Predictive Model Performance: Offline and Online Evaluations

Jeonghee Yi* , Ye Chen, Jie Li, Swaraj Sett, Tak W. Yan
Microsoft Corporation
1065 La Avenida st.
Mountain View, CA 94043
{jeyi, yec, lijie, swasett, takyan}@microsoft.com

ABSTRACT

We study the accuracy of evaluation metrics used to estimate the efficacy of predictive models. Offline evaluation metrics are indicators of the expected model performance on real data. However, in practice we often experience substantial discrepancy between the offline and online performance of the models.

We investigate the characteristics and behaviors of the evaluation metrics on offline and online testing both analytically and empirically by experimenting them on online advertising data from the Bing search engine. One of our findings is that some offline metrics like AUC (the Area Under the Receiver Operating Characteristic Curve) and RIG (Relative Information Gain) that summarize the model performance on the entire spectrum of operating points could be quite misleading sometimes and result in significant discrepancy in offline and online metrics. For example, for click prediction models for search advertising, errors in predictions in the very low range of predicted click scores impact the online performance much more negatively than errors in other regions. Most of the offline metrics we studied including AUC and RIG, however, are insensitive to such model behavior.

We designed a new model evaluation paradigm that simulates the online behavior of predictive models. For a set of ads selected by a new prediction model, the online user behavior is estimated from the historic user behavior in the search logs. The experimental results on click prediction model for search advertising are highly promising.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

*Contact author

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

Keywords

evaluation metric, offline evaluation, online evaluation, AUC, RIG, log-likelihood, prediction error, simulated metric, online advertising, sponsored search, click prediction

1. INTRODUCTION

In the field of machine learning, evaluation metrics are often used to judge and compare the performance of predictive models on benchmark datasets. It is quite clear that good quantitative assessments of their accuracies are essential to build successful predictive systems. Though a large array of evaluation metrics are already available [5, 13] and *de facto* standard metrics may exist for specific prediction problems, they do not come without limitations and drawbacks. Previous research has shown that some metrics may overestimate the model performances for skewed samples [9, 10, 14], and there exist variations of a metric that lead to different results under certain circumstances like cross validation [14].

For a typical machine learning problem, training and evaluation (or test) samples are selected randomly from the population the model needs to be built for, and predictive models are built on the training samples. Then the learned models are applied on the evaluation data, and the qualities of the models are measured using selected evaluation metrics. This is called *offline evaluation*.

In addition, highly complex modern applications, such as search engines like Google and Bing, and online shopping engines like Amazon and eBay, often conduct online evaluations of best performing offline models on a controlled AB testing platform (*online evaluation*). The online AB testing platform may set up two isolated testing environments that are identical except one is set up with the baseline (or control) model, and the other one with the new model to be tested. They send a predefined amount of live traffic to each environment for the same time period. The differences in online user behaviors, such as clicks and the number of searches per user, and some other performance metrics, such as revenue per search, are evaluated to determine whether the difference is statistically significant before making a final launch decision of the new model. The assumption here is that the online performance metric would be better, if the new model delivered better quality results.

One problem with the model evaluations in reality is that sometimes the improvement of model performance in offline evaluation does not get realized as much, or sometimes gets reversed in online evaluation. Unlike static offline evaluation, online testing even under the controlled environment is highly dynamic, of course, and many factors not consid-

ered during the offline modeling play a role in the results. Nevertheless, these observations raise a question if there exist fundamental biases or limitations of the offline evaluation metrics that lead to such discrepancies.

Another problem is comparing performance of predictive models built with different kinds of data, especially data with rare events. Rare events occur in disproportionately lower frequency than the counterparts, thus result in skewed sample distributions between the classes. This is a quite common phenomenon in real world problems. Examples of rare events include clicks on web search result links, clicks on display ads, and making a purchase after clicking on product ads. Previous research has shown that some metrics may overestimate the model performance for skewed samples[9]. The observations lead into the following questions. With the bias, how can we interpret and compare the model performance applied to different kinds of data ? For example, when we build prediction models for text ads and display ads, can we use the offline metrics as comparative measures to predict their true performance ? Suppose we know the true performance of a model, and we get equivalent offline metrics of the other model. Can we estimate the true performance of the other model ? If we can't, what kind of metrics should we use instead ?

We propose a new model evaluation paradigm: *simulated metrics*. We implemented auction simulation for offline simulation of online behaviors and used the simulated metrics to estimate the online model performance of click prediction models. Since simulated metrics are designed to simulate online behaviors, we expect they would suffer less from the performance discrepancy problem. Also, since simulated metrics directly estimates the online metrics such as user CTR (Click-Through Rate), they can be directly comparable even if they are for models built on different kinds of data.

The contributions of this paper are four-fold:

- We analyze the characteristics and limitations of offline evaluation metrics and share our findings about their behaviors on offline and online data.
- We share our experience on training, evaluation, and deployment of click prediction models for production online advertising system on the Bing search engine, and offer best practice guidelines for large-scale predictive model evaluation.
- To our knowledge, this is the first paper in open literature that proposes and applies simulated metrics as model evaluations paradigm.
- To our knowledge, this is again the first paper in open literature that analyzes the problems of offline evaluation metric behaviors that lead to online and offline performance discrepancy of predictive models.

The remaining parts of this paper are organized as follows. In the next section, we briefly review online advertising and binary classification error measurement. In Section 3, we survey predictive model evaluation metrics in open literature. We then review some of the metrics frequently used in the surveyed literature in Section 4. In Section 5, we describe the problems and limitations of the AUC and RIG measures on large scale click prediction models for sponsored search. In Section 6, we discuss the discrepancy of the offline and online performance of the models deployed on real-time production traffic of the Bing search engine. Finally, we summarize our findings and suggest best practice guidelines based on our analysis and lessons from real world experiences on online advertising data.

2. PRELIMINARIES

The target application of our study is online advertising. Some of the problem areas discussed in this study might be specific to the domain. In this section, we briefly review major areas of online advertising, and those who are interested may find excellent tutorials on the references provided below.

2.1 Online Advertising

Sponsored (or paid) search [11, 21, 28, 34] such as Google AdWords and Bing's Paid Search, is search advertising that shows ads alongside algorithmic search results on search engine results pages (SERPs). Sponsored search reaches out to people actively looking for information about products and services online, thus has relatively higher click-through rate (CTR) compared to other types of advertising.

Advertisers bid on keywords through a Generalized Second-Price (GSP) auction [11]. Bidders with highest *rank scores* (r) win the auction:

$$r = b \cdot p^{\alpha} \tag{1}$$

where b is a bid amount, p is estimated position-unbiased CTR, and α is a parameter, called *click investment power*. If α>1, the auction prefers ads with higher estimated CTRs, otherwise, ads with higher bids. Rank score is estimated CTR weighted by *cost per click* bid.

Ads are allocated in the descending order of estimated rank scores, and the auction winners pay price per click (a.k.a. cost per click, or CPC) for their ad impression only when people click on their ads. In a GSP auction, CPC depends on the next higher bidder's bid amount, c_i:

$$c_i = \frac{b_{i+1} \cdot p_{i+1}^{\alpha}}{p_i^{\alpha}}$$

User clicks are highly dependent on the position of the ads[7, 15]. Typically ads shown on the section above algorithmic search results (called *mainline*) get higher CTR than those shown to the right of the algorithmic results (called *side bar*). Within the same section, the higher the ad location, the more clicks it gets for the same ad.

Display ads[32] are graphical ads that appears on websites, content pages, or applications such as instant messaging, email, etc. *Contextual ads*[7], such as Google AdSence or Bing's Contextual Search are contextually optimized ads placed on publisher's sites often with customized look and feel of the publisher's site.

Accurate estimation of the probabilities of user clicks is critical for the efficiency of ad exchange [25]. The problem of estimating click probabilities has been studied extensively both for algorithmic search [24, 30, 31, 36] and for ads[6, 8, 16, 27].

2.2 Binary Classification Error Measurement

Consider a feature vector $\mathbf{x}$, and observed binary responses, $y \in \{0, 1\}$. $\mathbf{x}$ is considered as a realization of a random vector $\mathbf{X}$, and y as a Bernoulli random variable $\mathbf{Y}$. The class 1 probability $\eta = P[\mathbf{Y} = \mathbf{1}]$ is a function of $\mathbf{x}$: $\eta(\mathbf{x}) =$

$P[\mathbf{Y} = \mathbf{1}|\mathbf{X} = \mathbf{x}]$. A binary classifier predicts samples with $\eta(\mathbf{x}) > \mathbf{c}$ as class 1, where c is a parameter: otherwise, predicts as class 0.

The efficacy of the predictions is estimated using various criteria including the primary criteria such as prediction error, and surrogate criteria such as *log-loss* and *squared error loss* [4]. Primary criteria are used to estimate the class directly, and surrogate criteria are to estimate the class prediction probability. Prediction (or misclassification) error is intrinsically unstable for estimating model performance. Instead, log-loss and square error loss are often used for probability estimation and boosting, and defined as follows:.

- Log-loss:

$$L(y|p) = -log(p^y(1-p)^{1-y}) \\ = -ylog(p) - (1-y)log(1-p)$$

- Squared error loss (or quadratic loss):

$$L(y|p) = (y-p)^2 = y(1-p)^2 + (1-y)p^2$$

where p is the estimated probability of $\eta(\mathbf{x})$. The equality of squared error loss holds only for binary classifiers: i.e., $y \in \{0, 1\}$.

Log-loss is the negative log-likelihood of the Bernoulli model. Its expected value, -$\eta log(p)$ - $(1-\eta)log(1-p)$, is called *Kullback-Leibler loss*[19] or *cross-entropy*.

2.3 Experimental Data Set

Throughout the paper we show motivating examples and the analyses of the click prediction model performance on Microsoft Bing search engine. We sampled data from Bing's sponsored search logs during the time period of Jun. thru Aug., 2012. We used two sets of data: one sampled from paid search data on Bing, and another from the contextual ads on partner websites on Microsoft publisher network.

3. SURVEY OF METRICS

We studied papers from the proceedings of the International World Wide Web Conference (WWW), the ACM International conference on Web Search and Data Mining Conference (WSDM), and the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Conference (SIGKDD) in years 2011 and 2012 in the area of algorithmic search and online advertising. We manually categorized the topic areas of the papers and the evaluation metrics they used. Table 1 summarizes the results.

There are four major topical categories we found: recommendation, search, online advertising, and CTR estimation. Search and online advertising are further divided into subcategories. The count of higher level category is sum of the counts of its sub-categories.

The categories of metrics are divided into offline and online metrics. Online metrics include model performance statistics such as ad impression yield, ad coverage, and user reaction metrics, such as CTR and the length of user sessions. Offline metrics are categorized into the following six types [18, 1, 26, 35]:

- probability-based: AUC, MLE (Maximum Likelihood Estimator), *etc*
- Log Likelihood-based: RIG, cross-entropy, *etc*
- PE (prediction Error): MSE (Mean Square Error), MAE (Mean Absolute Error), RMSE (Root Mean Square Error), *etc*
- DCG-based: DCG (Discounted Cumulative Gain), NDCG (Normalized DCG), RDCG (Relative DCG), *etc*
- IR (Information Retrieval): Precision/Recall, F-measure, AP(Average precision), MAP(Mean Average Precision), RBP (Rank-Based Precision), MRR (Mean Reciprocal Rank), *etc*
- misc: everything else that does not belong to one of the other categories

NDCG is a *de facto* standard metric of choice for search ranking algorithms. Even though probability based metrics are relatively popular for advertising domain, there still doesn't exist a single metric that dominates the domain like NDCG for search ranking problems,. Despite previous research that suggest AUC is much more reliable[3, 29, 22], there were only 2 papers we found that measured AUC. We applied AUC on the click prediction (*pClick*) problem on advertising domain, and found that it was one of the most reliable metrics, but not without problems. We will discuss about individual metrics in detail in the next section.

4. EVALUATION METRICS

We focus our review on the metrics primarily used for click prediction problems. A click prediction model estimates position-unbiased CTRs of ads for the given query. We treat it as a binary classification problem.

We exclude NDCG from our review because it is designed to prefer a ranking algorithm that places more relevant results at earlier ranks. As discussed in section 2.1, in search advertising, the ranks are determined not by the *pClick* (i.e., the estimated click) scores, but by the rank scores. Therefore, measuring the performance of *pClick* by the rank orders using NDCG is inappropriate.

We also exclude Precision-Recall (PR) analysis on our review because there is a connection between PR curve and ROC (Receiver Operator Characteristic) curve, thus a connection between PR curve and AUC [9]. Davis and Goadrich show that a curve dominates in ROC space if and only if it dominates in PR space [9].

4.1 AUC

Consider a binary classifier that produces the probability of an event, *p*. p and 1-p, the probability the event does not occur, represent the degree to which each case is a member of one of the two events. A threshold is necessary in order to predict the class membership. AUC, or the *Area under the ROC (Receiver Operating Characteristic) Curve*[12, 33], provides a discriminative measure across all possible range of thresholds applied to the classifier.

Comparing the probabilities involves the computation of four different fractions in a confusion matrix: the true positive rate (TPR) or *sensitivity*, the true negative rate (TNR) or *specificity*, the false positive rate (FPR) or *commission errors*, and false negative rate (FNR) or *omission errors*. These four scores and other measures of accuracy derived from the confusion matrix such as *precision*, *recall*, or *accuracy* all depend on the threshold.

Table 1: A summary of evaluation metrics used by papers accepted to the WWW, the ACM WSDM, and the ACM SIGKDD conferences in years 2011 and 2012 in the area of algorithmic search and online advertising.

	Offline Metrics						Online	Total
	Probability	Log Likelihood	PE	NDCG	IR	misc		
Recommendation	1	1	2	3	3	1		11
Search	1	2		10	16	4	1	34
ranking	1	1		8	7	1	1	
personalized search				1	4			
social search ranking						1		
result clustering					4	1		
query classification/suggestion					1	1		
topic assignment		1						
distributed search				1				
Online Advertising	6	1	2		1	6	2	18
pricing/bid estimation	1		1			1	1	
ad auction						2		
bid agents			1				1	
targetting	3					3		
commerce	2	1		1	2			
CTR Estimation (Algo + Ads)		3	1	2			1	7
Total	8	7	5	15	20	11	4	70

The ROC curve is a graphical depiction of *sensitivity* (or TPR) as a function of commission error (or FPR) of a binary classifier as its threshold varies. AUC is computed as follows:

- sort records with descending order of the model predicted scores
- calculate TPR and FPR for each predicted value
- plot ROC curve
- Calculate the AUC using trapezoid approximation

Empirically, AUC is a good and reliable indicator of the predictive power of any scoring model. For sponsored search, AUC, especially AUC measured only on mainline ads, is one of the most reliable indicators of the predictive power of the models. A good model (AUC>0.8) usually has statistically significant improvement if AUC improves by 1 point (0.01).

The benefits of using the AUC for predictive modeling include:

- AUC provides a single-number discrimination score summarizing overall model performance over all possible range of thresholds. This enables avoiding the subjectivity in the threshold selection.
- It is applicable to any predictive model with scoring function.
- The AUC score is bounded between [0,1] with the score of 0.5 for random predictions, and 1 for perfect predictions.
- AUC can be used for both offline and online monitoring of predictive models.

4.2 RIG

RIG (*Relative Information Gain*) is a linear transformation of log-loss [15, 36]:

$$\begin{aligned} RIG &= 1 - \frac{log\ loss}{Entropy(\gamma)} \\ &= 1 - \frac{-c \cdot log(p) - (1-c)log(1-p)}{-\gamma \cdot log(\gamma) - (1-\gamma)log(1-\gamma)} \end{aligned} \quad (2)$$

where c and p represent observed click and $pClick$, respectively. γ represents the CTR of the evaluation data.

Log-loss represents the expected probability of click. Minimizing log-loss means that $pClick$ should converge to the expected click rate and the RIG score increases.

4.3 MSE

MSE (Mean Squared Error) measures the average of squared loss:

$$MSE(P) = \frac{\sum_{i=1}^{n}(c_i \cdot (1-p_i)^2 + (1-c_i) \cdot p_i^2)}{n}$$

where p_i and c_i are $pClick$ and the observed click, respectively, of sample i.

NMSE (Normalized MSE) is MSE normalized by CTR, γ:

$$NMSE(P) = \frac{MSE(P)}{\gamma \cdot (1-\gamma)}$$

4.4 MAE

Mean Absolute Error (MAE) is given by:

$$MAE(P) = \frac{1}{n}\sum_{i=1}^{n} e_i$$

where $e_i = \mid p_i - c_i \mid$ is an absolute error.

MAE weighs the distance between the prediction and observation equally regardless of the distance to the critical operating points. MAE is commonly used to measure forecast error in time series analysis.

Empirically it also has a good performance on estimating the *pClick* model efficacy for sponsored search. It is one of the most reliable metrics together with AUC.

4.5 Prediction Error

Prediction Error (PE) measures average *pClick* normalized by CTR:

$$PE(P) = \frac{avg(p)}{\gamma} - 1$$

PE becomes zero where the average *pClick* score exactly estimates the CTR. On the other hand, PE could be still very close to zero even when the estimated *pClick* scores are quite inaccurate with mix of under- and over-estimation of the probability as long as the average is quite similar to the underlying CTR. This makes prediction error quite unstable, and it can not be used to estimate the classification accuracy reliably.

4.6 Simulated Metric

Although online experiments on controlled AB testing environment provides the real performance metrics of models under comparison by user engagement, AB testing environments are pre-set with a fixed set of parameter values, thus the model performance metrics on the testing enviroment is only for the given set of operating points. Conducting online experiments over numerous sets of operating points is not practical because online experiment is not only very time consuming, but also could be very expensive in terms of both user experience and revenue, if the new model underperforms.

Instead of using expensive and time consuming online evaluation, the performance of a model over the entire span of feasible operating points can be simulated using the historic online user engagement data. Kumar, *et. al.* developed an online performance simulation methods for federated search [20].

Auction simulation, first, reruns ad auctions offline for the given query and selects a set of ads based on new model prediction scores and/or various sets of operating points.

We implemented auction simulation [15] using sponsored search click logs data and produced various simulated metrics. Auction simulation, first, reruns ad auctions offline for the given query and selects a set of ads based on the new model prediction scores. During the simulation, user clicks are estimated using historic user clicks of the given (query, ad) pair available in the logs as follows:

- If the user click data of the (query, ad) pair is found in the logs at the same ad diaplay location (call it *ad-position*) as the simulated ad-position, the historic CTR is directly used as the expected CTR.

- If the (query, ad) pair is found in the logs, but the simulated ad-position is different from the position in the logs, the expected CTR is calibrated by the position-biased historic CTR (or *click curve*). Typically, mainline ads get drastically higher CTR than sidebar ads[1] for the same (query, ad) pair, and ads at a higher location within the same ad block gets higher CTR for the same (query, ad) pair [7].

- If the predicted (query, ad) pair does not appear in the historic logs, the average CTR (called *reference CTR*) of ads on the ad-position is used.

Click curve and reference CTR are derived from the historic user responses in the search advertising logs.

Empirically, auction simulation produces highly accurate set of ads selected by the new model for the given set of operating points. Simulated metric often turns out to be one of the strongest offline estimators of online model performance.

[1]Sidebar ads are ads shown on the ad block at the right side of algorithmic search results.

5. EXPERIENCES WITH THE METRICS ON REAL-WORLD PROBLEMS

In this section we analyze the behaviors, limitations and drawbacks of various metrics in detail in the context of click prediction for search advertising. Note that we do not mean to suggest these metrics be dismissed all together due to the limitations and drawbacks. We rather suggest the metrics be carefully applied and interpreted, especially on the circumstances where the metrics may produce misleading estimations.

5.1 AUC

While AUC is a quite reliable method to assess the performance of predictive models, it still suffers from drawbacks under certain conditions of sample data. The assumption that AUC is a sufficient test metric of model performance needs to be re-examined [23].

First, it ignores the predicted probability values. This makes it insensitive to the transformation of the predicted probabilities that preserve their ranks. On one hand, this could be an advantage as it enables comparing tests that yield numerical results on different measurement scales. On the other hand it also is quite possible for two tests to produce dramatically different prediction output, but with similar AUC scores. It is possible that a poorly fitted model (overestimating or underestimating all the predictions) has a good discrimination power [17], while a well-fitted model has poor discrimination if probabilities for presences are only moderately higher than those for absences, for example.

Table 2 shows an example of a poorly fitted model that has even higher AUC score where a large number of negative samples have very low *pClick* scores, thus lower CTR. This has an effect of lowering the FPR in the relatively higher range of *pClick* scores, thus raising the AUC score.

Second, it summarizes the test performance over the entire spectrum of the ROC space including the area one would rarely operate on. For example, for sponsored search, placing an ad in mainline impacts the CTR significantly, while it is not as much of a concern how the predicted CTR fits to the actual CTR once it is shown on mainline or where it is not shown at all. In other words, the extreme right and left side of the ROC space are generally less useful. Baker and Pinsky proposed *partial ROC curves* as an alternative to entire ROC curves[2].

It has been observed that higher AUC does not necessarily mean better ranking always. As shown in Table 3, changes in the sample distribution on either end of FPR impacts the AUC score quite substantially. Nevertheless the impact on the performance of the model in terms of CTR could be the same especially at the practical operating point of

Table 2: The AUC Anomaly 1: A poorly fitted model has even higher AUC in the presence of a large number of negative samples concentrated on the low end of *pClick* score range. (The first table shows a betterl-fitted model.)

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	300	9,700	0.030000	0.2500	0.0086	0.0011	**1.0**
0.020000	200	9,800	0.020000	0.4167	0.0173	0.0029	**1.0**
0.010000	100	9,900	0.010000	0.5000	0.0260	0.0040	**1.0**
0.005000	500	99,500	0.005000	0.9167	0.1142	0.0624	**1.0**
0.000100	100	999,900	0.000100	1.0000	1.0000	0.8499	**1.0**
total	1,200	**1,128,800**			**AUC**	**0.9193**	

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	300	9,700	0.030000	0.2500	0.0010	0.0001	**1.0**
0.0200 00	200	9,800	0.020000	0.4167	0.0019	0.0003	**1.0**
0.010000	100	9,900	0.010000	0.5000	0.0029	0.0004	**1.0**
0.005000	500	99,500	0.005000	0.9167	0.0127	0.0070	**1.0**
0.000100	100	9,999,000	0.000010	1.0000	1.0000	0.9461	**10.0**
total	1,200	**10,127,900**			**AUC**	**0.9540**	

Table 3: The AUC Anomaly 2: Changes in the sample distribution on either end of FPR impacts the AUC score quite subtantially, although the actual model performance is quite similar at the practical operationg point.

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	3,000	97,000	0.030000	0.4545	0.0093	0.0021	**1.0**
0.020000	2,000	98,000	0.020000	0.7576	0.0188	0.0057	**1.0**
0.010000	1,000	99,000	0.010000	0.9091	0.0283	0.0079	**1.0**
0.005000	500	99,500	0.005000	0.9848	0.0379	0.0091	**1.0**
0.000010	100	9,999,900	0.000010	1.0000	1.0000	0.9548	**1.0**
total	6,600	**10,392,500**	0.000635		**AUC**	**0.9797**	

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	3,000	97,000	0.030000	0.4545	0.0093	0.0021	**1.0**
0.020000	2,000	98,000	0.020000	0.7576	0.0188	0.0057	**1.0**
0.010000	1,000	99,000	0.010000	0.9091	0.0283	0.0079	**1.0**
0.005000	100	9,999,900	0.000010	0.9242	0.9904	0.8820	**500.0**
0.000010	500	99,500	0.005000	1.0000	1.0000	0.0092	**0.0**
total	6,600	**10,392,500**	0.000635		**AUC**	**0.9069**	

Table 4: The AUC Anomaly 3: A poorly fitted model has the same AUC as a well-fitted model. (The first table shows a better-fitted model.)

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	300	9,700	0.030000	0.2500	0.0086	0.0011	**1.0**
0.020000	200	9,800	0.020000	0.4167	0.0173	0.0029	**1.0**
0.010000	100	9,900	0.010000	0.5000	0.0260	0.0040	**1.0**
0.005000	500	99,500	0.005000	0.9167	0.1142	0.0624	**1.0**
0.000100	100	999,900	0.000100	1.0000	1.0000	0.8499	**1.0**
total	1,200	**1,128,800**			**AUC**	**0.9193**	

Avg pClick	# clicks	# no-clicks	Actual CTR	TPR	FPR	Trapezoid	$\frac{Avg\ pClick}{ActualCTR}$
0.030000	300	97,000	0.003083	0.2500	0.0086	0.0011	**9.7**
0.020000	200	98,000	0.002037	0.4167	0.0173	0.0029	**9.8**
0.010000	100	99,000	0.001009	0.5000	0.0260	0.0040	**9.9**
0.005000	500	995,000	0.000502	0.9167	0.1142	0.0624	**10.0**
0.000100	100	9,999,000	0.000010	1.0000	1.0000	0.8499	**10.0**
total	1,200	**11,288,000**			**AUC**	**0.9193**	

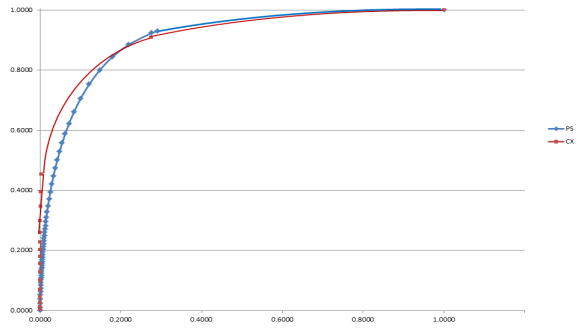

Figure 1: ROC curves of sponsored search and contextual ads

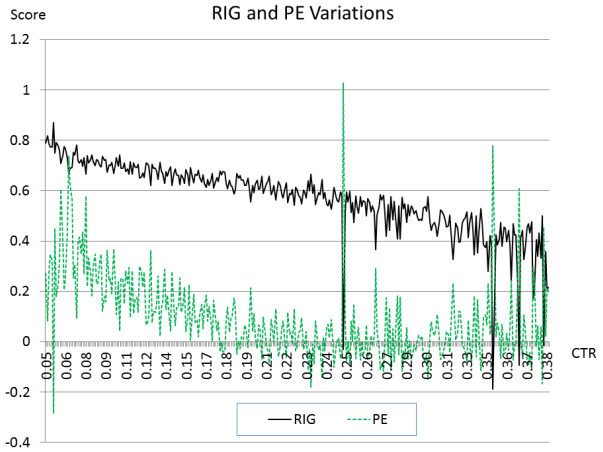

Figure 2: The RIG and PE scores over varying CTR of sample data: The RIG scores drop with increasing CTR.

the threshold. Because AUC does not discriminate the various regions of the ROC space, a model may be trained to maximize the AUC score just by optimizing the model performance on the either end of data. This may lead to lower than expected performance gain on the real online traffic.

Third, it weighs omission and commission errors equally. For example, in the context of sponsored search, the penalty of not placing the optimal ads in mainline (omission error) far exceeds the penalty of placing a sub-optimal ads (commission error). When the misclassification cost are unequal, summarizing over all possible threshold values is flawed.

Lastly, AUC is highly dependent on the underlying distribution of data. The AUC measures computed for two datasets with different rate of negative samples would be quite different. See Table 4. A poorly fitted model with lower intrinsic CTR has the same AUC as a well-fitted model. This also implies that higher AUC score for a model trained with higher rate of negative samples does not necessarily imply the model has better predictive performance. Figure 1 plots the ROC curves of *pClick* models for sponsored search and contextual ads. As indicated on the figure the AUC score of contextual ads model is about 3% higher than AUC of sponsored search, even though the former is less accurate: $\frac{avg\ pClick}{actual\ CTR} = 1.02$ for sponsored search vs 0.86 for contextual ads.

5.2 RIG

One problem with RIG is, like AUC, it also is highly sensitive to the underlying distribution of evaluation data. Since the range of the RIG scores of evaluation data vary quite widely depending on the data distribution, one may not be able to judge how good a prediction model is just by having the RIG scores.

Figure 2 illustrates how RIG (solid curve) and PE (dotted curve) varies over a typical CTR range of interest. We observe the RIG scores drop as the CTR of the dataset increases even with the same prediction model. The prediction error plotted on Figure 2 roughly indicates how close the prediction score is to the true CTR. As expected, the click prediction error is higher with the low *pClick* score range.

This behavior coincides with our earlier observations on various click prediction data sets with varying level of the intrinsic CTR. The observations suggest the followings in practice:

Table 5: Offline and online metrics of a new model (model-2) compared to the baseline model.

	offline metrics		Online metrics	
	AUC	RIG	CY	Mainline CY
model-2 metrics	8.6%	19.5%	-9.96%	-8.07%

- One should not use the face value of the RIG scores directly to compare two prediction model performances if the scores are from multiple data sets with different distribution.
- The RIG scores can be used to compare the relative performance of multiple models trained and tested on the same data.
- A RIG score in isolation is not informative enough to estimate the performance of the prediction model, as the score not only depends on the quality of the model performance, but also is heavily biased by the data distribution.

6. OFFLINE AND ONLINE PERFORMANCE DISCREPANCY

A more significant problem with the offline evaluation metrics in practice is the discrepancy in performance between the offline and online testing. There are cases where a predictive model that achieved significant gain on offline metrics does not perform as well or sometimes even underperform when deployed on the online testing environment.

Table 5 summarizes offline and online metrics of a click prediction model built with sponsored search data from the Bing search engine, and tested on online AB testing environments on Bing with real-time user traffic. Click yield (CY) is a metric of online user clicks that measures the number of clicks on ads per search page views. Mainline CY is the number of clicks on mainline ads per search page views. The new model experienced significant drop in user clicks over the baseline model on online environment even though both

Table 6: Simulated metric of a new model (model-2) compared to the baseline model.

	offline metrics		simulated metrics		online metrics	
	AUC	RIG	CY	Mainlin CY	CY	Mainline CY
model-2 metrics	0.2%	78%	40%	46%	39%	44%

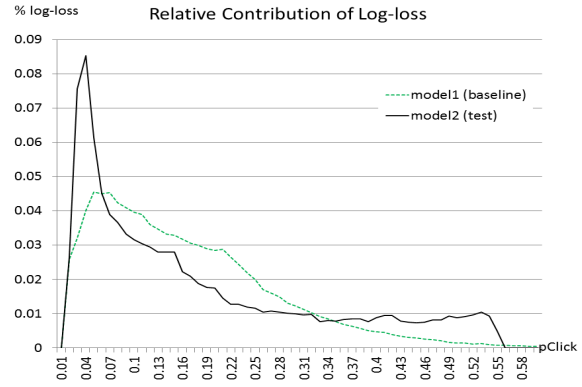

Figure 3: Relative contribution of log-loss over the typical range of *pClick* of interest.

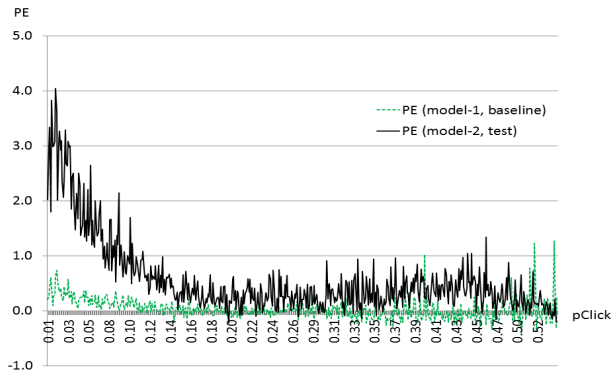

Figure 4: pClick prediction error(PE)

AUC and RIG exhibited significant gain on offline evaluation data.

Figure 3 compares log-loss [4] of two click prediction models (model-1 for baseline and model-2 for test) of each quantile within the typical range of *pClick* scores of interest. Model-2 substantially overestimates the *pClick* scores on the quantiles in lower *pClick* score range, and for the quantiles with higher *pClick* scores with much less degree of over-estimation. Figure 4 plots the prediction error of the same data with similar pattern.

Over-estimation of click probability on higher range of *pClick* scores, in practice, makes less impact on online performance than over-estimation on low *pClick* score range, because ads in the high *pClick* score range would have been most likely selected by either model. And once shown to the user, user clicks are mostly determined by the ad-position and relevances of the ads, rather than the assigned *pClick* scores.

On the other hand, over-estimation of *pClick* scores on the low *pClick* range could make significant negative impact on online metrics by giving low quality ads higher chance to be selected compared to the base model. The lower quality ads selected due to over-estimated *pClick* scores would result in lower rate of user clicks, thus hurting the online metric.

Most of the offline metrics including RIG and AUC are not able to capture thess behaviors, as the metrics cumulates the impact throughout the entire range of *pClick* scores.

6.1 Simulated Metrics

We computed the simulated metric by auction simulation as described in Section 4.6. The experimental results of the simulated click metrics along with the offline and online metrics are summarized in Table 6. We first trained a new model and optimized parameter settings that offer the best expected user click metric by auction simulation based on historic logs data. The click metrics with the best performing operating points of the models are reported as simulated metric in the table. We then set up AB testing environments with the best settings and ran the online AB testing experiments to get the online metrics. You can see that the online metrics highly coincides with the simulated metrics, while the improvements on AUC and RIG metrics differ drastically.

7. SUMMARY AND DISCUSSIONS

We reviewed and investigated the behaviors of various offline metrics for predictive models, especially in the context of click prediction for search advertising. To summarize:

- Simulated metrics are one of the most reliable metrics in predicting online performance of click prediction models. And the simulation of online behaviors is quite useful for various tasks including performance estimation and auction optimization.
- For click prediction models, AUC estimates model efficacy better than other offline metrics. Especially, AUC measured only on mainline ads are most reliable for search advertising. Nevertheless, AUC alone is not sufficient enough to estimate the model performance reliably.
- Both RIG and the AUC are highly sensitive to the class distribution of the evaluation data.
- Cross comparison of model performance by the AUC or RIG scores may be misleading if the class distributions of evaluation data are different.
- It is suggested to measure model performance in various quantiles, and carefully analyze how the change of model behavior over the range of quantiles would impact in the online environment. One may review various metrics together to discover any mismatch in the results, which may suggest some problems in the metrics.

8. REFERENCES

[1] A. Ashkan and C. L.A. Alarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proc. of the WWW Conference*, pages 407–416, 2011.

[2] S. G. Baker and P. F. Pinsky. A proposed design and analysis for comparing digital and analog mammography: social receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association*, 96(454):421–428, 2001.

[3] J. R. Beck and E.K. Shultz. The use of relative operating characteristic (roc) curves in test performance evaluation. *Archive of Pathological Lab Medicine*, 110(10):13–20, Oct. 1986.

[4] A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation: Structure and applications. technical report, 2003.

[5] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *ACM SIGKDD Conference*, pages 69–78, 2004.

[6] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *Proc. of the WWW Conference*, 2008.

[7] Y. Chen, P. Berkhin, J. Li, S. Wan, and T. W. Yan. Fast and cost-efficient bid estimation for contextual ads. In *Proc. of the WWW Conference*, 2012.

[8] Y. Chen, D. Pavlov, M. Kapralov, and J. F. Canny. Factor modeling for advertisement targeting. In *Proc. of NIPS*, 2009.

[9] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proc. of the 23'rd ICML Conference*, pages 233–240, 2006.

[10] C. Drummond and R. Holte. Explicitly representing expected cost: an alternative to roc representation. In *ACM SIGKDD Conference*, pages 198–207, 2000.

[11] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

[12] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers, 2003.

[13] C. Ferry, J. Hernandez-Orallo, and R. Modriou. An empirical comparison of performanc measures for classification. *Pattern Recognition Letters*, 30:27–38, 2009.

[14] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12:49–57, June 2010.

[15] T. Graepel, J.Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate eprediction for sponsored search advertising in microsoft's bing search engine. In *Proc. of the 27'th ICML Conference*, 2010.

[16] N. Gupta, U. Khurana, T. Lee, and S. Nawathe. Optimizing display advertisements based on historic user trails. In *Proc. of SIGIR Workshop on Internet Advertising*.

[17] D.W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley-Interscience Publication, 2002.

[18] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proc. of the ACM SIGIR Conference*, pages 41–48, 2000.

[19] S. Kullback and R.A. Leibler. On information and sufficiency. *The annals of Mathematical Statistics*, 22(1):79–86, 1951.

[20] A. Kumar, K. Pattabiraman, D. Brand, and T. Kanungo. Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. In *Proc. of the WWW Conference*, pages 67–76, 2011.

[21] S. Lahaie, D. M. Pennock, A. Saberi, and R. V. Vohra. Sponsored search auctions. *Algorithmic Game Theory*, 2007.

[22] P. Langley. Crafting papers on machine learning. In *Proc. of the 17'th ICML Conferende*, pages 1207–1212, 2000.

[23] J. M. Lobo, A. Jimenez-Valverde, and R. Real. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008.

[24] V. Murdock M. Ciaramita and V. Plachouras. Online learning from click data for sponsored search. In *Proc. of the WWW Conference*, 2008.

[25] R. McAfee. The design of advertising exchanges. *Review of Industrial Organization*, pages 1–17, 2007.

[26] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. on Information Systems*, 27(1):1–27, 2008.

[27] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *Proc. of the Int'l Workshop on Data Mining for Online Advertising and Internet Economy*, 2007.

[28] P. Papadimitriou and H. Garcia-Molina. Sponsored search auctions with conflict constraints. In *Proc. of the ACM WSDM Conference*, pages 4–14, 2012.

[29] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. of the 15'th ICML Conference*, pages 445–453, 1998.

[30] M. Regelson and D. C. Fain. Predicting click-through rate using keyword clusters. In *Proc. of the ACM Electronic Commerce Conference*, 2007.

[31] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proc. of the WWW Conference*, 2007.

[32] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proc. of the ACM WSDM Conference*, 2012.

[33] John A. Swets. Measuring the accuracy of diagnostic systems. *science*, 240(4857):1285–93, 2008.

[34] H. R. Varian. Position auctions. *Int'l Journal of Industrial Organization*, 25(6):1163–1178, 2007.

[35] E. M. Voorhees. The trec-8 question answering track report. In *Proc. of the TREC Conference*, 1999.

[36] C. Xiong, T. Wang, W. Ding, Y. Shen, and T. Liu. Relational click prediction for sponsored search. In *Proc. of the ACM WSDM Conference*, pages 493–502, 2012.