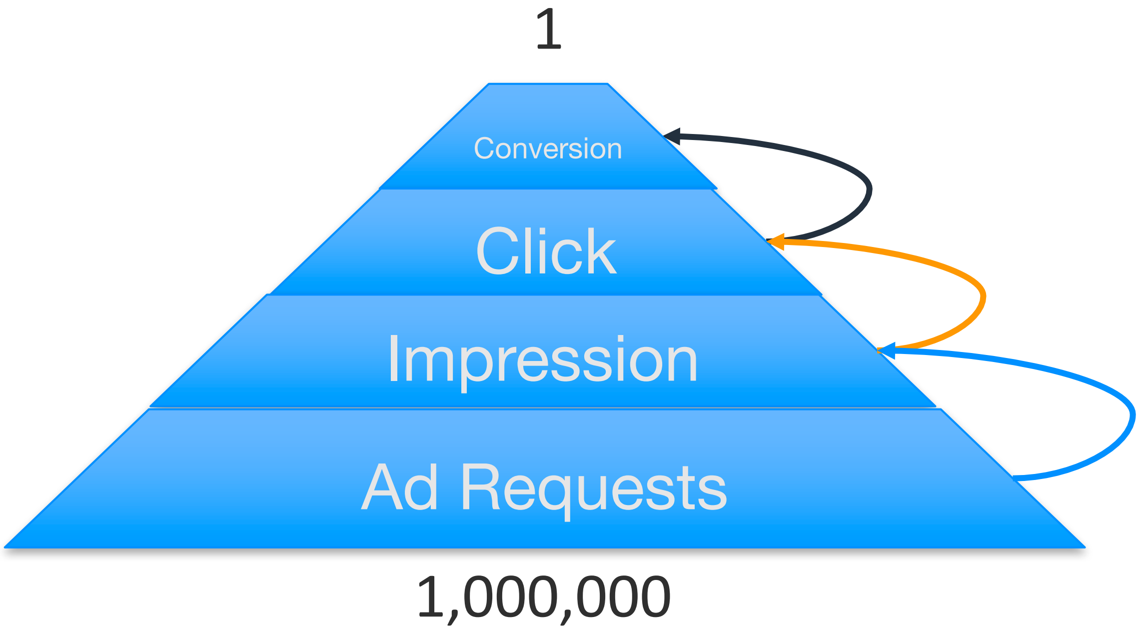
Ranking and Calibrating Click-Attributed Purchases in Performance Display Advertising

Sougata Chaudhuri, Abraham Bagherjeiran (*), and James Liu
A9 Advertising Science, A9.com (An Amazon Subsidiary)

August 14, 2017

Conversion Funnel

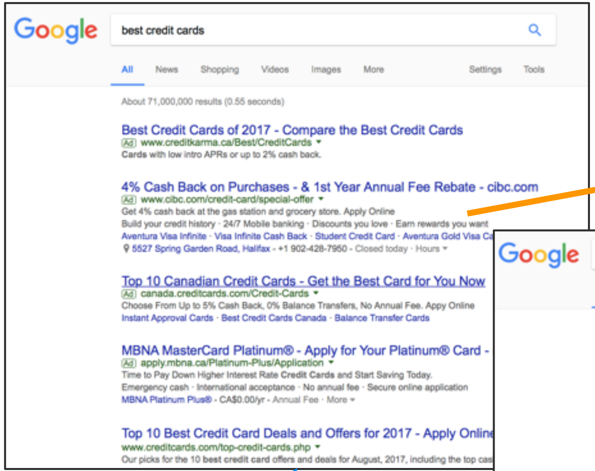


Advertising is a lossy business.

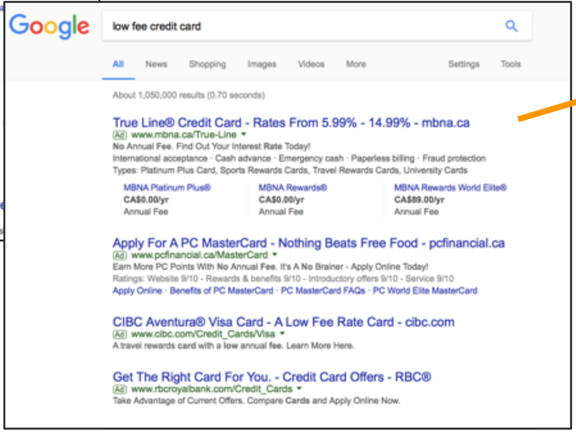
Sponsored Search

Not this talk

"best credit cars"



"low fee credit card"



Impression

Advertiser Page

Click

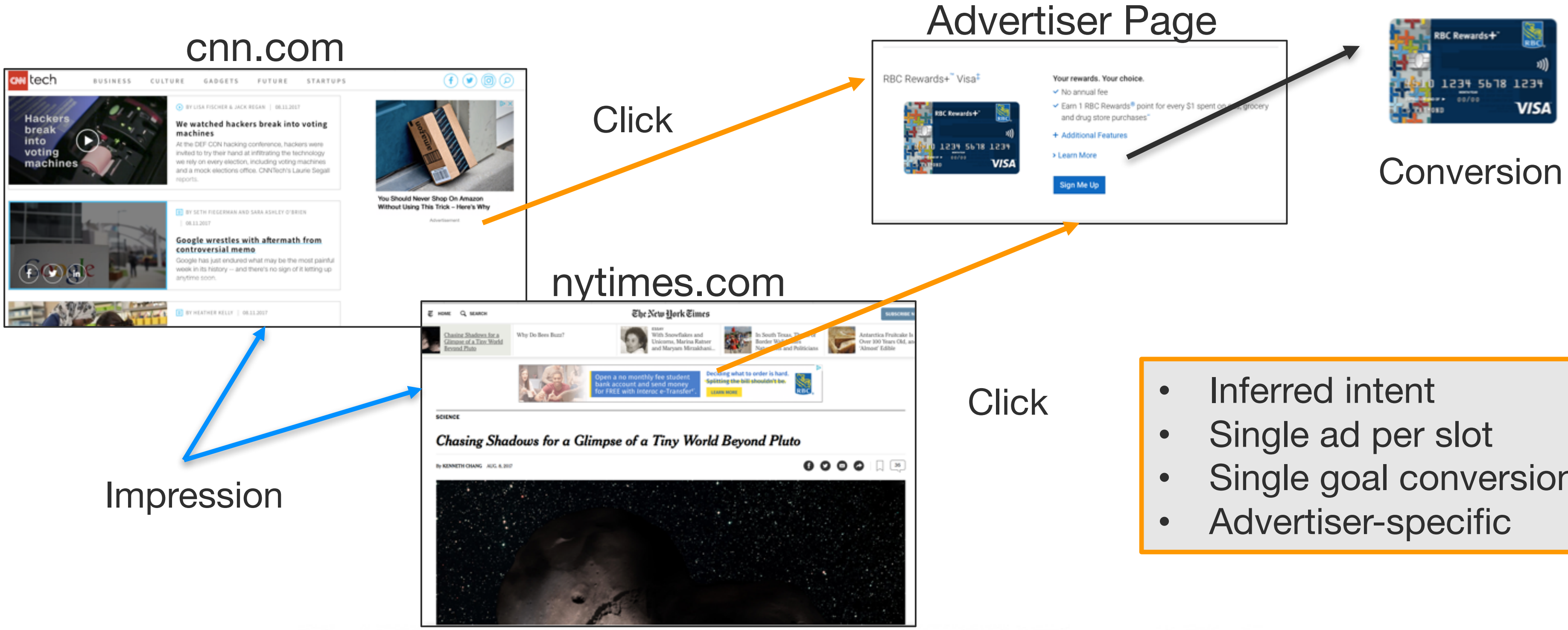
Conversion

- Direct intent
- Multiple ads per slot
- Single goal conversion
- Advertiser-specific

Funnel: Impression, click, conversion

Performance Display Advertising

Not this talk



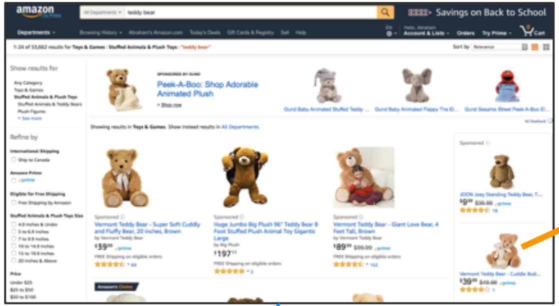
- Inferred intent
- Single ad per slot
- Single goal conversion
- Advertiser-specific

Funnel: Impression, click, conversion

Not this talk

Amazon Sponsored Products

Amazon Search



Amazon Detail Page

Click

Purchase

Impression

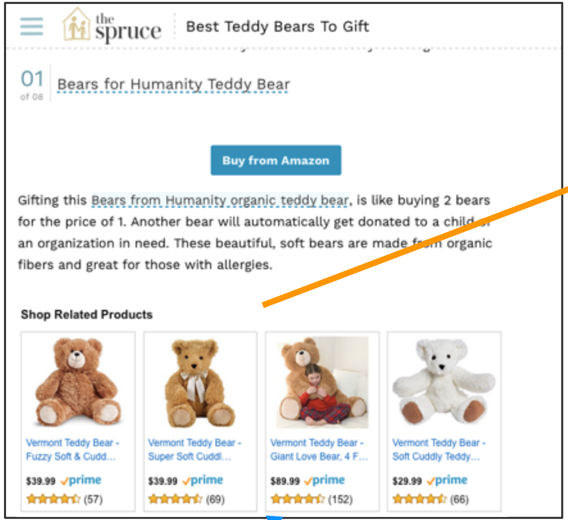
- Direct intent
- Multiple ads per slot
- Single sale
- Sales for merchant only

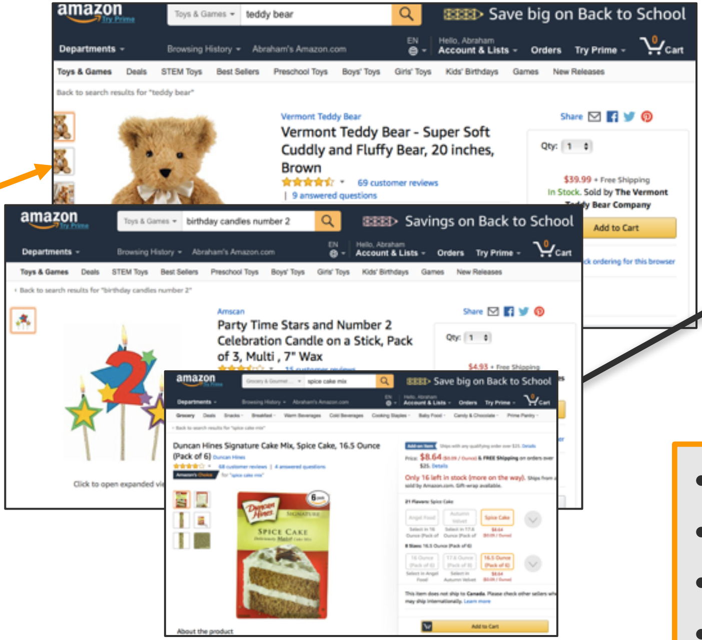
Purchase funnel: Impression, click, purchase

Amazon Contextual Ads

This talk

thespruce.com





Purchase Halo

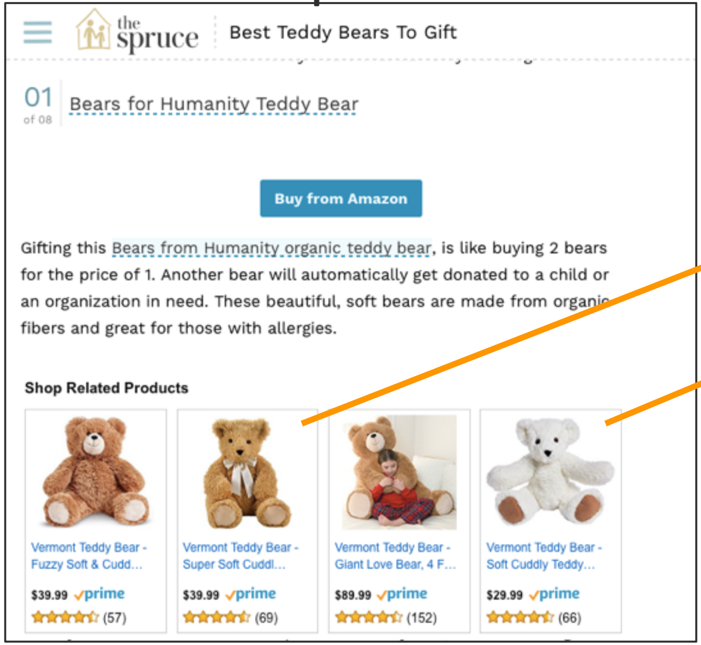
- Inferred intent
- Multiple ads per slot
- Complex goal
- All orders to Amazon

Impression

Purchase funnel: Impression, click, purchase(s)

Amazon Contextual Ads Problem

Some publisher



amazon

Purchase Halo

Preference: Purchases first, but clicks are good, too.

Problem Statement

- Input
  - User
  - Publisher page
  - List of ads
- Output
  - 5-10 ads, ranked by a score
- Objective
  - Maximize total expected value of purchase halo

$$x = \{\psi_1(user, pub), \psi_2(user, pub, ad), \psi_3(ad)\} \in \mathbb{R}^d,$$

Extracted interaction features

$$f_O(x^k)$$

Single ranking function score

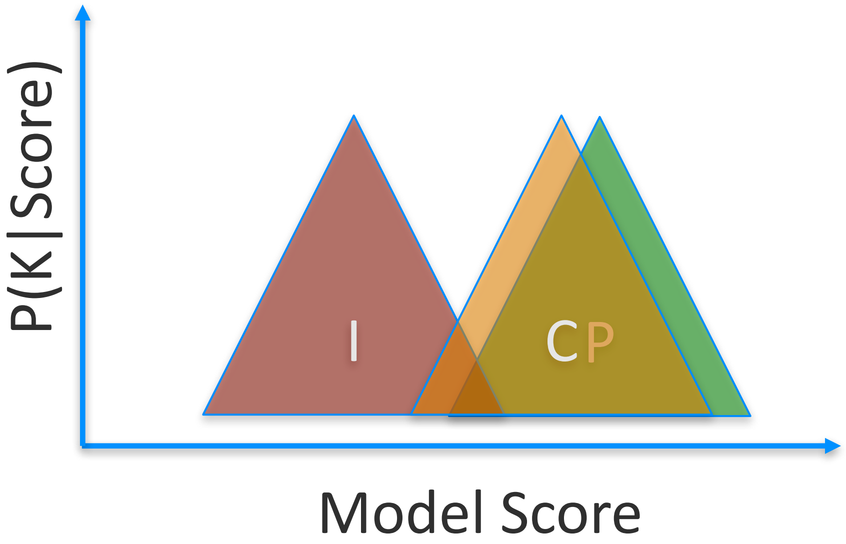
How should we setup the learning problem?

Related Work: Modeling with Preferences

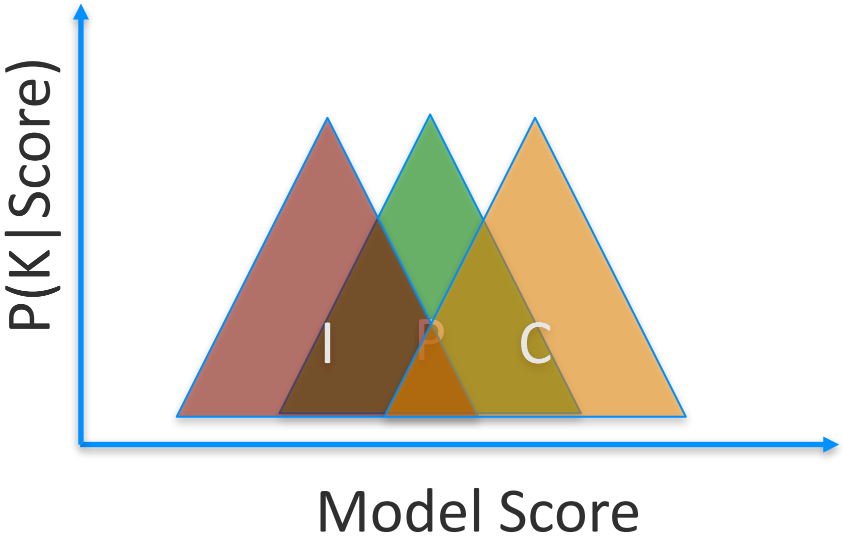
- Binary classification (with weights)
  - Purchase target only or click target only
- Compound models
  - P(Click) * P(Conversion)
- Pair-wise comparisons
  - Complex to evaluate
- Value Regression
  - How to capture value of clicks

Binary Classification

Assumed structure



Nested structure



Binary assumes that P and C are the same.

Binary Classification Only

- One-Step
  - I → C: Clicks v. Impressions
  - I → P: Purchases v. Impressions
- Evaluation
  - I → C: Great at predicting clicks, 17% worse at predicting purchases
  - I → P: Great at predicting purchases, 23% worse at predicting clicks

Does I → P predict the “good clicks” vs “bad”?

Why Binary Classification Isn't Enough

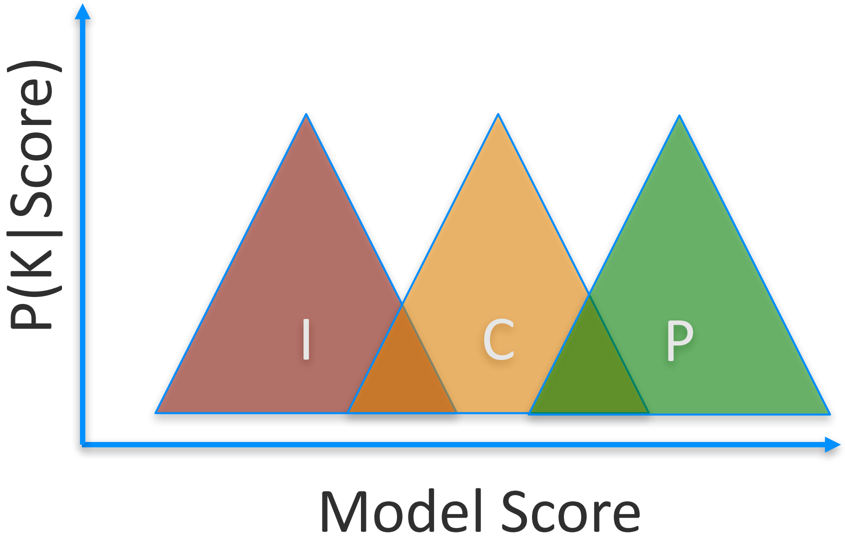
- Good clicks
  - In online tests, observed click rate went down
  - Overall post-click conversion rate also went down
  - Overall conversion rate went down
- Meaning
  - Nested relationship appears to be present

Ordinal Regression

- K nested classes
  - Impression
  - Click
  - Purchase
- Jointly train parallel linear models separating all classes

All clicks are equal, but some are better than others

Ordinal Classification



- Single score to separate multiple classes
- Preserves preferences
- Easy to evaluate

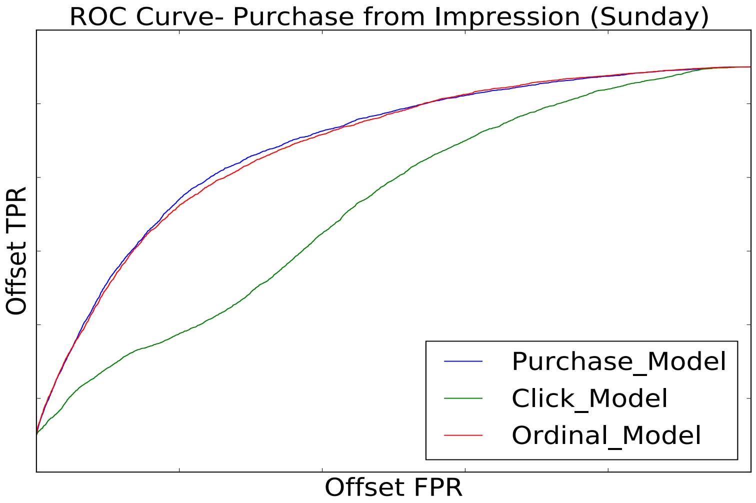
Binary assumes that C and P are dependent.

Binary v. Ordinal

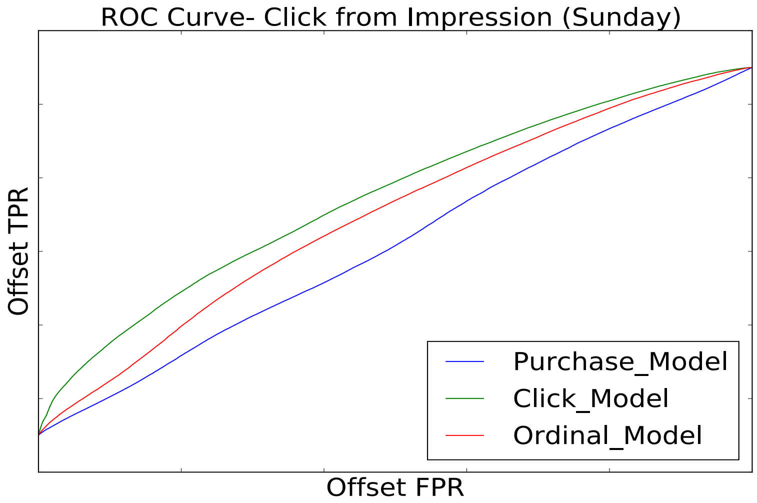
- Comparison
  - I → C: Clicks v. Impressions
  - I → P: Purchases v. Impressions
  - I → C → P: Ordinal
- Evaluation
  - I → C: Great at predicting clicks, 17% worse at predicting purchases
  - I → P: Great at predicting purchases, 23% worse at predicting clicks
  - I → C → P: 5% worse at predicting clicks, 1% worse at purchases

Ordinal is a good compromise between classes

I → P


ROC Curve- Purchase from Impression (Sunday)
Offset TPR
Offset FPR
Purchase_Model
Click_Model
Ordinal_Model

I → C


ROC Curve- Click from Impression (Sunday)
Offset TPR
Offset FPR
Purchase_Model
Click_Model
Ordinal_Model

Complications

- Training ordinal models
  - Extension to binary classification for linear models
  - Increases data training size
  - Increase efficiency of batch trainer with disk cache
- Data Preparation
  - Weigh classes careful to adjust for imbalance
- Calibration
  - Evaluated as a single model, score isn’t calibrated

Most of these complications are not too bad

Calibration

- Why is this a problem?
  - Sigmoid isn’t good at small probability values (10^{-6})
  - Other link functions possible
- Model and distribution stability
  - Data fluctuations, cold start
  - Training / Test distribution differences
- Sometimes you need a probability score
  - First price auction: P(Purchase) * Sales
  - Small errors in price = Big problems

Despite what you’ve heard, growing amount of ad auctions are closer to 1st price than 2nd price.

Calibration isn't solved

- Few solutions everyone uses
  - PAV, Isotonic, Platt
- How do you know it's working?
  - Log loss: $\frac{-1}{n} \sum_{i=1}^{n} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$
  - What's the ground truth? What if there is only a few events?
- Highly sensitive to binning strategies
  - 3% Log loss improvement by changing binning

Summary and Extensions

- Summary
  - Ordinal regression is a good strategy for ranking with several objectives
- Additional event types for the full funnel
  - Halo purchase
  - Exact purchase
  - Viewable impressions
  - Ad interactions

Appendix

Compound Models

- Multiple two models
  - P(Click) * P(Conversion | Click)
- Benefits
  - Use different features or datasets for each model
- Problems
  - How to avoid compounding errors when ranking on the joint score?
  - When multiple ads are present, does not provide the right penalty for non-converting clicks
  - Unclear for margin-maximization models.

Very popular method but not a good fit for ranking

Problem

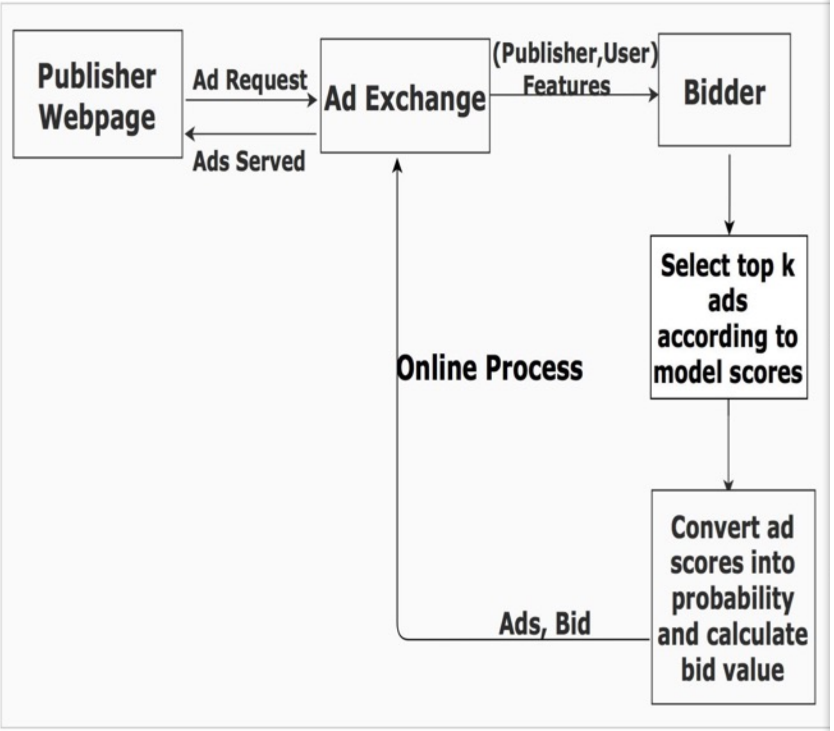
Select ads and calculate bid value to win ad impressions on publisher webpage.

Objective

Ads should lead to conversions/purchases after being clicked by user (click-attributed purchase)

Application

Amazon Associates Native Shopping Ads Program



General Overview of Online Interaction between Publisher, Ad Exchange and Bidder.

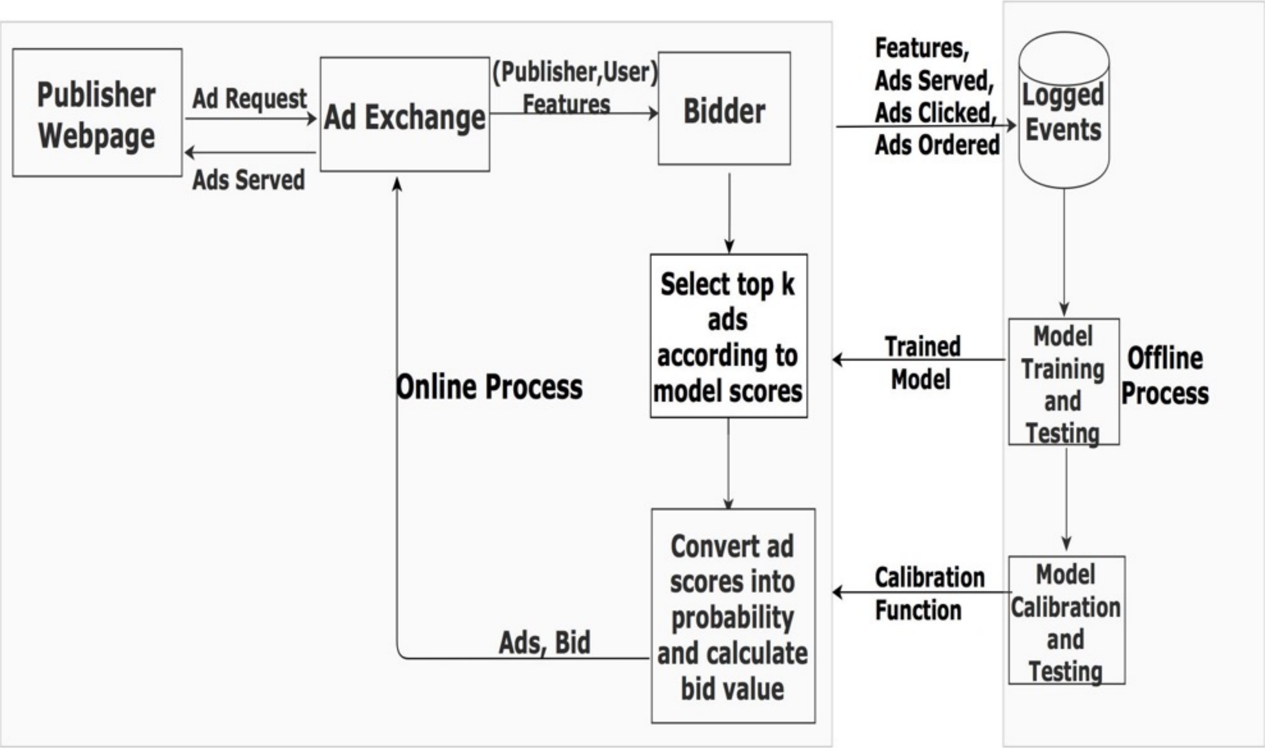
Challenges

- Model optimized for purchases also needs to be (near) optimal for clicks. Traditional binary classification models are not designed to optimize for both.

- Estimating the probability of purchases, which is extremely small, is difficult.

Our Approach

- Two stage modeling approach.
- Ad Ranking- single ordinal ranking model, which is optimized for purchases, while still being near optimal for clicks.
- Probability estimation- purchase purchase probability of top ranked ads are estimated by a calibration method, which combines a non-uniform binning strategy, in conjunction with continuous functions such as isotonic and polynomial regression and Platt scaling.



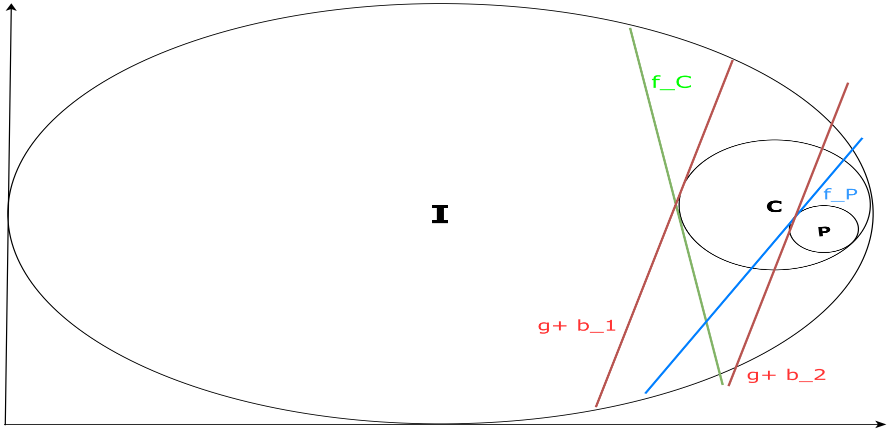
General Overview of Offline Model Training Pipeline and Online Interaction between Publisher, Ad Exchange and Bidder.

Definitions

- Purchase funnel: hierarchical events funnel from impression to click and eventually to a purchase, i.e.,

$$P \subset C \subset I$$

- Click-Through-Rate (CTR): $\frac{\text{No. of clicks}}{\text{No. of impressions}}$
- Conversion-Rate (CVR): $\frac{\text{No. of purchases}}{\text{No. of clicks}}$
- Purchase-Rate (CVI): $\frac{\text{No. of purchases}}{\text{No. of impressions}}$



Binary Classification and Ordinal Regression Models.

Ordinal Ranking Model: A function $f(\cdot)$ for an instance $a \in \mathbb{R}^d$ predicts a class $y \in \{1, 2, \ldots, K\}$, with classes ranked as $1 <= 2 <= K$. It is a natural fit for modeling purchase funnel by producing classes for an ad as follows:

$a \in I \setminus C \implies y = 1$ $a \in C \setminus P \implies y = 2$ $a \in P \implies y = 3$

The ordinal ranking model can actually be reduced to a binary classification problem and trained using well-tuned binary classification training scripts [1].

1. Ranking and Calibrating Click-Attributed Purchases in Performance Display Advertising- Chaudhuri et al., AdKdd and TargetAd, 2017.

Calibration

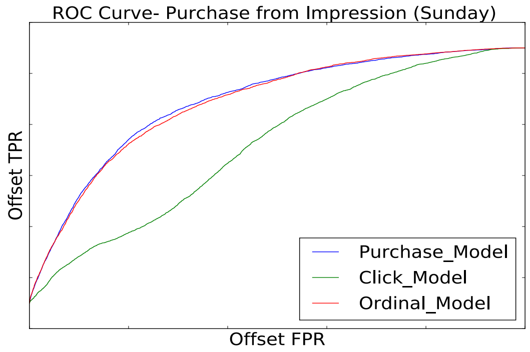
- The scores induced by ranking model is then calibrated to predict probability of purchases.

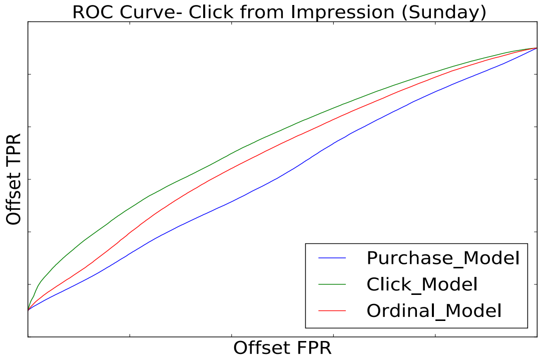
- Empirical probability of purchases is estimated from validation data, by a non-uniform binning strategy, which are then made continuous by fitting traditional regression based calibration functions like isotonic, quadratic and Platt-scaled.

Empirical Results

Prediction	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)

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 (numbers in bracket show std. dev.). All numbers have been expressed as % .


ROC Curve- Purchase from Impression (Sunday)
Offset TPR
Offset FPR
Purchase_Model
Click_Model
Ordinal_Model


ROC Curve- Click from Impression (Sunday)
Offset TPR
Offset FPR
Purchase_Model
Click_Model
Ordinal_Model

log-loss $$ll = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

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

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).

Thank You!