# Ranking-Based Evaluation of Regression Models

Saharon Rosset, Claudia Perlich, Bianca Zadrozny
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598
{srosset, reisz, zadrozny}@us.ibm.com

## Abstract

*We suggest the use of ranking-based evaluation measures for regression models, as a complement to the commonly used residual-based evaluation. We argue that in some cases, such as the case study we present, ranking can be the main underlying goal in building a regression model, and ranking performance is the correct evaluation metric. However, even when ranking is not the contextually correct performance metric, the measures we explore still have significant advantages: They are robust against extreme outliers in the evaluation set; and they are interpretable. The two measures we consider correspond closely to non-parametric correlation coefficients commonly used in data analysis (Spearman's $\rho$ and Kendall's $\tau$); and they both have interesting graphical representations, which, similarly to ROC curves, offer useful "partial" model performance views, in addition to a one-number summary in the area under the curve. We illustrate our methods on a case study of evaluating IT Wallet size estimation models for IBM's customers.*

## 1 Introduction

The standard approach to evaluating regression models on holdout data is through additive, residual-based loss functions, such as squared error loss or absolute loss. These measures are attractive from a statistical perspective as they have likelihood interpretations and from an engineering or scientific perspective because they often represent the "true" cost of the prediction errors.

In this paper we propose a different approach to the evaluation of regression models, through their success in ranking the test set observations in the correct order. There are several reasons why ranking-based evaluation of regression models is interesting. First, ranking may be the real goal in building the prediction model. That is, in some cases we may just want to be able to separate the entities which are likely to have high response from the ones which correspond to low response. An example is given in the case study we present in Section 5, where the goal is to identify companies with high potential IT spending (IT Wallet). These are preferred targets for vigorous sales efforts by IBM. If we assume that IBM's sales resources are fixed, then identifying the largest IT Wallets, regardless of their actual numeric value, is the best support a regression model can give.

Second, ranking-based measures are quite interpretable. The two main evaluation methods we consider allow us to draw interpretations and connections:

- We build graphical representations similar to ROC curves for classification [3], such that points on the graph represent different evaluations of model performance on "partial ranking", and the area under the curve is a "one-number" summary of overall model ranking performance.
- We can connect the one-number ranking performance summaries to commonly used non-parametric statistical correlation measures: Spearman's $\rho$ and Kendall's $\tau$. We can use the theory developed in the Statistics literature to calculate approximate, *assumption free* confidence intervals for $\tau$, useful in model evaluation and selection.

Third, ranking-based evaluation is robust. That is, it is only mildly affected by errors — even gross ones — both in the values of features and in the measured response. In Section 4 we define a concept of "evaluation robustness" and show that ranking-based evaluation methods are very robust, whereas commonly used additive, residual-based evaluation measures, like mean squared error or mean absolute error, are extremely non-robust for evaluation. Some residual-based methods, like median absolute error, are robust, but tend to ignore completely the error measure on many observations. As such, they can be argued to be "throwing the baby with the water". Ranking-based evaluation, on the other hand, considers the evidence from all

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

observations, while not allowing gross errors to affect the overall evaluation measure too strongly.

Our paper is organized as follows. In Section 2 we introduce an evaluation nomenclature, describe the standard, residual-based evaluation approaches, define the concept of ranking-based evaluation, and introduce our suggested ranking-based measures. We analyze these measures, their interpretations, their statistical properties, and their visualizations in Section 3. Section 4 is devoted to the topic of "evaluation robustness" — we define an evaluation robustness measure and show that the residual-based additive methods are non-robust, while ranking-based measures are highly robust. In Section 5, we present our motivating case-study, of evaluating IT Wallet estimation models for guiding IBM's Sales & Marketing efforts.

## 2 Residual- and Ranking-based evaluation of regression models

In this paper we concentrate on the *model evaluation* phase of supervised learning of regression models. We do not consider at all the process of *model building*, instead assuming that we are given a model $\hat{y} = m(\mathbf{x})$ which attempts to predict a response $y \in \mathbb{R}$ as a function of a feature vector $\mathbf{x}$. Assume further we have a *test set* of size $n$: $\{\mathbf{x}_i, y_i\}_{i=1}^n$, not used in modeling, which we want to utilize to evaluate the performance of model $m$ in predicting $y$. [1]

The standard approach to evaluating regression model performance is based on *additive* measures of error, depending on the residuals $r_i = y_i - m(\mathbf{x}_i)$:

$$\text{Mean prediction error} = \frac{1}{n}\sum_{i=1}^{n} L(r_i)$$

Some commonly used error measures are:

$$\begin{array}{rl} \text{Squared error loss:} & L(r) = r^2 \\ \text{Absolute error loss:} & L(r) = |r| \\ \epsilon\text{-insensitive loss:} & L(r) = \max(0, |r| - \epsilon) \end{array}$$

Other approaches to regression model evaluation include Regression Error Curves [8, 1], where the model is evaluated according to its error rate at different levels of "error tolerance"; and using medians of the absolute deviations (MAD), rather than their mean, as the error measure:

$$\text{MAD} = \text{Median}(|r_1|, ..., |r_n|) \quad (1)$$

[1] In reality, we may often have multiple models $m_1(\mathbf{x}), ..., m_k(\mathbf{x})$, and evaluate them as a way of facilitating a *model selection* decision. We may also employ cross validation instead of pure holdout data for model evaluation. Our methods and results apply to these situations as well (and indeed, we show a model selection application in Section 5). For simplicity of exposition, we stick here with the single model and holdout data case.

In this paper we consider, instead of these residual-based evaluation measures, the use of ranking-based measures, which evaluate the performance of the scoring model $m(\mathbf{x})$ in sorting the values of $y$ from "large" to "small". We assume, without loss of generality, that the test set is sorted in decreasing order of model scores, that is:[2]

$$m(\mathbf{x}_1) > m(\mathbf{x}_2) > ... > m(\mathbf{x}_n)$$

We rank the responses in the test set in decreasing order. Let $s_i$ be the rank of observation $i$ in this order:

$$s_i = |\{j \le n \mid y_i \le y_j\}|$$

Ranking-based evaluation uses only these ranks in evaluating model performance. We consider here two ranking-based evaluation measures and their interpretations. We start by defining the following intuitive statistics:

The number of ranking order switches:

$$T = \sum_{i<j} \mathbf{1}\{s_i > s_j\} \quad (2)$$

The weighted sum of order switches:

$$R = \sum_{i<j} (j - i) \cdot \mathbf{1}\{s_i > s_j\} \quad (3)$$

The first measure simply counts how many of the pairs in the test data are ordered incorrectly by the model $m(\mathbf{x})$. The second also considers these incorrect orderings, but weighs them by the difference in their model ranks, a measure of the magnitude of error being committed.

Next, we linearly transform these two measures to put them into the range $[-1, 1]$, where 1 corresponds to perfect model performance ($T, R = 0$) and $-1$ corresponds to making all possible errors, thus attaining perfect reverse ranking. It is easy to verify that $\max(T) = n(n-1)/2$, $\max(R) = n(n-1)(n+1)/6$. Thus we re-scale:

$$\hat{\tau} = 1 - \frac{4T}{n(n-1)} \quad (4)$$

$$\hat{\rho} = 1 - \frac{12R}{n(n-1)(n+1)} \quad (5)$$

This in fact gives us exactly Kendall's $\hat{\tau}$ and Spearman's $\hat{\rho}$ [6] — the measures for non-parametric correlation prevalent in Statistical data analysis tools. We use here the notation $\hat{\tau}$ and $\hat{\rho}$, to denote that these are the sample quantities (which [6] denote by $t, r$ respectively), as compared to the "population" quantities: $\tau = E\hat{\tau}$ , $\rho = E\hat{\rho}$.

In what follows, we will use, in addition to $\hat{\tau}$ and $\hat{\rho}$, their re-normalized versions: $\tilde{\tau} = 1/2 + \hat{\tau}/2$ and $\tilde{\rho} = 1/2 + \hat{\rho}/2$. Both of these are in $[0, 1]$ with 1 corresponding to perfect ranking and 0 to inverse ranking.

[2] We assume there are no ties in either the responses or the model scores. This is not critical but makes the presentation and discussion of ranking-based methods simpler.



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

## 3 Interpretations and visualizations

In this section, we present some of the interpretations and visualization approaches that we can apply to our ranking evaluation measures $\hat{\tau}$, $\hat{\rho}$.

### 3.1 Statistical properties of $\hat{\tau}$

Of the two measures we suggest, $\hat{\tau}$ is the more natural target for a statistical analysis. First, we observe that $\tilde{\tau}$ can be interpreted as the percentage of all pairs that are correctly ranked from the total of $n(n-1)/2$ pairs of observations [6]. Thus, $E\tilde{\tau}$ is the *probability* of correctly ranking a randomly drawn pair of observations. This gives $\tilde{\tau}$ a probabilistic interpretation, in the same spirit as the area under the ROC curve [2].

Second, we are interested the distribution of $\hat{\tau}$ and $\hat{\rho}$ and the uncertainty inherent to them. The moments of $\hat{\tau}$ and $\hat{\rho}$ under the relevant null assumptions ($\tau = 0$ and $\rho = 0$, respectively) are quite easy to calculate and a normal approximation gives a hypothesis testing methodology for the assumption of no correlation [6]. However, testing this assumption — that the ranking by scores and by actual test set responses are independent — is of little interest in most cases for model evaluation. We expect any reasonable prediction model to create rankings that are indeed correlated with the rankings by response.

Of much more practical interest are confidence intervals for the values of evaluation measure given its sample value (the non-null case), as they represent the uncertainty in the model evaluation based on a single test set. For residual-based measures, it is typically not possible to build confidence intervals without parametric assumptions and/or variance estimation. The non-parametric nature of $\hat{\tau}$ allows us to write a general expression for its variance [7]:

$$Var(\hat{\tau}) = \frac{8}{n(n-1)}\left(\pi_c(1-\pi_c) + 2(n-2)(\pi_{cc} - \pi_c^2)\right)$$

where $\pi_c = E\tilde{\tau} = 1/2 + 1/2\tau$ and $\pi_{cc}$ are two properties of the ranking function (we omit here the exact complicated definition of $\pi_{cc}$). We can replace these with their sample means, and after some algebra obtain [7]:

$$\hat{Var}(\hat{\tau}) = \left(\frac{2}{n(n-1)}\right)^2 \cdot 2 \cdot \cdot \left(2\sum_i C_i^2 - \sum_i C_i - \frac{(2n-3)}{n(n-1)}(\sum_i C_i)^2\right) \quad (6)$$

where $C_i = \sum_{j<i} \mathbf{1}\{y_i > y_j\} + \sum_{j>i} \mathbf{1}\{y_i < y_j\}$ is the number of observations that are "concordant" with observation $i$, that is, that their ranking relative to $i$ in the test data agrees with the ranking by model scores (as plotted in Figure 1 below).

Since $\hat{\tau}$ is asymptotically normal [6], we can obtain an approximate, consistent $1-\alpha$ confidence interval for $\tau$ as:

$$\hat{\tau} \pm Z_{1-\alpha/2} \cdot \sqrt{\hat{Var}(\hat{\tau})}$$

We are not aware of a similar calculation for $\hat{\rho}$.

### 3.2 Visualizations

Visualizations of ranking performance can often provide additional insights about model performances. We will use the data from the case study in Section 5 to illustrate the different visualization approaches. This dataset was collected from a survey of 500 firms reporting the amount of money allocated in 2003 to the purchase of IT goods. The 500 observations are considered the ground truth against which we wish to evaluate an IBM internal model.

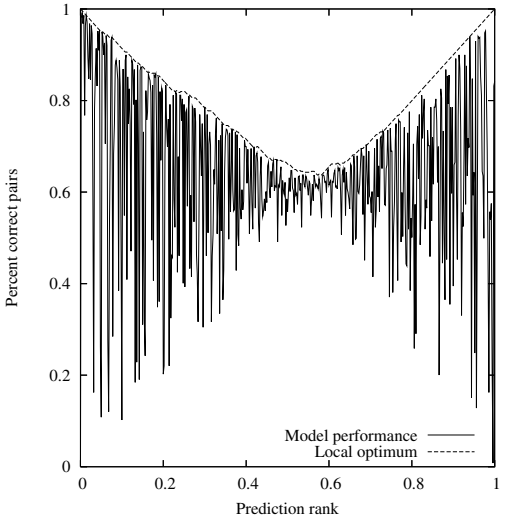


**Figure 1. Percent of correctly ranked pairs involving a particular prediction.**

Starting from the largest model prediction $m(\mathbf{x}_1)$ we calculate in decreasing order for each observation $\mathbf{x}_i$ the percentage of correctly ranked pairs $(y_i, y_j)$ over all $j \neq i$. Figure 1 shows this percentage of correctly ranked pairs as a function of the rank. The area above the curve is the sum of the percent incorrectly ranked pairs, which is equal to $2 \cdot T/n(n-1)$. Therefore, the area under the curve equals $\tilde{\tau}$. The dashed line corresponds to a *locally* optimal performance. A perfect model would show a constant performance of 100% correctly ranked pairs. But given that the model is not perfect and makes predictions that are sometimes too large or too small, even a perfect prediction for a particular observation with $m(\mathbf{x}_i) = y_i$ will have a number



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

of inversely ranked pairs due to errors of the other predictions. The upper limit of the performance for a given prediction, keeping everything else constant, is therefore not 100% but determined by the model performance of predictions around it. The interpretation of the locally optimal performance is therefore the highest achievable percentage of correctly ranked pairs if $m(\mathbf{x}_i)$ could be placed arbitrarily, given all other model predictions.

The performance in a particular region of the graph is characterized by two properties of the plot, 1) the distance of the upper envelope from the 100% line, and 2) the distance of the actual performance from the upper envelope. A particular region with a performance that on average remains very close to the upper envelope has a nearly optimal local ranking and is only disturbed by bad predictions that were either larger or small than the predictions of the region.

In summary, this graph provides a number of relevant insights for model evaluation and analysis:

- It shows the variability of the ranking performance for different prediction regions (large versus small predictions) in the vertical width of the performance band (average distance from the upper envelope). In the shown example, the model ranking of the top 3% cases on the left side of the graph appears to be very good, whereas the ranking of the small predictions is comparatively bad.
- It shows large outliers with less than 50 percent correct pairs. These predictions are very strongly out of place in terms of their ranking and may deserve some special analysis.

The next two graphs are related to $\tilde{\rho}$. For Figure 2 we transform the original regression results into $n-1$ classification results, where we discretize the observations into a binary class variable $c^{(i)}(y_j) = 1$ iff $y_j \geq y_i$ for all possible cutoffs $1 < i < n$.

For each classification $\mathbf{c}^{(i)}$ we can evaluate the model performance using the area under the ROC curve (AUC, [2]). The probabilistic interpretation of the AUC is the probability that a pair of observations with opposite class labels is ranked correctly. Since $\text{AUC}_i$ only considers pairs with different class labels under cutoff $i$, the number of pairs used to calculate $\text{AUC}_i$ is equal to $i \cdot (n-i)$ and the number of incorrectly ranked pairs under this cutoff is therefore $(1\text{-AUC}_i) \cdot i \cdot (n-i)$.

The total number of times a pair of observations will be assigned opposite class labels across all cutoffs is equal to the rank difference: a neighboring pair $(k, k+1)$ receives opposite class labels only if the cutoff is equal to k whereas the extreme pair $(1, n)$ will have opposite classes for all $n-1$ cutoffs. Given our definition in (3) this implies that $\sum_i (1 - AUC_i) \cdot i \cdot (n-i) = R$.

Since each $\text{AUC}_i$ is rescaled by a different factor $i \cdot (n-i)$, a graph of the $\text{AUC}_i$ as a function of the cutoff $i$ would not have an area equal to $\tilde{\rho}$. In order to achieve a direct correspondence with $\tilde{\rho}$ we have to allocate $i \cdot (n-i)$ units to $\text{AUC}_i$ by rescaling the x-axis accordingly. Figure 2 shows such a transformation of the x-axis and has an area under the curve of $\tilde{\rho}$. This graph confirms our earlier notion that the model performs better in ranking large outcomes and worse in ranking small outcomes.

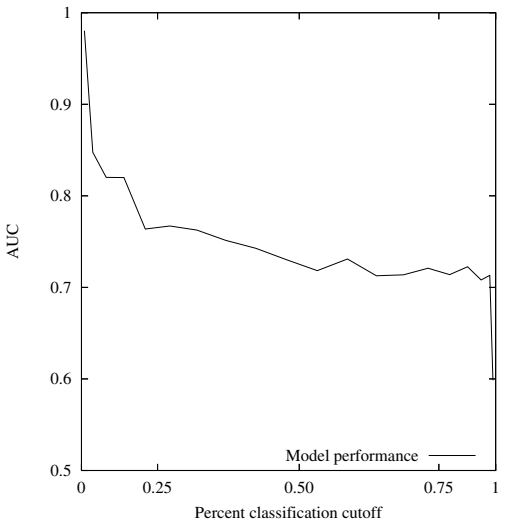


**Figure 2. AUC as a function of cutoff position above which the truth will be assigned to class 1 adjusted for the number of pairs. Note the nonlinear transformation of the x-axis.**

The plot in Figure 3 is very close in spirit to a lift curve. After sorting the model predictions in decreasing order, we plot the cumulative inverse rank of the truth $p_i = \sum_{j=1}^{i}(n - s_j + 1)$ for increasing cutoffs $i$ in percent. Using the inverse rank emphasizes the model performance on the largest predictions that is shown in the bottom left of the graph. The model performance is bounded above by the optimal ranks $p_i = \sum_{j=1}^{i}(n - j + 1)$ and below by the cumulative worst (inverse) ranking $w_i = \sum_{j=1}^{i} j$. The area under the model curve is given as $\sum_{i=1}^{n}(n-i)(n-s_i)$ and can be shown to be equal to $n^3 - n^2 - n - R + \sum_{i=1}^{n} i^2$ exposing the relationships to $\tilde{\rho}$ (see [6] for details).

## 4 Evaluation robustness and ranking-based evaluation

A commonly used definition of robustness in model fitting uses the concept of *fitting breakdown point*. A simplified version of the breakdown point definition of [5, p. 98] reads:



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

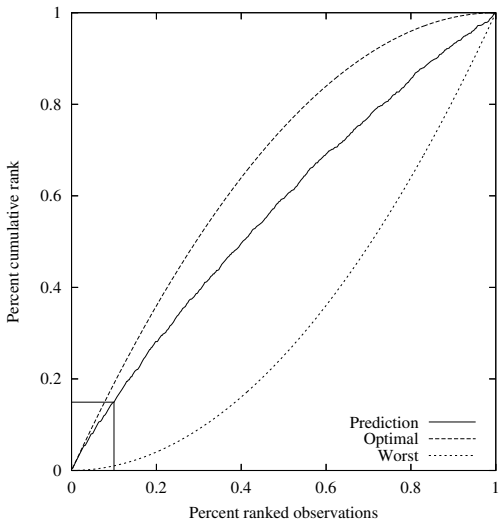


**Figure 3. Lift-curve of the cumulative rank. With about 10% of the data the model is able to capture 15% of the cumulative rank, about 4% less than the optimal.**

> The breakdown point of a fitting procedure is the % of data points that must be arbitrarily badly corrupted before the fitted model is arbitrarily badly corrupted.

Using this definition, we can show that linear regression with squared error loss has a breakdown point of $1/n$. Thus, this is a non-robust procedure — one corrupted data point can affect the fitted model arbitrarily badly. Linear regression with absolute loss, on the other hand, has a breakdown point of "almost" $1/2$. This is a robust fitting procedure, since as long as less than half of the data points are corrupted we are guaranteed to remain "reasonably close" to the uncorrupted solution (see [5] and references therein for details). As we will see below, absolute loss *is not* robust for evaluation.

We would like to define a similar notion of robustness for evaluation. We are not aware of any work in the literature on that topic. Here we suggest one such notion. Consider the following evaluation process:

**Inputs:**
Regression model: $m : \mathbb{R}^p \to \mathbb{R}$
Test set of size $n$ : $\{\mathbf{x}_i, y_i\}_{i=1}^n$
Continuous evaluation loss measure: $\mathcal{L} : \mathbb{R}^n \cdot \mathbb{R}^n \to \mathbb{R}$

**Evaluation procedure:**
Apply model to test set to get predictions: $\mathbf{m} = (m(\mathbf{x}_1), ..., m(\mathbf{x}_n))'$
Apply evaluation measure to model: $e = \mathcal{L}(\mathbf{m}, \mathbf{y})$

We are now ready to define *evaluation breakdown*

**Definition 1** *Let* $M = \sup_{\mathbf{u}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^n} \mathcal{L}(\mathbf{u}, \mathbf{v})$ *(possibly* $M = \infty$*). The* evaluation breakdown *of an evaluation metric* $\mathcal{L}$*, denoted* $\mathcal{R}(\mathcal{L})$ *is the smallest % of test data points that need to be arbitrarily corrupted to guarantee that for any* $c < M$ *and any test data we can get* $\mathcal{L}(\mathbf{m}, \mathbf{y}) > c$*.*

The following simple result is an immediate consequence of this definition:

**Proposition 1** *If* $\mathcal{L}$ *is:*

- *A non-negative, additive function of the residuals, that is* $\exists L : \mathbb{R} \cdot \mathbb{R} \to R$ *such that* $\mathcal{L}(\mathbf{m}, \mathbf{y}) = \sum_i L(y_i - m(\mathbf{x}_i))$*, with* $L(u) \geq 0$ , $\forall u \in \mathbb{R}$
- *Unbounded (that is,* $M = \infty$*)*

*Then* $\mathcal{R}(\mathcal{L}) = 1/n$ *for a test set of size* $n$*. That is, a single corrupted data point is enough to guarantee arbitrarily bad evaluation.*

**Proof** Because of the additivity, $M = \infty$ implies $\sup_{u\in\mathbb{R}} L(u) = \infty$ as well. Given a value $c < \infty$, let $u_0$ be such that $L(u_0) > c$. Now, if we choose any observation $(\mathbf{x}_k, y_k)$ and corrupt its response $y_k$ to $\tilde{y}_k = m(\mathbf{x}_k) + u_0$ we get the evaluation score:
$L(\tilde{y}_k - m(\mathbf{x}_k)) + \sum_{i\neq k} L(y_i - m(\mathbf{x}_i)) \geq L(u_0) > c$ ■

This result illustrates that squared error loss, absolute loss and any other unbounded function of the residuals are all *evaluation non-robust* and have a breakdown point of $1/n$ for evaluation. This is less trivially also true of the area over the REC curve (AOC), suggested by [1] as a one-number summary of model performance. The AOC can be bounded from below by an additive function:

$$AOC \geq \frac{1}{n} \max_i |y_i - m(\mathbf{x}_i)| \geq \frac{1}{n^2} \sum_i |y_i - m(\mathbf{x}_i)|$$

and the resulting lower bound is evaluation non-robust, which clearly leads to the AOC being evaluation non-robust as well.

On the other hand, MAD (1) is clearly a robust evaluation measure, since corrupting almost half of the data arbitrarily still guarantees that the median is in the uncorrupted half, and so is not significantly affected. It can be argued, though, that by ignoring completely the large absolute residuals, MAD is "too robust" and does not penalize a model for making gross errors. Our suggested ranking-based evaluation measures are also robust. The constant $M$ of Definition 1 is finite and to get arbitrarily close to it we clearly need to corrupt almost all our data, in some cases all of it, to make sure that the order of all pairs is incorrect. More convincingly, we can derive an $O(\frac{1}{n})$ bound on the effect of



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

any outlier on the overall model evaluation score (a property not shared by any of the residual-based methods, including MAD), as follows:

**Proposition 2** *Given a test set of size $n$, if we corrupt at most $k$ observations arbitrarily, the change in both the measures $\tilde{\tau}$ and $\tilde{\rho}$ is $O(1/n \cdot k)$.*

**Proof** Denote the original label vector by $\mathbf{y}$ and the corrupted one by $\mathbf{y}^*$. If we consider our two measures $T, R$ as defined in (2, 3), we can see that:

$$\begin{aligned} |T(\mathbf{y}^*) - T(\mathbf{y})| &\leq \sum_{i=1}^{k}(n-i) \leq k \cdot n \\ |R(\mathbf{y}^*) - R(\mathbf{y})| &\leq k(\sum_{i=1}^{n-k} s_i + n) \leq kn\left(\frac{n+3}{2}\right) \end{aligned}$$

where the second calculation uses the equivalent formulation $R = \sum_i i^2 - \sum_i i \cdot s_i$, proven in [6], and the fact that the change in the ranks of the corrupted observations is no more than $n$, while the ranks of the non-corrupted observations can be changed by at most $k$.

Plugging this into the definitions of $\tilde{\tau}, \tilde{\rho}$ we get:

$$\begin{aligned} |\tilde{\tau}(\mathbf{y}^*) - \tilde{\tau}(\mathbf{y})| &= \left|\frac{2(T(\mathbf{y}^*) - T(\mathbf{y}))}{n(n-1)}\right| \leq \frac{2 \cdot k}{n-1} \\ |\tilde{\rho}(\mathbf{y}^*) - \tilde{\rho}(\mathbf{y})| &\leq \frac{6 \cdot k}{n-1} \cdot \left(\frac{1}{n+1} + \frac{1}{2}\right) \end{aligned}$$

■

## 5 Case study: evaluation of IT Wallet estimation models at IBM

The wallet of a customer (or potential customer) is defined as the total amount that a customer spends in a certain product category in a given time frame. There are many possible uses for wallet estimates, which include targeting sales/marketing actions towards large wallet customers, detecting partial defection of customers and rewarding sales representatives according to the share of the wallet of a customer that they attain. Recent marketing literature demonstrates that knowing the customer's share of wallet is important for customer relationship management [4].

In certain industries, customer wallets can be easily obtained from public data. For example, in the credit card industry, the card issuing companies can calculate the wallet size using credit records from the major credit bureaus. For most industries, however, no public wallet information is available at the customer level. In this case, a model that relates the available information about the customer to the wallet needs to be created. A common approach is to obtain actual wallet information for a random subset of customers through a survey and build a regression model.

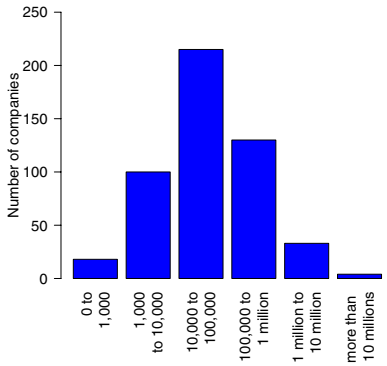


**Figure 4. IT Wallet Distribution. Note that the x-axis is on a logarithmic scale.**

At IBM, a variety of approaches have been considered for estimating the wallet of customers for information technology (IT) products, including heuristic approaches and predictive modeling. Given the variety of models, there was a pressing need for an objective comparison of their performance. For this reason, a survey was conducted to obtain actual 2003 wallet figures for a set of 500 companies that are customers or potential customers of IBM in the US, for three product categories: hardware, software and services. In this case study, we do not discuss building models for wallet estimation, but restrict ourselves to the important problem of evaluating and comparing models in a meaningful and robust fashion.

### 5.1 Advantages of ranking-based evaluation

Figure 4 shows the distribution of IT wallets obtained in the survey. Note that this distribution is long-tailed, that is, there are many companies with relatively small wallet size and a few companies with very large wallet size. Therefore, evaluation measures such as mean squared error and mean absolute error can be greatly influenced by a small subset of companies that have very large wallets and for which the models are more likely to make larger absolute errors. On the other hand, measures such as median squared error can completely ignore the performance of the model on the companies with large IT wallet size, which are usually the most important customers. An approach that is often used to mitigate the effects of a skewed distribution (especially in modeling) is to transform the numbers to a logarithmic scale. This approach, however, is not adequate for the evaluation of wallet models, since log-dollars is a unit that does not have a clear financial meaning and, therefore, cannot be used in conjunction with other financial variables such as



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

| Measure | Entire Set | Excl. One | %Change |
|---|---|---|---|
| RMSE | \$3,872,481 | \$1,940,415 | 49.8% |
| MAE | \$568,281 | \$421,198 | 25.9% |
| $\hat{\tau}$ | 0.2845 | 0.2896 | 1.80% |
| $\hat{\rho}$ | 0.4032 | 0.4115 | 2.05% |

**Table 1. The effect of the single most influential observation on different measures.**

budgets and costs.

For some practical applications, such as targeting large wallet customers, ranking the customers according to wallet size is all that is needed. To make this more concrete, let us assume that a certain budget of $B$ dollars is available for targeting customers and that the cost of targeting a customer is fixed at $c$ dollars. Then, we can target at most $k = \lfloor \frac{B}{c} \rfloor$ customers. Given that the number of customers we can target is fixed at $k$, it is easy to see that the best strategy given the wallet estimates is to target the customers with the top $k$ wallet values. Because $B$ and $c$ can vary over time, we need a complete ranking of customers to be able to threshold at any point.

Thus, we can argue that evaluating wallet models using ranking-based measures is advantageous for at least two reasons. First, as demonstrated in Section 4, they are robust, which is especially important when we have skewed distributions. Second, they evaluate the appropriate performance measure, at least for targeting applications.

### 5.2 Empirical robustness of different performance measures

To illustrate the sensitiveness of different performance measures to a single observation, we calculate the root mean squared error (RMSE), mean absolute error (MAE), Kendall's $\hat{\tau}$ and Spearman's $\hat{\rho}$ measures of the model that predicts the software wallet for the companies included in the survey. Then, we recalculate the measures excluding the single company that has the largest influence on each measure. For additive measures this is the company whose estimated wallet is the furthest from the actual wallet. For ranking measures we do an exhaustive search to determine the removal that leads to the biggest change. The results are shown in Table 1.

The percent change characterizes empirically the degree to which each measure is affected by a single company. For RMSE and MAE a single company is responsible for 49.8% and 25.9% of the total loss, respectively, while for Kendall's $\hat{\tau}$ and Spearman's $\hat{\rho}$ no company is responsible for more than 2.05% of the total correlation. These empirical results are in agreement with the theoretical results from Section 4 showing that the ranking-based measures are more robust than residual-based measures.

### 5.3 Model comparison using different performance measures

Here, we use residual-based and ranking-based performance measures to compare two existing IBM Software Wallet models, which in this study we refer to as $M_1$ and $M_2$.

In table 2 we show the RMSE, MAE, Kendall's $\hat{\tau}$ and Spearman's $\hat{\rho}$ measures of the two models for the 500 companies included in the survey. According to these results, $M_1$ is slightly outperforming $M_2$ for the residual-based measures (RMSE and MAE), where smaller is better, and it is greatly outperforming $M_2$ for the ranking-based measures ($\hat{\tau}$ and $\hat{\rho}$), where larger values are better.

A natural question to ask at this point is whether the differences in performance between these models are significant. In order to measure statistical significance, we create 100 bootstrap samples from the original survey of 500 companies. Using the bootstrap samples, we can obtain estimates of the standard deviation of the difference between the performance of the two models for each measure. Assuming a gaussian distribution, we compute p-values using these estimates of the standard deviation and the actual difference in performance from the original sample. The last three columns of Table 2 summarize these results.

From these results, we can conclude that the difference in performance between the two models is statistically significant for the ranking-based measures and not statistically significant for the residual-based measures. This is a consequence of the fact that we have a few extreme outliers in the test sample. The outliers considerably affect the residual-based measures and make them unreliable and inappropriate for model comparison. On the other hand, the ranking-based measures are much more stable and robust to the outliers, allowing us to be confident that model $M_1$ is indeed performing better than $M_2$ in terms of ranking.

Using the bootstrap samples, we can estimate the variance of $\hat{\tau}$ and compare it to the analytical expression (6). The numbers (for model $M_2$) are indeed very close: the empirical estimate is $0.00103$ and the analytical estimate is $0.00097$.

### 5.4 Visualizations

The graphs depicted in Section 3 use data from this case study, more specifically from model $M_1$. By investigating these graphs, we can make the following observations about this model:

- From Figures 1 and 2 we can see that the model is able to rank better the examples with the largest predictions



Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 \$20.00 © 2005 IEEE

| Measure | $M_1$ | $M_2$ | Absolute Difference | Std. Dev. | p-value |
|---|---|---|---|---|---|
| RMSE | \$3,775,592 | \$3,872,481 | \$96,890 | \$92,649 | 0.14778 |
| MAE | \$568,281 | \$593,406 | \$25,124 | \$33,498 | 0.22663 |
| $\hat{\tau}$ | 0.2845 | 0.1692 | 0.1154 | 0.0304 | 0.00007 |
| $\hat{\rho}$ | 0.4032 | 0.2503 | 0.1529 | 0.0426 | 0.00016 |

**Table 2. Performance measures and statistical analysis for two IT Wallet models.**

than the examples with the smallest predictions. This is to be expected because loss functions used for modeling usually emphasize on the largest potential absolute errors. For targeting, this is a good scenario, since we are interested in picking the top $k$ customers, for a certain (usually small) $k$. From Figure 3, we see that indeed the model is doing well in identifying the top customers, up to the top 10% customers.

- In Figure 1, we see that there are some examples that are completely out of place in terms of their ranking. Analyzing the common characteristics of these examples could lead to conclusions about the shortcomings of this model and possible ways to improve its ranking performance

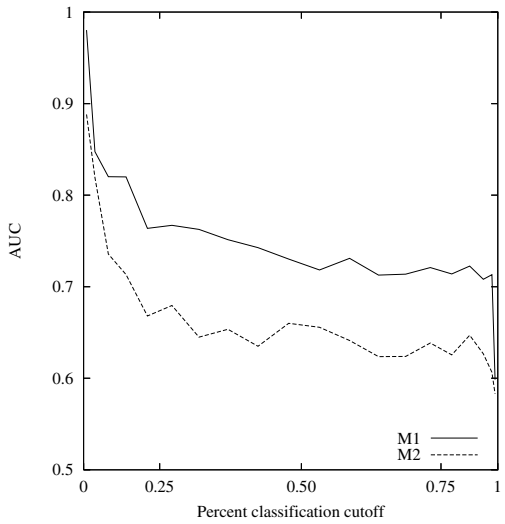


**Figure 5. Comparative model performance in terms of AUC across all cutoffs.**

Figure 5 presents the comparative AUC's as a function of cutoff position for both models. The graph shows that model $M_1$ is consistently outperforming model $M_2$ for all the cutoff points. A bootstrap analysis of the performance differences between $M_1$ and $M_2$ shows that within a 0.1 to 0.9 cutoff range $M_1$ outperforms $M_2$ more than 95% of the time. This fits in nicely with the significance results for $\hat{\rho}$ in Table 2, which implies that the area under $M_2$ is significantly smaller than under $M_1$ (given the interpretation we presented in Section 3.2).

## 6 Conclusion

We have presented ranking-based evaluation strategies for regression models that are often appropriate for marketing tasks (as shown in our case study) and are more robust to outliers than traditional residual-based performance measures. Other advantages of our evaluation approach include the proposed visualization methods. These can provide insights about local model performance and outliers. Further contributions of this work include the definition of evaluation robustness as a property of different evaluation measures, and the presentation of confidence intervals for one of the suggested ranking measures.

## References


[1] J. Bi and K. P. Bennett. Regression error characteristic curves. In *Proceedings of ICML-03*, 2003.

[2] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[3] J. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[4] R. Garland. Share of wallet's role in customer profitability. *Journal of Financial Services Marketing*, 8(3):259–268, March 2004.

[5] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley & Sons, 1986.

[6] M. Kendall and J. M. Gibbons. *Rank Correlation Methods*. Edward Arnold, 1990.

[7] N. E. Noether. *Elements of Nonparametric Statistics*. Wiley and Sons, 1967.

[8] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.




Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 \$20.00 © 2005 **IEEE**