# Evaluation of Prediction Models for Marketing Campaigns

Saharon Rosset
Amdocs Ltd.
and Stanford University
saharonr@amdocs.com

Einat Neumann
Amdocs (Israel) Ltd.
and Tel-Aviv University
einatn@amdocs.com

Uri Eick
Amdocs (Israel) Ltd.
urieick@amdocs.com

Nurit Vatnik
Amdocs (Israel) Ltd.
nuritv@amdocs.com

Izhak Idan
Amdocs (Israel) Ltd.
izhaki@amdocs.com

## ABSTRACT

We consider prediction-model evaluation in the context of marketing-campaign planning. In order to evaluate and compare models with specific campaign objectives in mind, we need to concentrate our attention on the appropriate evaluation-criteria. These should portray the model's ability to score accurately and to identify the relevant target population. In this paper we discuss some applicable model-evaluation and selection criteria, their relevance for campaign planning, their robustness under changing population distributions, and their employment when constructing confidence intervals. We illustrate our results with a case study based on our experience from several projects.

## Keywords

Model Evaluation, Marketing Campaigns, Performance Measures, Confidence Intervals.

## 1. INTRODUCTION

When dealing with marketing applications, such as campaign management, the issue of evaluating prediction models is twofold. First, the evaluation has to be statistically sound, allowing us to compare models, choose among them and estimate their expected future performance. Second, and perhaps more important, we need to evaluate models with regard to the way they will be utilized from a business perspective. For example, suppose we are building a scoring model to predict voluntary churn (customer's propensity for disconnecting services) in order to identify the target population for a retention campaign. If in the campaign we intend to contact only the 2% of our customers who are at highest churn risk, it seems unreasonable to evaluate a suggested model using accuracy over a full test data set. The model's performance on 98% of the population is irrelevant to the campaign goal. [6] and [8], among others, present flexible and efficient techniques for evaluating models with regard to a wide variety of goal functions. However, we have found the statistical analysis of the most relevant scores for planning campaigns to be lacking, and have compiled an array of tools and techniques to fill the gaps.


KDD 01 San Francisco CA USA


In this paper we discusses some of the approaches we take when evaluating model-performance in the context of campaign planning and executing. We also present statistical issues that arise when attempting to combine relevance and rigor in the evaluation process. The main results we present are:

- Description of the requirements from appropriate evaluation techniques for campaign planning and comparison of various relevant evaluation measures (Section 2).
- Methodology for applying some of the evaluation measures (Section 3). This includes issues such as, score adjustment, distribution of scores and methods for constructing confidence intervals.
- A case study (Section 4), illustrating the importance and usefulness of combining contextual and statistical considerations in model-evaluation.

## 2. MODEL EVALUATION

We begin our discussion at the point where a scoring model has been constructed. We disregard the method or algorithm that were used to create the model, and concentrate on the means for evaluating it, given the campaign objectives. A different approach would be to consider the objectives while constructing the model ([1] and [4]). Once we have a candidate model, we want to estimate its expected performance on unlabeled data. Our standard model evaluation methodology is:

1. Evaluate the models' performance on an independent test set (labeled data that has been set aside beforehand and not used in training the model).
2. Adjust the models' score to fit the full population distribution, in case it is expected to be different from the sample distribution used for training and test (Section 3.1).

We focus our discussion on the performance measures, which are of interest for campaign planning and analysis, and their statistical properties.

### 2.1 Planning Campaigns

When planning a campaign, one seeks to identify individuals most likely to respond to the campaign. Due to budget restrictions the number of individuals to be approached in the campaign is limited. Thus there is a need for a good model for selecting the target segment and its performance on the rest of the population is of little or no consequence. The success of such a model is usually measured by the amount of responders captured within the targeted population. This amount can be measured in two different ways:

- How much better are we doing by using our model to select the target population relative to a random selection of the target population. This measure is known as the *Lift*. For example: instead of reaching 2% of the responders when approaching randomly 2% of the population, we could reach 16% of the responders by approaching the top model-scored 2% of the population. In this case we are improving the random model by 8 times, i.e. the lift at 2% is 8.
- How frequently do we expect to encounter a responder when running our campaign? This measure is expressed by the *Response Rate*. For example: when carrying out a telemarketing campaign, we may be interested in knowing once in how many calls we should expect a responder.

These two measures capture the essence of a models' usefulness for campaign planning, from two business perspectives. The measures are also mathematically equivalent but have a different behavior in the face of changing population distribution, as described in Section 2.2.1.

## 2.2 Performance Measures

Having established the need for adapted model evaluation in the context of campaign planning and mentioned some useful measures, we now concentrate on the statistical properties of evaluation measures, and consider their robustness in changing population distributions. We roughly divide them into two categories: *overall* performance measures and measures calculated *per cutoff points*. The evaluation process commences with sorting all test entities according to their model-produced scores. This ranked list serves as the basis for calculation of all possible performance measures, together with the following terminology:

- $A$, $B$ – total number of responders and non-responders, respectively.
- $A_j$, $B_j$ – total number of responders and non-responders, respectively, in the $j$-th top quantile.
- $j*(A+B)$ or $(A_j + B_j)$ – all cases in the $j$-th top quantile
- $A/(A+B)$ – overall response rate

### 2.2.1 Measures at Pre-Specified Cutoff Points

Response Rate

The Response Rate (RR) represents the responders' percentage you reach out of all the customers approached in a campaign. Customers approached in a campaign are the $j$-th top quantile and the RR is the response rate within that quantile: $RR_{(j)} = A_j / (A_j + B_j)$

This measure is useful for calculating the expected profit from a campaign, however, it is extremely sensitive to the overall response rate. It drops almost linearly with the drop of the *overall* response rate in the population. Thus, models created based on populations with different response rates can not be compared using this measure, without applying an appropriate normalization. Furthermore, if the response rate in the "future" population on which a campaign is going to be run is unknown, there is no way to get a reliable estimate of the RR for that campaign.

Lift

The Lift measures the ratio between the RR and the overall response rate:

$$Lift_{(j)} = \frac{RR_{(j)}}{\left(\frac{A}{A+B}\right)} = \frac{\left(\frac{A_j}{A_j+B_j}\right)}{\left(\frac{A}{A+B}\right)} = \frac{\left(\frac{A_j}{A}\right)}{\left(\frac{A_j+B_j}{A+B}\right)} = \frac{\left(\frac{A_j}{A}\right)}{j}$$

The Lift shows directly how much better would be a campaign based on the model than a campaign based on a random selection. Thus it is also a useful measure for campaign evaluation, from a different perspective than RR. The Lift is somewhat sensitive to the overall response rate (it mildly increases as the overall response decrease), but much less than the RR. Despite this slight disadvantage it is an intuitive evaluation criterion and a common (and sensible) choice.

Response Non-Response Ratio

The Response to Non-Response Ratio (RNR) is the ratio between the percentage of all responders and the percentage of all non-responders in the top $j$-th quantile: $RNR_{(j)} = \left(\frac{A_j}{A}\right) \Big/ \left(\frac{B_j}{B}\right)$

Statistically the RNR is robust and independent of the overall response rate. Thus, it is easy to compare model performance with it. However, there is no intuitive way to define the RNR in terms of planning campaigns and it is hard to use it in order to illustrate campaign effectiveness.

Table 1 summarizes the advantages and limitations of each measure.

**Table 1. Comparison of Cut-Point Measures**

| Measure | Sensitivity to population response rate | Interpretation for campaign evaluation |
|---|---|---|
| RR | Extremely Sensitive | Frequency of encountering a responder |
| Lift | Somewhat Sensitive | Improvement over a random model |
| RNR | Invariant | No immediate interpretation |

### 2.2.2 Overall measures

In some cases it is difficult or undesirable to decide in advance on a specific size for the target population, or a specific model may be used for many campaigns. In such cases, it may be preferable to estimate the models' performance simultaneously with regard to a whole range of potential targets.

Misclassification Rate

If the task is classification, i.e. each entity is to be classified as a responder or a non-responder, it is necessary to set a threshold score. The Misclassification Rate (MCR) is the percentage of entities classified incorrectly among all entities (an alternative suggested by [5] is the Misclassification Cost that weights costs into the error calculation). In the campaign-planning context the MCR is usually inappropriate since a campaign inherently focuses on some small sub-populations and not the entire population. Further discussion of its inadequacy can be found in [6].

Receiver Operating Characteristic (ROC) Curve

In order to define the ROC Curve we first present the following definitions:

*Sensitivity* – the percent of responders classified as responders.

*Specificity* – the percent of non-responders classified as non-responders.

When classifying, increasing the cutoff-point increases *Sensitivity* and decreases *Specificity*. ROC curve is a plot of the *Sensitivity* against *1-Specificity* at many cut-points. The area under the curve (AUC) is a measure of a models' ability to separate responders from non-responders. When comparing two models by their ROC curves, we're actually comparing their RNR at all possible cutoff points simultaneously. Further discussion can be found in [6], who also introduce the ROC convex hull method, for comparing a large number of classifiers.

Gain Chart

A Gain Chart (a.k.a Cumulative Lift chart) is a graph displaying the proportion of all responders vs. the proportion of the population (the quantile) sorted according to the model scores. Had the population been sorted randomly we would have expected each quantile to include the same response proportion. Similarly to the ROC curve, a Gain Chart displays the Lift in all quantiles simultaneously. The area under the curve is a measure of the models' relative ability to identify responders. In the direct marketing domain this chart is sometimes referred to as the *Pareto Curve* since it expresses a similar notion as the "80/20 rule" [7].

The relationship between the two graphs

ROC and Gain Chart are methods for evaluating the ranking performance of models, thus they make good criteria when the aim is to identify high or low tendency among the whole population. [8] demonstrates that the two diagrams are equivalent and presents several significance tests for the difference between AUCs that can be used to determine differences between overall performance of models. As with the RNR and the Lift, ROC Curve is good for comparing models, especially when response rate may vary, and the Gain Chart is good for evaluating campaign targeting effectiveness.

# 3. PREDICTING MODEL PERFORMANCE

The performance measures, discussed in details in the previous section, are usually calculated on a test sample data set. These measures need to be adjusted to the full population, in case its distribution is different from the sample distribution used for training and test (the sample may be biased due to intentional under-sampling of the larger class). The transformation method is presented in Section 3.1. The next stage after the transformation is to calculate reliable predictions based on the performance measures for future data. We do that by building proper Confidence Intervals (CI's). Methods for building exact and approximate CIs are introduced in Section 3.2. We limit the discussion to the Lift and RR measures. Note that in the previous section lift was defined as $(A_j/A)/j$. Assuming j is fixed and to simplify the formula, this section refers to $(A_j/A)$ as the lift.

## 3.1 From Sample to Population

Models are usually evaluated on a test set (TS) put aside from the sample population. In many cases the response rate in the full population (FP) is very small but in the sample the 2 classes are much more balanced. Suppose we sort the TS according to the model-scores and calculate certain measures at each percentile. Due to the different response rates, TS percentiles do not correspond to percentiles in the FP. It is therefore necessary to translate the percentiles from the TS to the FP and we call this procedure the "Inverse Transformation". [9] introduce a similar issue - they account for the differences between the training-set consideration test-set distributions when measuring error rate.

The following quantities are given prior to performing the transformation:

$A$, $B$ - the number of responders and non-responders in the FP, respectively.

$a$ , $b$ - the number of responders and non-responders in the TS, respectively.

$a_i$ , $b_i$ - the number of responders and non-responders in percentile i in the TS, respectively.

The first step is to extrapolate each percentile pair $(a_i, b_i)$ in the TS to $(\hat{A}_i, \hat{B}_i)$ in the FP, where $\hat{A}_i = a_i(A/a)$ and $\hat{B}_i = b_i(B/b)$. Each extrapolated percentile pair $(\hat{A}_i, \hat{B}_i)$ does not add up to a FP percentile, so TS percentiles are merged or split in order to attain FP percentiles.

Lets assume that instead of TS single percentiles we are using accumulated TS percentiles and we want to transform them into accumulated FP percentiles. Let *d(j)* be the number of top TS percentiles necessary to comprise the top j FP percentiles. Note that *d(100) = 100* and that d(j) might not be integer. The estimator for the number of responders and non-responders in the top j FP percentiles are:

$$\hat{A}_j = \sum_{i=1}^{d(j)} \hat{A}_i = \frac{A}{a} a_{d(j)} \quad (1)$$

$$\hat{B}_j = \sum_{i=1}^{d(j)} \hat{B}_i = \frac{B}{b} b_{d(j)} \quad (2)$$

Thus, the estimated lift for the FP at percentile j is $\hat{A}_j / A$ and the estimated RR for the FP at percentile j is $\hat{A}_j /(A_j + B_j) = \hat{A}_j / j(A+B)$.

## 3.2 Confidence Intervals

Percentile point-estimators are not sufficient for evaluating the model predictive ability. When attempting to predict a model's performance on future data we also build confidence intervals. There are two main uses for CI's in the context of model evaluation:

- The Lower Bound (LB) of a one-sided CI can be used to give a realistic (more conservative) approximation for the future performance of a model.
- A two-sided CI for the difference between scores of two models can be used to compare their performance.

Here we present the generation of one-sided CI's for a single model, but the extensions to two-sided CI's are trivial. The proportions we are interested in estimating are:

Lift : $$p_j^{(1)} = \frac{A_j}{A} \quad (3)$$

Response Rate: $$p_j^{(2)} = \frac{A_j}{A_j + B_j} \quad (4)$$

In Section 3.2.1 we discuss the exact distribution of the estimators and in Section 3.2.2 we propose alternative distribution approximations to use when constructing CIs for these proportions. Section 3.2.3 compares the different approaches with empirical results.

### *3.2.1 The Exact Distribution of the Lift and the Response-Rate Estimators*

Our aim is to construct CIs for the lift and the RR based on the sample lift and sample RR.

Sample Lift: $$\hat{p}_j^{(1)} = \frac{a_d}{a} \quad (5)$$

Sample Response Rate: $$\hat{p}_j^{(2)} = \frac{a_d}{a_d + b_d} = \frac{a_d}{d(a+b)} \quad (6)$$

Note, that instead of *d(j)* we use *d* in this section (assuming j is fixed ). In order to use these estimators, it is actually necessary to know the distribution of $a_d$ . Since we're not dealing with infinite populations but rather the total test population sizes *a* and *b* are given, we are in a "hyper-geometric"-like setting. The hyper-geometric (HG) distribution is inappropriate, of course, since it inherently assumes we are selecting a sub-population at random. We are selecting a sub-population according to our model, which we hope and assume is non-random. What we need is a "biased hyper-geometric" (BHG) distribution:

$a_d \sim BHG\ (a+b,\ a,\ d(a+b),\ p_j^{(1)})$

Which would have the mean $a \cdot p_j^{(1)}$ compared to $a \cdot d$ in the HG distribution, and variance approximately $a \cdot p_j^{(1)}(1-p_j^{(1)})$, representing our knowledge of the overall sums *a* and *b*.

We have not found an appropriate distribution in the statistics literature. We are currently working on formulating such distribution and consider this an interesting research issue.

### *3.2.2 CIs Based on Approximations*

Since the formula of the exact distribution is not known explicitly, in practice we can use either Binomial or HG "like" approximations for constructing CIs.

#### *Binomial CIs*

Lift Confidence Interval

From (1) and (5): $$\hat{p}_j^{(1)} = \frac{a_d}{a} = \frac{\hat{A}_j}{A} \quad (7)$$

We assume that $a_d \overset{\circ}{\sim} Bin\left(a, p_j^{(1)}\right)$ and use the following approximation - $\hat{p}_j^{(1)} \overset{\circ}{\sim} Normal\left(p_j^{(1)}, \frac{p_j^{(1)}(1-p_j^{(1)})}{a}\right)$.

Thus, we can directly calculate the LB of a one-sided CI for the lift (using $\hat{p}_j^{(1)}$ instead of the unknown $p_j^{(1)}$ in the variance):

$$LB_j^{(1)} = \hat{p}_j^{(1)} - Z_{1-\alpha}\sqrt{\frac{\hat{p}_j^{(1)}(1-\hat{p}_j^{(1)})}{a}} \quad (8)$$

Response Rate Confidence Interval

The sample RR (6) can also be expressed as following:

$$\hat{p}_j^{(2)} = \frac{1}{1+\frac{b_d}{a_d}} \quad (9)$$

We assume that $a_d \overset{\circ}{\sim} Bin\left(a_d + b_d, p_j^{(2)}\right)$ and again use the normal approximation: $\hat{p}_j^{(2)} \overset{\circ}{\sim} Normal\left(p_j^{(2)}, \frac{p_j^{(2)}(1-p_j^{(2)})}{a_d + b_d}\right)$.

The LB of a matching one-sided CI is:

$$LB_j^{(2)} = \hat{p}_j^{(2)} - Z_{1-\alpha}\sqrt{\frac{\hat{p}_j^{(2)}(1-\hat{p}_j^{(2)})}{a_d + b_d}} \quad (10)$$

But we are interested in a different proportion:

$$\hat{p}_j^{(2^*)} = \frac{\hat{A}_j}{\hat{A}_j + \hat{B}_j} = \frac{\left(\frac{A}{a}\right)a_d}{\left(\frac{A}{a}\right)a_d + \left(\frac{B}{B}\right)b_d} = \frac{a_d}{a_d + factor \cdot b_d} = \frac{1}{1 + factor \cdot \frac{b_d}{a_d}}$$

where $$factor = \frac{\left(\frac{B}{b}\right)}{\left(\frac{A}{a}\right)}. \quad (11)$$

from (9): $$\frac{b_d}{a_d} = \frac{1-\hat{p}_j^{(2)}}{\hat{p}_j^{(2)}}$$

so from (9) and (11):

$$\hat{p}_j^{(2^*)} = \frac{1}{1 + factor\left(\frac{1-\hat{p}_j^{(2)}}{\hat{p}_j^{(2)}}\right)} = \frac{\hat{p}_j^{(2)}}{\hat{p}_j^{(2)} + factor\left(1-\hat{p}_j^{(2)}\right)} = f(\hat{p}_j^{(2)})$$

$\hat{p}_j^{(2^*)}$ is a monotonic increasing function of $\hat{p}_j^{(2)}$, therefore it is possible to calculate a LB for $p_j^{(2^*)}$ based on the LB for $p_j^{(2)}$:

$$LB_j^{(2^*)} = f(LB_j^{(2)}) = \frac{LB_j^{(2)}}{LB_j^{(2)} + factor\ (1 - LB_j^{(2)})} \quad (12)$$

That is the proper way for constructing this CI, since our uncertainty is at the TS level. Thus, the CI should be calculated based on the TS quantities and then "inverse transformed" to the FP.

The Connection between the two Binomial CI's

The sample lift and the sample RR can each be represented as a monotonic increasing function of the other. It is therefore possible to calculate the CI for each based on the CI of the other. Under certain conditions, doing that might result in a shorter CI. We illustrate this notion only on constructing the Lift CI based on the RR CI, but it obviously works in the opposite direction as well.

From (5) and (6): $$\hat{p}_j^{(1)} = g(\hat{p}_j^{(2)}) = \frac{d(a+b)}{a}\hat{p}_j^{(2)} \quad (13)$$

From (13) and (10):

$$LB_j^{(1^*)} = g(LB_j^{(2)}) = \frac{d(a+b)}{a} LB_j^{(2)} =$$

$$= \frac{d(a+b)}{a}\left( \hat{p}_j^{(2)} - Z_{\dots}\sqrt{\frac{\hat{p}_j^{(2)}(1-\hat{p}_j^{(2)})}{d(a+b)}} \right) =$$

$$= \dots = \hat{p}_j^{(1)} - Z_{\dots}\sqrt{\frac{\hat{p}_j^{(1)}(1-\hat{p}_j^{(2)})}{a}} \qquad (14)$$

Comparing (6) and (12):

$$LB_j^{(1)} = \hat{p}_j^{(1)} - Z_{\dots}\sqrt{\frac{\hat{p}_j^{(1)}(1-\hat{p}_j^{(1)})}{a}}$$

$$LB_j^{(1^*)} = \hat{p}_j^{(1)} - Z_{\dots}\sqrt{\frac{\hat{p}_j^{(1)}(1-\hat{p}_j^{(2)})}{a}}$$

The alternative $LB_j^{(1^*)}$ will achieve a shorter CI under the following condition:

$$(1-\hat{p}_j^{(2)}) < (1-\hat{p}_j^{(1)}) \Leftrightarrow \hat{p}_j^{(2)} > \hat{p}_j^{(1)} \Leftrightarrow \frac{a_j}{d(a+b)} > \frac{a_j}{a}$$

$$\Leftrightarrow d(a+b) < a \Leftrightarrow d < \frac{a}{(a+b)}$$

Obviously, this shift from one proportion to another would produce the same LB if we used the exact distribution. However, since the binomial approximations are conservative (with larger variance), we believe that given a precentile, it's acceptable to choose the method that yields a shorther CI.

### *Hyper-Geometric "like" CIs*

When applying the normal approximation, a less conservative approach would be, to use the HG-like approximate variance instead of the binomial variance:

$$\hat{p}_j^{(1)} \overset{\circ}{\sim} Normal\left( p_j^{(1)}, \frac{p_j^{(1)}(1-p_j^{(1)})(1-p_j^{(2)})}{a} \right) \quad \text{and}$$

$$\hat{p}_j^{(2)} \overset{\circ}{\sim} Normal\left( p_j^{(2)}, \frac{p_j^{(2)}(1-p_j^{(2)})(1-p_j^{(1)})}{a_d + b_d} \right).$$

Just as demostrated for the binomial based CIs we construct parallel HG based CIs $LB_j^{(1)}$ and $LB_j^{(2^*)}$ for the Lift and the RR, respectively.

$$LB_j^{(1)} = \hat{p}_j^{(1)} - Z_{1-\alpha}\sqrt{\frac{\hat{p}_j^{(1)}(1-\hat{p}_j^{(1)})(1-\hat{p}_j^{(2)})}{a}}$$

$$LB_j^{(2^*)} = f(LB_j^{(2)}) = \frac{LB_j^{(2)}}{LB_j^{(2)} + factor \quad (1 - LB_j^{(2)})}$$

where $LB_j^{(2)} = \hat{p}_j^{(2)} - Z_{1-\alpha}\sqrt{\frac{\hat{p}_j^{(2)}(1-\hat{p}_j^{(2)})(1-\hat{p}_j^{(1)})}{a_d + b_d}}$

In this case constructing a LB for the lift based on the RR (and vice versa) will produce the same LB as constructing it directly.

### *3.2.3 Comparing the CIs with Empirical Results*

Our experience shows that the CIs built based on approximations are generally satisfactory. Furthermore, we made an experiment with real data that gave an idea of the quality of the practical methods we suggest. We implemented bootstrap sampling [2] and compared the various CIs to the empirical bootstrap distribution. Table 2 displays the Lift 99% CI-LB based on the binomial approximation directly (B), via the RR (B*), and based on the HG-like approximation (HG). These LBs are compared to the 1% bootstrap quantile statistic. For example, at the 3rd percentile the Lift LBs obtained by the B, B* and HG approximations are 0.170, 0.239, 0.234, respectively, and all three are more conservative than the Lift LB obtained by the bootstrap - 0.278.

**Table 2. Comparing methods for constructing Lift LBs**

| Percentile | B | B* | HG | Bootstrap |
|---|---|---|---|---|
| 1% | 0.040 | 0.102 | 0.102 | 0.123 |
| 3% | 0.170 | 0.239 | 0.243 | 0.278 |
| 5% | 0.279 | 0.336 | 0.344 | 0.373 |
| 10% | 0.438 | 0.471 | 0.479 | 0.503 |
| 20% | 0.628 | 0.635 | 0.649 | 0.680 |
| 50% | 0.970 | 0.954 | 0.972 | 0.965 |

The experiment conclusions are:

- All methods generated very close LBs – not seen in Table 2, which describes only the coverage of the CI's.
- For most percentiles, at least 99% of the lift observations according to the bootstrap distribution where higher than the approximated LBs. This means that in this example (for most percentiles), the CI's produced by the practical methods have a confidence level of at least 99%.

## 4. CASE STUDY

This work was done by the Data Mining group, which is part of the R&D Business Insight unit at Amdocs Ltd. Amdocs is a leading provider of CRM, Billing and Order Management solutions to the communications and IP industry worldwide. The Amdocs Business Insight™ suite includes solutions for customer retention, reduction of bad debt, credit scoring, collection optimization, customer profitability analysis, sales analysis (cross/up sell) and relationship optimization.

In this section we describe, our experience in projects where we have found some of the concepts described in this paper to be useful. The type of model we consider is a prediction model for a retention campaign, in which responders are potential churners and the overall response rate is the overall churn rate.

The performance of a suggested new model was compared to that of a legacy model. Initially the legacy model's RR at 10% was 2.75 times better than the new model and thus it was concluded (mistakenly) that the legacy's prediction is better. However, when the actual evaluation process was investigated, it turned out that the two models were evaluated based on different test populations. The population used to evaluate the legacy's model had a 4.5 times higher overall churn rate. Given that fact, the RR was not the appropriate comparison criterion, since it is biased in

favor of the model evaluated based on a population with higher churn rate. We considered the lift as an alternative, since we use it often. When we calculated the lift, it turned out that the new model's lift was 1.62 times higher. Yet, as stated previously this measure too is not completely objective and independent of the population distribution. The lift for a smaller churn rate population is slightly higher. In order to compare the legacy and the new model, we had to use a more robust measure so we chose the RNR, which for the new model was 1.57 times greater than the legacy model. Table 3 summarizes the values of each measure for each model.

**Table 3. Evaluating two models based on populations with different churn rates (measured at the 10th percentile)**

| Model | Churn Rate | RR | Lift | RNR |
|---|---|---|---|---|
| Legacy Model | 1:60 | 6.6% | 4 | 4.2 |
| New Model | 1:273 | 2.4% | 6.5 | 6.6 |

To demonstrate the principle difference between the measures we introduce a model we built for churn prediction. For this single model we calculate each measure and display its value after performing the inverse transformation (twice according to two different churn rates). Figure 1 demonstrates the extreme sensitiveness of the RR, the relative robustness of the lift, and the high robustness of the RNR.

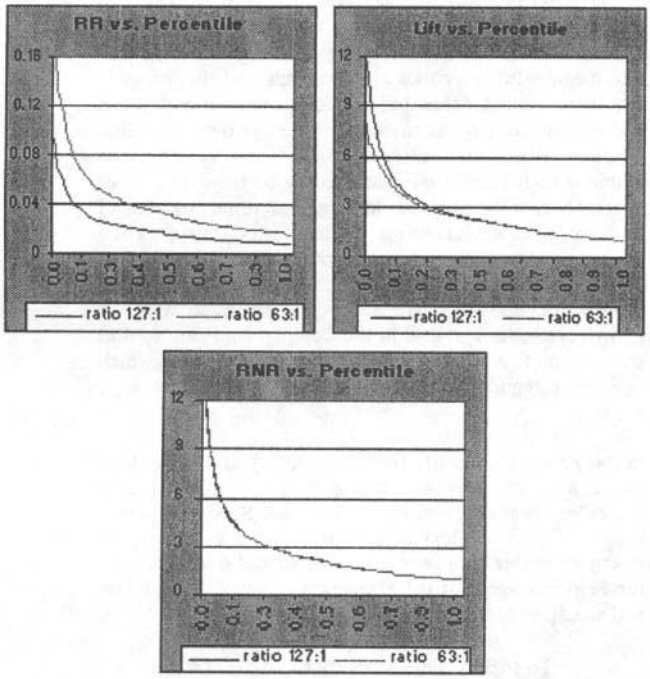


**Figure 1. The impact of changing churn-rate over model-evaluation measures**

**RR, Lift and RNR are portrayed vs. the ranked population (by percentiles). The two lines show the same model results transformed into two different populations using their different churn rates.**

## 5. CONCLUSION

In this paper we discussed a few model-evaluation criteria, their robustness under changing population distributions, and their relevance for campaign planning. We demonstrated how problematic a comparison based on a non-robust measure like *RR* may be, and commended the use of the *Lift* and *RNR* measures. Still in many cases *RR* is what the user expects. Discussing the robustness of a measure is usually not the best strategy for convincing the end-user which model to choose, therefore we introduced the *inverse transformation*. Inverse transformation, which was previously described for the purpose of transforming test-set class-distribution into full-population distribution, helps create a uniform presentation. After performing the proper transformation, it is possible and correct to compare models (with no bias) based on the RR. Such uniform presentation would be more intuitive for decision-makers.

## 6. ACKNOWLEDGMENTS

We would like to thank Gadi Pinkas, Aron Inger and Isabel Sasoon of Amdocs (Israel) Ltd. for their counseling.

## 7. REFERENCES

[1] Bhattacharyya, S. (1998). Direct Marketing Response Models using Genetic Algorithms. In *Proceedings of KDD-98*, pp.144-148. Menlo Park, CA: AAAI Press.

[2] Efron, B. and Tibshirani, R. (1993). Introduction to the Bootstrap. Chapman & Hall, New-York.

[3] Ling, C. X.; Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. In *Proceedings of KDD-98*, pp. 73-79. Menlo Park, CA: AAAI Press.

[4] Masand, B. and Piatetsky-Shapiro, G. (1996). A Comparison of Approaches for Maximizing Business Payoff of Prediction Models. In *Proceedings of KDD-96*, pp.195-201. Menlo Park, CA: AAAI Press.

[5] Pazzani, M.; Merz, \c.; Murphy, P.; Ali, K.; Hume, T. and Brunk, C. (1997). Reducing Misclassification Cost. In *Proceedings of the 11th International Conference on Machine Learning*, pp.217-225.

[6] Provost, F. and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distribution. In *Proceedings of KDD-97*, pp. 43-48. Menlo Park, CA: AAAI Press.

[7] Roberts, M. L. and Berger, P. D. (1999). Direct Marketing Management, pp. 97-101. Upper Saddle River, N.J.: Prentice Hall.

[8] Rosset, S. (1999). Ranking - Methods for Flexible Evaluation and Efficient Comparison of Two-Class Models. http://www-stat.stanford.edu/~saharon/finalthesis.zip.

[9] Weiss, G. M. and Provost, F. (2001). The Effects of Class Distribution on Classifier Learning. Technical Report ML-TR-43, Department of Computer Science,Rutgers University. http://www.cs.rutgers.edu/~gweiss/papers/class-distr.pdf.