$

Table 1 is a lift table for a predictive model. In this case, a model was applied to 20,900 records of which 1,312 were targets, Model performance was measured in increments of 5%, which is typical. The percentage of targets in the population will be called the base rate and denoted as b. In this example, b=6.28%.

In this table, a row for pth percentile describes the properties of the first p of the list, scored by decreasing model score. The columns are:

- **recs**: the number of records;
- **hits**: the number of targets correctly identified;
- **hits%**: the percentage of hits, which is the number of hits divided by the number of records;
- **lift**: the lift, equal to (**hits%**)/b.
- **model CPH** and **optimal CPH** are defined below.

Table 1: A sample lift table

%	recs	hits	hits%	lift	model CPH	optimal CPH
5	1045	277	26.51	4.22	0.211	0.796
10	2090	378	18.09	2.88	0.288	1
15	3135	481	15.34	2.44	0.367	1
20	4180	563	13.47	2.15	0.429	1
25	5225	646	12.36	1.97	0.492	1
30	6270	711	11.34	1.81	0.542	1
35	7315	770	10.53	1.68	0.587	1
40	8360	821	9.82	1.56	0.626	1
45	9405	872	9.27	1.48	0.665	1
50	10450	929	8.89	1.42	0.708	1
55	11495	979	8.52	1.36	0.746	1
60	12540	1029	8.21	1.31	0.784	1
65	13585	1080	7.95	1.27	0.823	1
70	14630	1134	7.75	1.23	0.864	1
75	15675	1180	7.53	1.20	0.899	1
80	16720	1208	7.22	1.15	0.921	1
85	17765	1237	6.96	1.11	0.943	1
90	18810	1258	6.69	1.07	0.959	1
95	19855	1286	6.48	1.03	0.980	1
100	20900	1312	6.28	1	1.000	1

E.g., the top 10% of the list have 2,090 records and have 378 targets (hits), which gives the hit rate of

$$\frac{378}{2090} = 0.1809 = 18.09\%$$

The lift at 10% is $18.09/6.28 = 2.88$.

2.1 CPH: Cumulative Percent Hits measure

The lift curve usually goes from a number less than $1/b$ down to 1 and its properties depend on b. It is easier to analyze a closely related measure, called *Cumulative Percent Hits* or CPH, which always varies from 0 to 1.

DEFINITION 1. *Let* $\mathrm{CPH}(M,p)$ *be the cumulative percent hits in the top* p *percent of the model-score-sorted list for the model* M.

We will use both percentages (as in $\mathrm{CPH}(M,5\%)$) and fractions (as in $\mathrm{CPH}(M,0.05)$) as equivalent ways to indicate the selected sub-list.
The relationship between CPH and Lift is

$$\mathrm{CPH}(M,p) = p \times \mathrm{Lift}(M,p)$$

In table 1, for example,

$$\mathrm{CPH}(M,10\%) = \frac{378}{1312} = 0.288$$

By definition, for any model M,

$$\begin{aligned} \mathrm{CPH}(M,0) &= 0 \\ \mathrm{CPH}(M,100\%) &= \mathrm{CPH}(M,1) = 1 \end{aligned}$$

For an idealized random list R, and for any percentage p, we expect

$$\mathrm{CPH}(R,p) = p$$

We realize that an actual random selection will vary slightly from these percentages, but it is convenient to use this idealized random model for a comparison.
Let

$$\begin{aligned} T &= \text{number of targets}, \\ N &= \text{total number of records}, \\ b &= \frac{T}{N} = \text{base rate}. \end{aligned}$$

Let $\mathrm{CPH}(B,p)$ be the maximum possible CPH at p, corresponding to the optimal (best) model B. We can see that if $b < 5\%$, then

$$\mathrm{CPH}(B,5\%) = 1$$

If $5\% < b < 10\%$, then

$$\begin{aligned} \mathrm{CPH}(B,5\%) &= \frac{N \times 0.05}{T} \\ \mathrm{CPH}(B,10\%) &= 1, etc. \end{aligned}$$

2.2 SumCPH measure

Because CPH curve increases from 0 to 1 as p, percentage of list selected, goes from 0 to 1, a natural measure is the area under the CPH curve.

DEFINITION 2. *We define* $\mathrm{SumCPH}(M)$ *as the area under the* CPH *curve of the model* M, *i.e.*

$$\mathrm{SumCPH}(M) = \int_0^1 \mathrm{CPH}(M,p)dp$$

In practice we only have a finite number of points, so instead of the integral we will compute SumCPH as a sum instead of an integral, as discussed below.
For an idealized random model R, we have $\mathrm{CPH}(R,p) = p$ (the diagonal line). It is easy to see that $\mathrm{SumCPH}(R)$, which is the area under the diagonal line, is equal to 1/2.
For an optimal model, the highest CPH would be obtained if all the targets are grouped in the beginning of the list.
Figure 1 shows an example of an optimal model where $b = 0.03$. Here the optimal CPH increases in a straight line from 0 to 1 as p goes from 0 to 3%, and CPH stays at 1 as p goes from from 3% to 1. For an arbitrary b, the optimal CPH will increase in a straight line from 0 to 1 as p goes from 0 to b. The area under this line is $b \times 1/2$. Then CPH = 1 as p goes from from b to 1, giving area $(1-b) \times 1$. Hence, the area under the optimal CPH curve would be

$$\mathrm{SumCPH}(B) = b/2 + (1-b) \times 1 = 1 - b/2 \qquad (2)$$

To summarize, we can expect $\mathrm{SumCPH}(M)$ to be between $1/2$ and $1-b/2$, and higher SumCPH generally corresponds to a better model.

2.3 L-quality measure

However, SumCPH by itself is not an appropriate measure of overall lift quality for several reasons.
First, SumCPH is not intuitive - the meaning of 0.69 is not clear. Second, SumCPH by itself does not tell us how close it is to the maximum. The higher the base rate, the lower is the maximum possible SumCPH and the measure of lift quality should account for it.

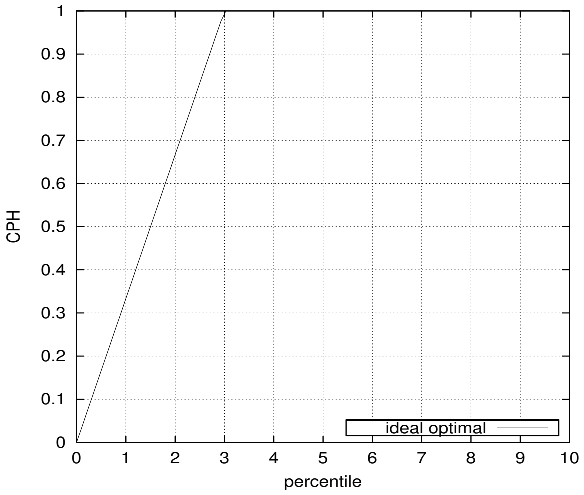

Figure 1: SumCPH of the optimal model

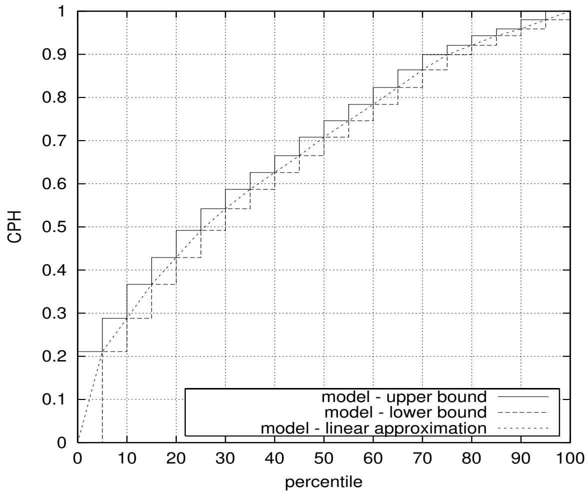

Figure 2: Upper and lower bounds on SumCPH

We would like a measure that, for the random model R and the optimal model B, would satisfy:

$$\text{L-quality}(R, w) = 0 \quad (3)$$
$$\text{L-quality}(B, w) = 1 \quad (4)$$

We propose a measure, which we call *L-quality*, satisfying these requirements, and defined as

$$\text{L-quality}(M) = \frac{\text{SumCPH}(M) - \text{SumCPH}(R)}{\text{SumCPH}(B) - \text{SumCPH}(R)}$$

Since $\text{SumCPH}(R) = 1/2$ and $\text{SumCPH}(B) = 1 - b/2$, we can rewrite this definition as

$$\text{L-quality}(M) = \frac{\text{SumCPH}(M) - 1/2}{1 - b/2 - 1/2} = \frac{2\,\text{SumCPH}(M) - 1}{1 - b}$$

L-quality can be viewed as a relative quality of the model, with the optimal model having a quality of 1 (or 100%), and the random model having a quality of 0. Negative quality means the model predictions are worse than random, which indicates data problems or bugs.
We defined L-quality for an exact SumCPH, which assumes that CPH value is known for every point. In practice, CPH is usually known only at certain points (quantiles). For example, in table 1 CPH value is known only in 5% intervals. The next section examines how to best approximate SumCPH from such data.

2.4 Upper and Lower bounds for SumCPH From Quantile Data

Because CPH is an accumulation of a non-negative quantity (number of hits), it is monotonically non-decreasing. Therefore, for any q such that $5\% < q < 10\%$, we have

$$\text{CPH}(M, 5\%) < \text{CPH}(M, q) < \text{CPH}(M, 10\%)$$

Thus, $\text{CPH}(M, 5\%)$ is a lower bound for $\text{CPH}(M, q)$, and $\text{CPH}(M, 10\%)$ is an upper bound. Figure 2 shows the upper and lower bounds on CPH for the model in table 1.
Using approach, we can estimate the step-wise lower and upper bounds, (denoted $\text{SumCPH}_{\text{HI}}$ and $\text{SumCPH}_{\text{LO}}$) for SumCPH. For each quantile, the $\text{SumCPH}_{\text{HI}}$ approximation will take the highest CPH value in that quantile, and $\text{SumCPH}_{\text{LO}}$ will take the lowest CPH value.
Let $\text{SumCPH}(M, p\%)$ denote the area under CPH curve from from 0 to $p\%$.
Let $\text{SumCPH}_{\text{HI}}(M, p\%, w)$ (resp. $\text{SumCPH}_{\text{LO}}(M, p\%, w)$) be an upper (resp. lower) bound for $\text{SumCPH}(M, p\%)$, measured with a step of w.
E.g., the upper and lower bounds for $\text{SumCPH}(M, 10\%)$ are

$$\begin{aligned}\text{SumCPH}_{\text{HI}}(M, 10\%, 0.05) &= 0.05\,\text{CPH}(M, 5\%) + \\ &\quad 0.05\,\text{CPH}(M, 10\%) \\ \text{SumCPH}_{\text{LO}}(M, 10\%, 0.05) &= 0.05\,\text{CPH}(M, 0\%) + \\ &\quad 0.05\,\text{CPH}(M, 5\%)\end{aligned}$$

More generally, let the model performance be measured in steps of w, i.e. w, $2w$, $3w$, ..., 1, where the number of steps is $S = 1/w$. Then

$$\begin{aligned}\mathrm{SumCPH_{HI}}(M,1,w) &= w\sum_{i=1}^{S}\mathrm{CPH}(M,iw)\\ \mathrm{SumCPH_{LO}}(M,1,w) &= w\sum_{i=1}^{S}\mathrm{CPH}(M,(i-1)w)\end{aligned}$$

Note that $\mathrm{CPH}(M,1)=1$. Subtracting the above two equations, we get

$$\mathrm{SumCPH_{HI}}(M,1,w)-\mathrm{SumCPH_{LO}}(M,1,w) = w\,\mathrm{CPH}(M,1)=w$$

We can apply these formulas to the CPH of the model M_1 in the table 1, getting

$$\begin{aligned}\mathrm{SumCPH_{HI}}(M_1,1,0.05) &= 0.691\\ \mathrm{SumCPH_{LO}}(M_1,1,0,05) &= 0.641\end{aligned}$$

If we apply the same step-wise approximation to an ideal random model R (where $\mathrm{CPH}(R,p)=p$), measured with a step of w, then we get

$$\begin{aligned}\mathrm{SumCPH_{HI}}(R,1,0.05) &=\\ 0.05\times(0.05+0.10+...+0.95+1) &=\\ 0.05\times 0.05\times(21\times 20)/2 = 0.525\end{aligned}$$

Likewise, $\mathrm{SumCPH_{LO}}(R,1,0.05)=0.475$.
For an arbitrary step width w,

$$\begin{aligned}\mathrm{SumCPH_{HI}}(R,1,w) &= 1/2+w/2\\ \mathrm{SumCPH_{LO}}(R,1,w) &= 1/2-w/2\end{aligned}$$

This confirms our intuition that the upper and lower bounds are about $w/2$ away from the center and suggests that the central estimate would be between them.
We can use high and low bounds for SumCPH to define similar high and low bounds for L-quality.

$$\begin{aligned}\text{L-quality}_{\mathrm{HI}}(M,1,w) &= \frac{\mathrm{SumCPH_{HI}}(M,1,w)-1/2}{1-b/2-1/2}\\ \text{L-quality}_{\mathrm{LO}}(M,1,w) &= \frac{\mathrm{SumCPH_{LO}}(M,1,w)-1/2}{1-b/2-1/2}\end{aligned}$$

2.5 Linear Estimate for SumCPH From Quantile Data

We can estimate $\mathrm{SumCPH}(M,1,w)$ (the area under CPH curve from p=0 to 1) by approximating CPH curve between adjacent quantiles with a straight line. Let such estimate be called $\mathrm{SumCPH_{lin}}(M,1,w)$. We note that the area under the straight line would be exactly the average between the upper bound and the lower bound estimates. Thus,

$$\begin{aligned}\mathrm{SumCPH_{lin}}(M,1,w) &=\\ w\sum_{i=1}^{S}\frac{\mathrm{CPH}(M,iw)+\mathrm{CPH}(M,(i-1)w)}{2} &=\\ w\frac{\mathrm{CPH}(M,0)+2\sum_{i=1}^{S-1}\mathrm{CPH}(M,iw)+\mathrm{CPH}(M,1)}{2}\end{aligned}$$

Since $\mathrm{CPH}(M,0)=0$ and $\mathrm{CPH}(M,1)=1$, we can simplify the above to

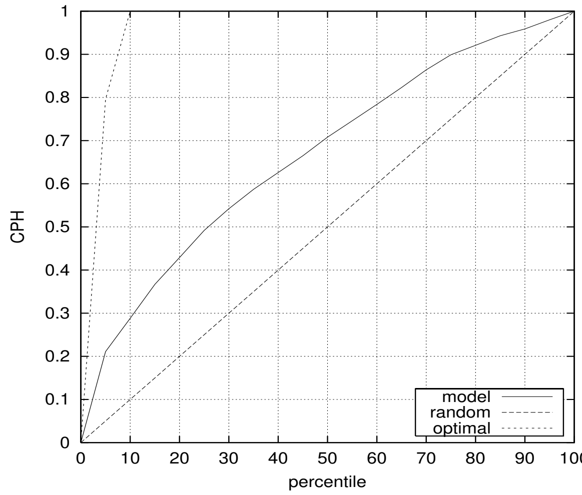

Figure 3: Optimal, Model, and Random CPH curves using linear approximation

$$\mathrm{SumCPH_{lin}}(M,1,w) = w(\sum_{i=1}^{S}\mathrm{CPH}(M,iw)-1/2) \quad (5)$$

Substituting into the definition of L-quality the values of $\mathrm{SumCPH}(R)=1/2$ and $\mathrm{SumCPH}(O)=1-b/2$, and using the linear estimate for SumCPH, we get

$$\text{L-quality}_{\mathrm{lin}}(M) = \frac{\mathrm{SumCPH_{lin}}(M,1,w)-1/2}{1-b/2-1/2} \quad (6)$$

Figure 3 shows the CPH values for a sample, random, and optimal models, using a linear approximation.

2.6 Variability of L-quality

One important question is how accurate and stable are L-quality estimates. We note that while the maximum error in $\mathrm{SumCPH_{lin}}$ is $w/2$, the average error expected from a linear approximation would be much smaller ($O(w^2)$ for smooth lift curves).
For a number of lift tables we had data at a 1% granularity. We compared L-quality obtained from a 1% table to $\mathrm{SumCPH_{lin}}$ obtained using only 5% steps.
Empirically, it seems that $\text{L-quality}_{\mathrm{lin}}(M,1,5\%)$ is about 0.2-0.5% smaller than $\text{L-quality}_{\mathrm{lin}}(M,1,1\%)$. The reason could be that the true CPH curve is usually convex and the linear approximation slightly underestimates it.
A separate question is how much will L-quality change if a model is applied to different random samples from the same population. We are in the process of doing such experiments.

Table 2: Comparing L-quality at 1% and 5% granularity

data	step	LQ-high	LQ-line	LQ-low
sr1-m31	1%	47.0	45.9	44.8
	5%	50.8	45.4	40.1
	diff		0.5	
sr2-m03	1%	47.7	46.6	45.5
	5%	51.6	46.3	40.9
	diff		0.3	
sr1-m35	1%	34.4	33.3	32.3
	5%	38.4	33.1	27.9
	diff		0.2	
sr2-m05	1%	47.0	45.9	44.8
	5%	50.8	45.4	40.1
	diff		0.5	
sr4-m26	1%	37.1	36.1	35.0
	5%	40.9	35.6	30.2
	diff		0.5	

3. USING L-QUALITY

For table 1, we have $\text{SumCPH}_{\text{HI}} = 0.692$, $\text{SumCPH}_{\text{lin}} = 0.667$, and L-quality $= 35.6\%$.

In [4] we analyzed a number of lift curves and found that a typical lift curve T has $\text{CPH}(T,p) = \sqrt{p}$, which corresponds to L-quality$(T) = 34.6\%$.

Thus if a model has L-quality below 34.6%, it can probably be improved.

In our practice, we found that a good model rarely achieves a L-quality above 50%. L-quality of 80% and above is usually an indication of a leaker in the model.

We were able to apply L-quality to lift curves (in 5% increments) from the KDD-Cup 1997 competition. The L-quality of the top 3 winners was very close:

- UrbanScience (Gainsmarts tool): 43.3%
- Charles Elkan (BNB tool): 42.7%
- SGI (Mineset tool): 41.7%

The average L-quality of the remaining 13 competitors was only 19.1%, so the 3 winners were indeed significantly better than the average.

Our preliminary results suggest that a difference between the first two L-quality values, measured with a 5% step is not significant. Thus L-quality measure confirms UrbanScience and Charles Elkan tie for the first place, while SGI was very close but still slightly behind in third place.

4. CONCLUSION

In this paper, we addressed the problem of measuring an overall quality of the lift curve. We formally defined such measure, called L-quality, which has a number of good properties, including intuitive clarity, and ability to compare lift curves when cost information is not available. We also analyzed how to estimate true L-quality from quantized lift tables and analyzed typical and KDD-Cup winning L-quality values.

5. REFERENCES

[1] `http://www.kxen.com`

[2] Ismail Parsa "KDD-Cup 1997 results", `http://www.kdnuggets.com/news/97/n25.html#item2`

[3] Foster Provost, Tom Fawcett "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions", Proceedings of KDD-97, AAAI Press, 1997.

[4] Gregory Piatetsky-Shapiro, Brij Masand "Estimating Campaign Benefits", Proceedings of KDD-99, ACM Press, 1999.