Ranking and Calibrating Click-Attributed Purchases in Performance Display Advertising

Conference Paper · July 2017

DOI: 10.1145/3124749.3124755

3 authors, including:

Sougata Chaudhuri
University of Michigan
15 PUBLICATIONS **14** CITATIONS

SEE PROFILE

Ranking and Calibrating Click-Attributed Purchases in Performance Display Advertising

Sougata Chaudhuri
A9.com, Palo Alto
sougata@a9.com

Abraham Bagherjeiran
A9.com, Palo Alto
abagher@a9.com

James Liu
A9.com, Palo Alto
jameschl@a9.com

ABSTRACT

In performance display advertising, bidders compete on behalf of advertisers for ad impressions, that is, the opportunity to display relevant ads on a publisher website. We consider bidding on behalf of online retailers who buy ad impressions hoping to realize value only from purchases attributed from clicks. The bidder has a two stage problem. In the first stage, the bidder has to select a small subset from a large selection of ads, with the selected ads most likely to lead to purchases. In the second stage, the bidder has to estimate the purchase probability of the selected ads, which can then be used to create bid values. The challenge in the first stage is that a model optimized for purchases also needs to be (near) optimal for clicks, due to the click attribution constraint. The challenge in the second stage is that true probability of purchases is extremely small, and is difficult to accurately model. We propose a ranking model, followed by a calibration method, to sequentially address the two stage problem. We describe how ordinal ranking is a natural fit for the ad selection problem and how to learn a single model by optimizing for purchases, while being (near) optimal for clicks. We then propose a calibration method, which comprises of a novel *non-uniform binning* technique for empirical probability estimation, in conjunction with calibration functions such as isotonic and polynomial regression and Platt scaling. We provide empirical results on logged events from a major ad network, that demonstrate the superiority of ordinal model over binary classifiers for ranking ads and the superiority of our proposed calibration technique over traditional uniform binning based calibration technique.

1. INTRODUCTION

Performance display advertising consists of bidders competing on behalf of advertisers for ad impressions. Publishers solicit ads from multiple ad exchanges who, in turn, solicits bids of ad impressions from bidders. Each bidder uses the information sent by the ad exchange (user and website information) to select k ads from a plethora of ads ($k \sim 1-4$), calculates a bid value and sends it back to the ad exchange. We consider the setting where the winning bidder pays per impression but expects to earn revenue from post-impression performance. In contrast to pay-per-click advertising where the relative ordering of ads is of primary importance, pay-per-impression performance advertising requires accurate bids for an impression i based on its future expected value, defined as follows: $E[V(i)] = \sum_{a \in C(i)} Pr(a \mid i)V(a)$, where $C(i)$ denotes the space of all possible future outcomes after an impression is shown to a user, $Pr(a \mid i)$ indicates probability of outcome given impression, and $V(a)$ indicates the value of the outcome to the advertiser. For online retailers, the only outcome of interest is the purchase and its value is the sales price of the item. However, the advertiser decided how the purchase is attributed to the impression. We focus on the popular last-click purchase attribution model, where the advertiser only attributes the purchase to an ad if it was last clicked by the user before she made the purchase.

Abstracting out a number of architectural and implementation details of different ad networks, the bidder essentially has two sequential problems. The first is to build a model which would select a small number of ads out of a large selection, so that the selected ads are relevant. The second objective is to accurately estimate the purchase probability of the ad(s).

The challenge associated with the first problem is that a model optimized for purchases has to be (near) optimal for clicks. A common approach to predicting purchases is a binary classification model, which would predict whether a displayed ad would lead to a purchase. However, the model would likely fail to identify the larger set of ads which would at least be clicked, leading to opportunity cost (clicks might have intrinsic value to advertisers) and might drive down the overall volume of purchases. On the other hand, if a classification model is optimized for clicks, it might fail to distinguish the subset of ads which would likely lead to purchases. This would lead to a lot of ads being displayed which would not lead to purchases, and can be costly if the advertiser agrees to pay per click.

The challenge associated with the second problem is that the probability of displayed ads leading to purchases is extremely low (usually less than 0.01%). Accurately estimating the purchase probability is important because bids are calculated proportional to the predicted probabilities.

We propose a ranking model, followed by a calibration method involving a novel binning method, to sequentially address the two stage problem. Our contributions are as follows:

- We discuss how the hierarchical nature of events in the purchase funnel (displayed ad leads to click, which leads to purchase) naturally make ordinal ranking model suitable for the problem. The model is used to rank a large set of ads, with ads likely to lead to purchases ranked higher than ads likely to only lead to clicks, which are ranked higher than ads which are likely to be ignored.

- We propose a novel non-uniform binning technique, to bin the scores produced by the first stage model, and calculate empirical purchase probabilities corresponding to the binned scores. The binned data is then used to train popular calibration functions like isotonic regression, quadratic regression and Platt-scaled model to calibrate scores to probability of purchases.

- We conduct experiments on logged events of a major ad network and show how ordinal ranking model finds a better balance between predicting clicks and purchases, over separate click and purchase models based on binary classifiers. We also show that all three calibration functions, coupled with our non-uniform binning technique, perform significantly better than the corresponding functions coupled with traditional uniform binning.

2. PROBLEM FRAMEWORK

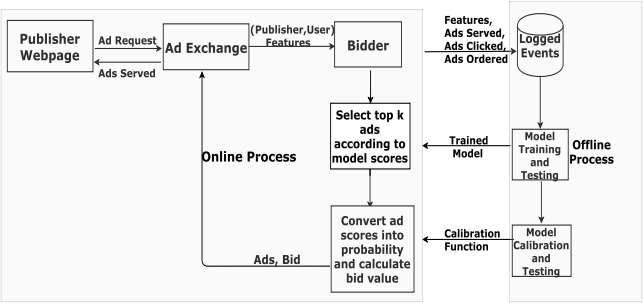

Figure 1: Offline Model Training and Online Interaction between Publisher, Ad Exchange and Bidder.

The following section provides an overview of performance display advertising, along with definition of various metrics. Fig 1 shows the framework of interaction between publisher, ad exchange and bidder, for major ad networks. When a user visits a publisher webpage, the publisher queries an ad exchange for ad impressions. The exchange sends information about user and publisher to one or more bidder. From a plethora of ads, the bidder selects the top few ranked ads, where ads are ranked according to scores induced by some model. The bidder then calculates a suitable bid value, according to probability of click/purchase, and sends the ads and bid value back to the ad exchange. If the bidder wins the auction (there may be multiple bidders which bid on the same ad exchange), then the ads are displayed on the publisher's webpage. Bidders compete on an expected cost per impression (eCPM) basis and ad exchanges or publishers can have very different auction and pricing strategies. Due to risk aversion among publishers, there is an increasing industry trend toward first-price auctions where bidders pay what they bid when they bid [7]. This entire interaction between publisher, ad exchange and bidder happens online, usually in less than 100ms. The information about user, publisher and ads displayed, ads clicked and ads that led to purchases are logged, and the bidder intermittently trains its models, offline, on the logs.

An *impression* event (**I**) is said to occur when an ad is displayed on publisher website. Similarly, a *click* event (**C**) is said to occur when an ad is clicked and a purchase event (**P**) is said to occur when an ad leads to a purchase. Since we assume the last-click attribution model, we will always consider that an ad can lead to a purchase only after it has been clicked, which, in turn, can only happen after it has been displayed. A *purchase funnel* is the hierarchical events funnel from impression to click and eventually to a purchase, i.e., $\mathbf{P} \subset \mathbf{C} \subset \mathbf{I}$ (when the events are represented as sets).

Note: Due to the hierarchical nature of the events, we will have the inherent understanding throughout the paper (unless stated otherwise), that an ad displayed means only displayed and not clicked, an ad clicked means displayed and clicked but product not purchased and an ad leading to purchase indicates the ad has moved through the entire purchase funnel.

The following metrics have been referred to in the paper: click-through-rate (CTR)= $\frac{\#clicks}{\#impressions}$, conversion-rate (CVR)= $\frac{\#purchases}{\#clicks}$, purchase-rate (CVI)= $\frac{\#purchases}{\#impressions}$. Note that CVI= CTR*CVR.

As mentioned before, the bidder trains the offline models for a two-fold problem. The first is to build a model that will (near) optimally predict which ad impression will go through a purchase funnel, which will only be clicked, and which will be untouched, and rank the ads accordingly. The second is to take the scores, induced on the selected ads by the model in the first stage, and calibrate it to actual CVI values, which will then allow suitable bid value calculation. In this paper, we void any detailed discussion on the types of features used in models. Our work focuses on the methods and should generalize to any reasonable set of features.

3. PREDICTING CLICK-ATTRIBUTED PURCHASES FROM IMPRESSIONS: ORDINAL REGRESSION MODEL

The prevalent technique in the literature is to separately predict clicks and purchases from impressions using separate binary classifiers.

Binary classifier: From the logged events data, for a user and publisher, and an ad a which was shown to the user, an instance x for the classifier is created as the following feature vector: $x = \{\psi_1(user, pub), \psi_2(user, pub, ad), \psi_3(ad)\} \in \mathbb{R}^d$, i.e., x is the feature vector constructed from one or more of user, publisher, ad and possibly their cross information. For a purchase prediction model, the class label is $y = 1$ if a led to a purchase, else $y = -1$. The training set $S = \{x_i, y_i\}_{i=1,\ldots,N}$ is then used to train a cost sensitive binary classifier $f_P()$ (which produces real valued score for x) by optimizing logistic loss:

$$\{C_1 \sum_{y_i=1} \log(1 + e^{-f(x_i)}) + C_{-1} \sum_{y_i=-1} \log(1 + e^{f(x_i)})\}$$

where C_1 and C_{-1} are differential weights to counter class imbalance. For a click prediction model (f_C), the labels are created as: $y = 1$ if a *at least* got a click, else $y = -1$, and the model is trained similarly. When a new request from an ad exchange comes in, the function f_P (resp. f_C) ranks a large set of possible ads, and selects the top few ads, with the expectation that top ranked ads are more likely to lead to purchases (resp. at least get a click) than ads lower down the list.

Drawback of binary classifier based models: The f_C function is not trained to differentiate between click events

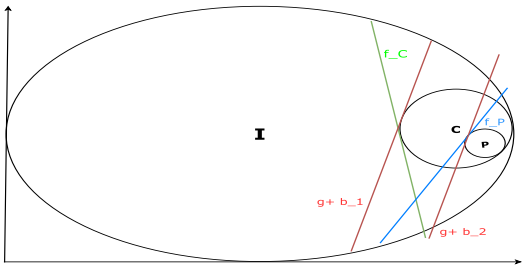

Figure 2: Binary Classification and Ordinal Regression Models. f_C is the binary classifier separating non-click from click events, while f_P is a binary classifier separating purchase from non-purchase events. g is the ordinal model, with bias terms b_1 and b_2, modeling the entire funnel. The geometry of the hyperplanes are such that ordinal model does as well as f_P in predicting purchases, while doing much better in predicting clicks. *Best viewed in color.*

which lead to purchases against "only click" events, while f_P function is not trained to differentiate between "at least click" and non-click events. Thus, neither of the models are geared towards ranking ads to separate click-attributed purchases, only clicks and only impression events.

Ordinal ranking model: In an ordinal ranking (regression) problem setting, a function f_O predicts, for an instance $x \in \mathbb{R}^d$, a label $y \in \{1, 2, \ldots, K\}$, where the classes are *ordered* as $1 <= 2 <= \ldots <= K$, with $K >= 2$. We can train an ordinal ranking function for the purchase funnel as follows: for a user and publisher, and an ad a which was shown to the user, the feature vector x is constructed in the same way as before. The class label is encoded as $a \in I \setminus C \implies y = 1$, $a \in C \setminus P \implies y = 2$ and $a \in P \implies y = 3$. The ordering is *natural* since advertisers care most for purchases, followed by "only click" event, which is more valuable than an ignored impression.

Training ordinal model using binary classifier training script: Due to widespread use of binary classifiers over the years, a number of papers have been written to train an ordinal model by reducing the ordinal setting to a binary classification setting. We will adopt a technique, whose generic version has been outlined in [4], that learns multiple parallel hyperplanes separating the different classes.

Each instance x is expanded, each class y is reduced to a binary class, and a differential weight is calculated as follows:

$$x^k = (x, e_k),\ y^k = 2 \cdot \mathbb{1}[k < y] - 1,\ w_y^k = |C_{y,k} - C_{y,k+1}|, \quad \text{for } k \in \{1, \ldots, K-1\}$$

where e_k is the standard basis vector in dimension $K-1$ and $C_{y,k} = \mathbb{1}[y \neq k]$. For e.g., with $K = 3$, an instance x, with $y = 1$, will expand into 2 instances: $x^1 = (x, 1, 0)$ and $x^2 = (x, 0, 1)$, with corresponding labels: $y^1 = -1, y^2 = -1$ and weights: $w_y^1 = 1$ and $w_y^2 = 0$. After the reduction, an ordinal function f_O is trained by optimizing the logistic loss, similar to a binary classifier:

$$\{C_1 \sum_{y_i^k = 1} w_{y_i}^k \log(1 + e^{-f(x_i^k)}) + C_{-1} \sum_{y_i^k = -1} w_{y_i}^k \log(1 + e^{f(x_i^k)})\} \tag{1}$$

Advantage of ordinal model over binary classification models: It is expected that the ordinal model will learn to adequately weigh the features which are discriminative for all the three classes in our problem setting; something that a binary classifier will fail to do.The ordinal function $f_O : x^k \mapsto \mathbb{R}$ decomposes into a scoring function $g(\cdot)$, which operates on the original feature vector ($g : x \mapsto \mathbb{R}$) and bias terms $b_1, b_2, \ldots, b_{K-1}$, which operate on the extensions, as follows: $f_O(x^k) = g(x) + b_k$. The vital component of the ordinal function is the function $g(\cdot)$, which is the hyperplane equivalent of the binary classifiers. When a new request from an ad exchange comes in, the function g ranks a large set of possible ads by inducing scores (similar to f_C and f_P), with the expectation that ads which are likely to lead to purchases will be ranked higher than ads which are likely to be only clicked, which in turn, are ranked higher than ads which are likely to be ignored. Fig 2 gives a pictorial representation of the hierarchical set of events, the classifiers f_C and f_P and ordinal model g (with bias terms).

Class imbalance and differential weights: A major problem for purchase modeling is that CVI values are extremely low (usually $< 0.01\%$); hence the number of purchase events is a small fraction of number of impression events in a training set. The prevalent technique, of training a binary classifier in presence of high class imbalance, is to sub-sample the negative examples, followed by differentially weighing the two classes during logistic loss optimization ($C_1 \gg C_{-1}$). Since the ordinal model is trained by reducing it to a binary classification problem, the net weights on negative examples ($y_i^k = -1$) become $C_{-1} \cdot w_{y_i}^k$ and the net weights on the positive examples ($y_i^k = 1$) become $C_1 \cdot w_{y_i}^k$. As the weights w_y^k are theoretically set, the only parameters that need to be varied are C_1 and C_{-1}, just like in binary classification [1].

4. CALIBRATION OF MODEL SCORES

The first stage model induces scores on the ads, and top ranked ads are selected. The objective for the bidder now is to estimate the CVI of the selected ad(s). Calibration is the technique of converting the model induced scores to actual probabilities.

While training the ordinal regression model, we heavily subsample the negative examples for artificial class balancing, which leads to the proportions of instances from different classes, in the training set, not matching the true class proportions. It is thus important to calibrate the scores to the true probabilities of purchases, which can then be used to set bid values. The natural approach of estimating probability, by passing the score of a test instance through the sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$) is not appropriate because the probabilities will be highly overestimated for moderately large scores (even for $x = 0$, $\sigma(x) = 0.5$, which is too high an estimate for purchase probability). We first empirically estimate true probabilities at different score ranges, by describing a novel *non-uniform binning* technique which is used to bin the scores of a validation set, and then describe various functions to establish a continuous relationship between scores and probabilities. Note that the calibration technique can be used for scores produced by either the ordinal model

[1]In order to train models on the data provided by the ad network, which was in terabytes and stored on hadoop clusters, we modified the Liblinear [1] package by introducing a number of optimizations and enhancements to the algorithm.

or the classification based purchase model f_P.

Binning technique for empirical probability estimation: Empirical probabilities are calculated by dividing the model induced scores into bins, and calculating the proportion of positive to negative class instances in each bin. For example, let $S_v = \{x_i, y_i\}_{i=1,\ldots,m}$ be a validation set, where $x \in \mathbb{R}^d$ are instances (feature vectors) and $y \in \{-1, 1\}$ are class labels, representing non-purchase/ purchase respectively. A model $f(\cdot)$ produces scores from the instances as $f(x_i) = s_i$. Assume that the scores are sorted in ascending order, and M bins are constructed, with bin edges set as $v_1 < v_2 < \ldots < v_{M+1}$. Then, the empirical probability of a bin $[v_j, v_{j+1}]$ is

$$\hat{p}_j = \frac{\#\{x_i, y_i | y_i = 1 \wedge \ f(x_i) \in [v_j, v_{j+1}]\}}{\#\{x_i, y_i | f(x_i) \in [v_j, v_{j+1}]\}}. \qquad (2)$$

Natural questions that arise during bin construction process are: a) how many bins to construct and b) what should be the distribution of the bins. For the traditional *equal width uniform binning technique* [2], too many bins would lead to a number of bins having 0 positive instances; hence 0 empirical probability, and too few bins would lead to most bins having close to global empirical probability (i.e., global CVI), and not be able to differentiate between high and low score zones. We propose the following technique:

1. Sort scores in ascending order and construct bins, with each bin having approximately equal number of positive instances $p > 0$. Each bin will thus vary in number of negative instances, and hence in bin width.

2. Find the appropriate number of positive instances p by doing a grid search (for eg. $5 \leq p \leq 20$), and setting the final value of p as one which produces the smallest *Kullback-Leibler divergence* between the bin empirical probabilities (Eq 2) and the probabilities from *sigmoid* function operating on the average score of each bin (i.e., average of the scores inside the bin), i.e., $\sigma(s_{avg})$.

We explain the reasoning behind the grid search technique. When we train models by optimizing logistic loss, we implicitly assume that the marginal distribution of y given x is the sigmoid function. Since the assumption is made on the distribution, produced from distorting the original distribution (distortion by subsampling negatives and differential weighting), we cannot expect the probability of a test instance to be the sigmoid transformation of the score (in fact, this is why sigmoid transformation of score leads to overestimation of probability). However, since the model is trained on logistic loss, we should reasonably expect the curve interpolating the empirical probability estimates, obtained from binning the scores produced by the model, to resemble a sigmoid more than any other geometric shape (say a monotonically increasing line). Thus, a reasonable way to bin the scores (and hence estimate empirical probabilities at different score range) is to force the interpolating curve to have minimum KL divergence to the sigmoid function, while still constrained by the estimated empirical probabilities, so as to not heavily overestimate CVI at different score range.

Establishing functional relationship between empirical probabilities and scores: After the binning process, we have point-wise bin average scores $\{b_i\}_{i\in[k]}$ and corresponding empirical probabilities $\{\hat{p}_i\}_{i\in[k]}$. The objective now is to establish a continuous function relating the two, so that a test score s can be matched to a probability value.

Isotonic regression: Isotonic regression is the technique of fitting a free-form line to a sequence of observations, with the free form-line being monotonically non-decreasing everywhere. Isotonic regression is solved by the Pool Adjacent Violator's algorithm (PAVA) [5], which in our case, given the empirical probabilities $\hat{p}_i$, will solve the following quadratic program: $\min_{p_1 \leq p_2 \ldots \leq p_k} \sum_{i=1}^{k} w_i(\hat{p}_i - p_i)^2$, where w_i are user defined weights. For test score $s \in [b_i, b_{i+1}]$, the predicted probability is calculated from linear interpolation between p_i and p_{i+1}.

Polynomial regression: A polynomial regression function between empirical probabilities and scores is modeled as follows: $\min_\beta \sum_{i=1}^{k} w_i(\hat{p}_i - \text{poly}(b_i)^\top \beta)^2$, where $\text{poly}(b_i)$ is the polynomial expansion of scalar b_i. A test score s is predicted to have a probability $\text{poly}(s)^\top \beta$.

Platt scaling: The Platt-scaling technique, as described in [6], learns scaling parameters a and b from the following relation: $\hat{p}_i = \frac{1}{1+e^{-ab_i+b}}$ (each $\hat{p}_i$ is weighted by user defined weight w_i. The parameters can be learnt in R, by using package *glm* with quasibinomial family). Probability of a test score s is given by $\frac{1}{1+e^{-as+b}}$.

Combining the modeling and calibration procedure: Assume that the bidder uses the ordinal model for ranking ads. When an ad request comes in from the exchange, the bidder first ranks a large set of ads using the ordinal model. Then, the top k ads ($k \sim 1-4$) are selected, which the bidder wants to display. The bid value is then calculated by estimating the CVI of the k ads, using the calibration method. The exact bid computation technique is left to the discretion of the bidder.

5. EXPERIMENTS

We tested our models on data collected from a major ad network. Since the data is highly proprietary in nature, we report only relative performance numbers.

5.1 Ordinal Model Compared to Classification Models

We collected logged data for impressions, clicks and purchases, across a whole week, for training ordinal and classification models. The raw training data had nearly hundred million impressions, with few million unique ads. The ads had a long tailed distribution, with a minority of ads displayed most frequently, while a majority were displayed only a few times. The global CVI was a small fraction of a percent. We conducted heavy sub-sampling (about 2% retention) of the negative instances, with uniform at random sampling of each day's data and stratified sampling while combining the whole week's data. Due to aggressive sub-sampling, the final training data had impressions in the order of a few million, clicks in the order of a few hundred thousand and purchases in the order of a few thousand. The number of unique features in the final training set was in the order of tens of millions. The feature set was a very simple one-shot encoding of attributes such as time of day, location, publisher and user information, etc. Each instance was represented as an extremely sparse binary feature vector (for e.g. : is the user from US, is the operating system Windows, etc). Joint user, publisher and ad features were composed by crossing the non-ad binary features with the

Table 1: Relative performance of 2 binary classification models (f_C and f_P) and ordinal regression model (f_O), in terms of AUC metric, averaged over 7 days. For $I \rightarrow C$ and $I \rightarrow P$ prediction, f_C and f_P perform best respectively, with their respective AUC numbers reported as 0. The numbers in the other columns show mean percentage gain in AUC compared to best numbers (numbers in bracket show std. dev.). f_O produces near optimal performance for $I \rightarrow P$ prediction, while performing substantially better than f_P for $I \rightarrow C$ prediction. All numbers have been expressed as %.

Predicting	f_C	f_P	f_O
$I \rightarrow C$	0	-17.2 (1.1)	-5.6 (0.5)
$I \rightarrow P$	-22.7 (3.1)	0	-0.85 (0.04)

ad id of the ad which was displayed for the corresponding impression.

5.2 Results

We trained three models; ordinal model (f_O), classification model for clicks (f_C) and classification model for purchases (f_P), and tested on logged data for 7 separate days of another week. As shown in Table 1, f_C degrades performance by 23% compared to f_P on the purchase prediction task ($I \rightarrow P$). Similarly, f_P is 17% worse than f_C on the click prediction task ($I \rightarrow C$). The ordinal regression is a good single predictor for both tasks. This is further confirmed from the ROC curves (Figure 3), which shows that when predicting purchases, f_O has significantly higher TPR (true positive rate) for same value of FPR (false positive rate) than f_C model, while nearly overlapping with the f_P model.

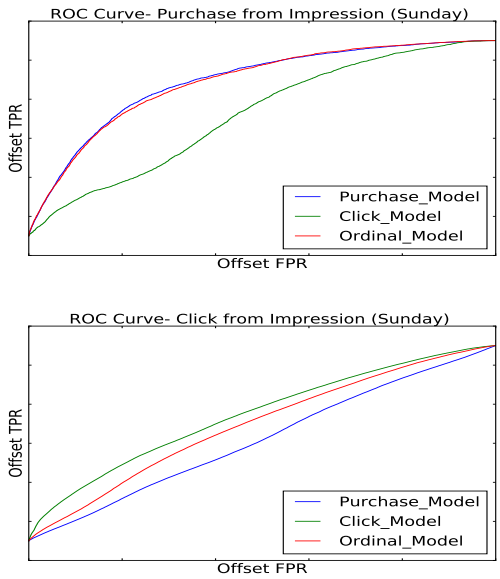

Figure 3: ROC curve for f_C, f_P and f_O models, predicting purchases from impression (**top**) and clicks from impressions (**bottom**), for a single test day (Sunday). The axes values have been offset by bias terms. *Best viewed in color.*

Table 2: Log-loss improvement for each calibration function, in conjunction with proposed non-uniform binning, over uniform binning, for CVI prediction. The results have been averaged over 5 days (numbers in bracket show std.dev). Baseline has been set to 0. All number have been expressed as %.

Binning	Isotonic	Quadratic	Platt-scaled
Uniform	0	0	0
Non-uniform	3.45 (2.55)	2.95 (1.94)	3.01 (1.75)

5.3 Model Score Calibration

We compared calibration performance, obtained using our proposed non-uniform binning technique against standard uniform binning technique, for all three calibration functions, in terms of predicting CVI. We used quadratic polynomial regression in our experiments, which produces a good balance between fit to training data and complexity of model. We trained the calibration functions on the ordinal model induced scores for 2 days of logged data, which we considered as validation set. The calibrations functions were then compared on logged data for 5 separate test days.

Training calibration functions: During training, all the three calibration functions were scaled by weights w_i, with $w_i = \sqrt{n_i}$, where n_i was the total number of instances in the corresponding bin. The weighing scheme put more emphasis on function fit where empirical probability was calculated with higher confidence (i.e. with more instances) .

5.4 Results

The proposed non-uniform binning and traditional uniform binning techniques were compared on test data sets using the log loss metric, which is defined as: $\frac{-1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)]$, where y_i is the label (purchase/ no purchase) and $\hat{p}_i$ is predicted purchase probability. We chose log loss instead of offline contextual bandit evaluation technique [3] because we did not have information about the policy which was used to serve the impressions which formed the logged events. Table 2 shows the performance improvement of each calibration function, when the two binning techniques are compared. 50 bins were used in the uniform binning technique. More bins led to a number of bins having no positive instances, and less bins produced worse performance. Each function performs better with the non-uniform binning technique. *The improvement is significant since even a fraction percent improvement in purchase probability prediction can have major implications on bidder's performance.*

Prediction in high score region: One key problem in performance ad bidding is deciding between undervalued and overvalued bids in the high score region. An impression with higher score is more likely to lead to a purchase, as per our ordinal model. However, during calibration on validation set, the number of instances (positive and negative) in high score region is much smaller than in the low score region, and hence, probability estimates in the high score region is constructed with low confidence, leading to possible calibration function misfit. Figure 4 shows that Platt scaled function predicts much higher purchase probability than Isotonic and quadratic regression functions, in

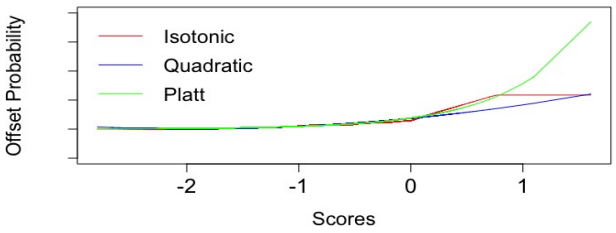

Figure 4: CVI prediction of Isotonic, Quadratic and Platt functions, trained in conjunction with non-uniform binning strategy, for a test day. *Best viewed in color.*

the high score region, and log loss may not be a suitable metric to judge whether the probability is being overestimated. *It remains a open question as to how to find a better way to fit calibration models in the high score region and how to quantify the fit using some novel metric.*

6. CONCLUSION

We discussed two key problems for bidders in performance display advertising; to optimize for post-click purchases while having good click prediction performance, and to calibrate model induced scores to true probability of purchases. We proposed the ordinal ranking model for the first problem and a novel binning technique, used in conjunction with isotonic, polynomial and Platt-scaled functions, for the second problem. Our extensive experiments showed that ordinal regression model is better suited to (near) optimally predict purchases, while having good click prediction performance, than separate binary classifiers. We also showed that our proposed non-uniform binning technique produces better result than standard uniform binning, for each of the calibration functions.

7. REFERENCES

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874, 2008.

[2] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of SIGKDD*, pages 768–776. ACM, 2012.

[3] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of WSDM*, pages 297–306. ACM, 2011.

[4] L. Li and H.-T. Lin. Ordinal regression by extended binary classification. In *NIPS*, pages 865–872, 2006.

[5] P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

[6] Y. Tagami, S. Ono, K. Yamamoto, K. Tsukamoto, and A. Tajima. Ctr prediction for contextual advertising: Learning-to-rank approach. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, page 4. ACM, 2013.

[7] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, page 3. ACM, 2013.