Advanced & Practical Issues on Ad Selection

A Data Scientist’s Perspective

Ad Ranking at Kakao

ben.hur & dan.lee @kakaocorp.com

if (kakao) dev 2019
Previously at if(kakao)...

You’ll find me at https://if.kakao.com/2018
\( eCPM = BA \times pCTR \)

- **Bid Amount**
- **Relevance** (Response rate)
  - \( \Pr(y = 1 \mid x_{UA}) \)
  - \( X \) (Embedding)
  - Training & Optimization
- **AutoBid**
Reserve price, feedback (HideAds), abusing

On-live, budget, inventory (format, size), time

Adv-set user segment → **Automatic (LAL)**

**Historical User-Ad interaction**

\[ eCPM = BA \times pCTR \times pCVR \]

Cut-off: eCPM, pCTR, pCVR, BA

Frequency capping, implicit feedback

Auction (Hard/Soft bid floor)
Targeting
discover your real audience
Male or young → Outdoor activity → Rider → Potential customers

- Gift for YOU
  - Buy one get one free
  - Shop Now

- It's Travel Time
  - Refresh yourself
  - Booking

- Congratulations!
  - Happy birthday
  - Purchase
Inventory buying

Static Info.
- Gender, age, region
- Interest

Context
- Placement (inventory)
- Current time & location
- Device/OS
- Wifi/Cellular

Custom
- Upload customers

Audience buying

Dynamic (behavior) Info.
- Site visit
- Product (Page) view
- Keyword query
- Category
- Cohort

Effective & Coverage

LookALike
$IV = (p - q) \log \frac{p(1 - q)}{(1-p)q}$

Impression ➔ Click ➔ Conversion
Candidate Selection
pick me pick me pick me up
Only 1,000+ creatives held 95% impressions.
vs 1M creatives

# creatives (1w, Mobile only)

10 50 100 200 500 500+
User Ad Creatives = T1 T2 T3 T4 A4 A8

ui

click{user, creative}
User Embedding
who are you? reveal yourself
Curse of Dimensionality

Over millions to billions of sparse encoding
Creative, subscription, KWD, …

User

Dim. Reduction

PCA / AE
Clustering
Hashing trick
Random Projection
SVD / [N/B]MF
**LDA** (topic modeling)
W2V

Embedding vector
Feature Embedding with Dimensionality Reduction

- Reliability / Speed / Scalability
- Robustness (+) vs Information loss (-)
- Abstraction (anonymity) vs Less interpretability (-)

Lessons learned

- 30 ~ 50 topics enough
- Multiple sources in one embedding? Not work properly
- How to retain previous dimension structure (topic semantics)
  - Syntactic hashing (short term) and re-training (long term)
RIG (Relative Information Gain)

LogLoss

# Topics

Baseline: 0.22
10: 0.21
20: 0.20
30: 0.18
40: 0.17
50: 0.16

LogLoss: 0.062

if (kakao) dev 2019
Deep… Prediction
you don't need to be too deep
\[ \text{Pr}(y = 1 | X) = \frac{1}{1 + \exp(-w^T x)} \]

\[ w^T x = w_3 M + w_{340} M + w_{3SC2} M + w_{3PF1} M + w_{3PF3} M \]
Demography

AD response

Subscription

AD

Primitive Embedding

Pooling

Deep & Cross Embedding

Prediction

Pr(Y = 1 | X)
Experimental Design & Considerations

Model Structure
- Hidden layers and nodes
  - 2 hidden layers
- Regularization
  - Dropout
  - Batch normalization
  - L2-norm
- Activation function
  - Relu
  - Elu
  - PRelu

Feature Embedding
- One-hot encoding dimension
  - Dim = k * N^{1/4} ***
  - k = 1 (e.g., 1M ≈ 32)
- Pooling of multi-hot features
  - Sum
  - Average
  - Max
  & more…

*** https://www.tensorflow.org/guide/feature_columns
Pros

• Higher prediction accuracy
• One-shot feature embedding while training

Cons

• Tough real-time update for rare cold-start data
• Worse calibration
  - Sensitive to hyper-parameter and training data
• Less interpretability
• Long serving latency → Candidate selection
Conversion Prediction
conversion is all we need
Impression: Branding, inventory, CPT/CPM

Click: Traffic, audience, CPC

Conversion: Purchase, right audience, CPA/CPS/AutoBid
$D = D_0 e^{-\lambda t}$

Survival Model

$$\Pr(Y = 1 \mid X) \iff \Pr(Y = 1 \mid X, D) * \Pr(D \mid X)$$
Dynamic Creative Optimization (DCO)

creatives need to be creative
Gift for YOU

It’s Travel Time

Buy one get one free

Shop Now

Congratulations!

Refresh yourself.

Booking

Happy birthday

Purchase
**Interests:** Travel, Photography

**Current location:** Seoul, Korea

**Recent query**
- Flight fare to Spain
- Barcelona tour spots
- La Liga fixtures

**Recent shopping**
- Soccer shoes, training ball

**Recent readings (KWD)**
- Soccer
- Lionel Messi
- Frenkie de Jong

**Airline, Travel Agency**

**Origin:** Incheon, ROK

**Destination:** Barcelona, ESP

**Content:** FC Barcelona (Soccer)

**Incheon to Barcelona, Spain**
Staring from 999,000 KRW

Visit & watch Messi’s play

**Book Now**

BAHN AIRLINE
AutoBid

show me the money
Daily traffic: 1,000,000
Avg(eCPM): 2,000
Conversion/Traffic: 0.01%

Daily budget: 1,000,000
Avg(pCTR): 1%

BA_{cpc}: 100? 200? 500?
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>BidAmount (BA)</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>pCTR</td>
<td>1% (0.01)</td>
<td>1% (0.01)</td>
</tr>
<tr>
<td>eCPM</td>
<td>1,000</td>
<td>5,000</td>
</tr>
<tr>
<td></td>
<td>$1,000 = 1,000 * 100 * 0.01$</td>
<td>$5,000 = 1,000 * 500 * 0.01$</td>
</tr>
<tr>
<td>[Avg. eCPM = 2,000]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected WinRate</td>
<td>10%</td>
<td>90%</td>
</tr>
<tr>
<td>Expected impression (Traffic * winRate)</td>
<td>100,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Spending (Budget: 1,000,000)</td>
<td>100,000</td>
<td>4,500,000</td>
</tr>
<tr>
<td></td>
<td>$100,000 = 100,000 * 0.01 * 100$</td>
<td>$4,500,000 = 900,000 * 0.01 * 500$</td>
</tr>
</tbody>
</table>
What is the optimal BA?
BA = 200?
Landscape Forecasting

Budget Smoothing

Traffic Selection

Pacing & Control

Historical data, ARIMA, Prophet
QC / FC / SSP
sometimes, nothing is better
Quality Control: Cut-off low performing ADs
Frequency Capping: Relieve Ad fatigue

Cumulative Clicks

Overall CTR

Average CTR

CTR

# impressions

1 2 3 4 5 6 7 8 9 10

Frequency Capping: Relieve Ad fatigue
Negative Feedback

- Hide (Do Not Show Ads)
- AdBlock
- **DNT** (Do Not Track)
- ITP (Intelligent Tracking Prevention)

- NDNC (No Response)

- Abusing
Few more things...
Random bucket  
(Multi-armed bandit)  
Thompson sampling

It's Travel Time  
Refresh yourself. Booking

10% traffic  
90% traffic

make unstable to make stable
Data Overload & Imbalance

Millions of clicks over billions of impressions

\[ q = \frac{p}{p + \frac{1 - p}{\omega}} \]
Research

- Problem & ideation
- Model validity
  - Log-loss, RIG
  - Simulation
- Offline Test
- Online Test
  - Validity & revenue
  - CTR, calibration
  - 0 Bucket
- Production
  - Complexity & Stability

DCO
CS / DNN
DFM
LAL / LDA / AB / FC / …
- Main bucket (control group)
- Current serving version

- Identical model to main bucket
- To check the effect of serving bias
- Do not reject null hypothesis ($A = A'$)

- Test bucket (treatment group)
- 10% (up-to 50%, except random bucket)
- Hours to weeks

- B'?

- Buckets are randomly assigned to users or traffic.
- User-based buckets are periodically re-assigned.

- 5 ~ 10%
- Exploration (i.e., cold-start), serving-unbiased, reference (worst case)
A Data Scientist’s Happiness Circuit

Better Model → More Clicks → More Revenue → Incentive?
Revenue (B / Y: 99.01%)

Observed CTR (B / Y: 112.83%)

Predicted CTR (B / Y: 113.28%)

Calibration (103.4 vs 102.6)

Calibration = Predicted CTR / Observed CTR
Serving Latency

- Dimensionality reduction (& feature selection)
- Negative downsampling
- Candidate selection
- Simple model $\Rightarrow$ proper model
  - Simple structure & less layers/nodes
- Binary representation

- Go language
- Scale-out
- …
Simple rules still rule. Heuristic forever!
Data is number, but not just number. Devil’s in detail.
Bad things happen. It’s normal.
Without loss, we can’t learn. Random works.
Seeing is believing. Check, check, and check.
Production is Alpha and Omega. It’s speed, stupid.
Algorithm? Data (feature work) is the KING.
Question!

Q's will set you free.
Data science is an art of manipulating data.
A data scientist is not an artist, but a data butcher.
References
- LAL https://bit.ly/2QzBfOa
- Thompson sampling https://stanford.io/2MdTgOJ