An Ensemble-based Approach to Click-Through Rate Prediction for Promoted Listings

Kamelia Aryafar, Senior Data Scientist, @karyafar
Devin Guillory, Senior Data Scientist, @dguillory
Liangjie Hong, Head of Data Science, @lhong

August 2017
Takeaways

• Etsy’s Promoted Listings Product
• System Architecture and Pipeline
• Effective Prediction Algorithms and Modeling Techniques
• Discuss Correlations between offline experiments and online performance
Promoted Listings

Background
Etsy: Background

Etsy is a global marketplace where users buy and sell unique goods: handmade or vintage items, and craft supplies.

Currently Etsy has > 45M items from 2M sellers and 30M active buyers
Promoted Listings: Background
Promoted Listings: How it works

• Sellers specify overall Promoted Listings budget (optional max bid per listing)
• Sellers cannot choose which queries they want to bid on.
• CPC is determined by a generalized second price auction.¹

\[
Score_i = Bid(L_i, Q_j) \times ExpectedCTR(L_i, Q_j)
\]
Sellers pay minimum bid required to keep their position

\[
Score_i = Bid(L_i, Q_j) \times \text{ExpectedCTR}(L_i, Q_j)
\]

\[
\text{CPC}_i = \frac{Score_{r+1}}{Score} \times Bid_i
\]
CTR Prediction Overview
Promoted Listings: System Overview
Data Collection
CTR Prediction: Data Collection

- Training Data: 30 Days Promoted Listings Data
- Balanced Sampling
- Evaluation Data: Previous Day Promoted Listings Data
Model Training
CTR Prediction: Modeling

- $P(Y|X) = p(\text{click} \mid \text{ad}_i) \sim \text{Logistic Regression}$
- Single Box training via Vowpal Wabbit
- FTRL-Proximal Algorithm to learn weights

$$w_{t+1} = \arg\min_w \left( g_{1:t} \cdot w + \frac{1}{2} \sum_{s=1}^{t} \sigma_s \|w - w_s\|_2^2 + \lambda_1 \|w\|_1 \right)$$
Inference
CTR Prediction: Inference

Scalding

Logs/Visits
HDFS

Instance creation

Featurization

ssh/curl

Modeling

Single box model with vw

CTR model on HDFS

Inference With UIE

Ads EFF

Inference (with UIE)

CTR model on HDFS

Prediction set HDFS

Model Service

Scalding job
Mapppers

Ads EFF
CTR Prediction: Scaling

- Calibrate predictions due to Balanced Sampling
- Fit predictions to previous day’s distribution

\[
\theta_{t-1} = N(\mu_{t-1}, \sigma^2_{t-1})
\]

\[
\theta_t = N(\mu_t, \sigma^2_t)
\]

\[
S_i = (P_i - \mu_t) / (\sigma_t/\sigma_{t-1}) + \mu_{t-1}
\]
Evaluation
CTR Prediction: Offline Performance

- Models trained over days [t-32, t-2],
- Model Evaluated over t-1
- Key Metrics:
  - Area Under Curve (AUC)
  - Log Loss
  - Normalized Log Loss
Online Performance

- Tracking offline metrics established AUC as target metric
- Single digit improvements in AUC -> Single Digit improvement in CTR
Ensemble-Based Model
Featurization

- Historical Features - based on promoted listing search logs that record how users interact with each listing
- Content-Based Features - extracted from information presented in each listing’s page
**Featurization: Historical Features**

- Per Listing Historical Features:
  - Types: (Impressions, Clicks, Cart Adds)
  - Transformations:
    - Log-Scaling: $\text{floor}(\log(\text{impressions}))$
    - Beta Distribution Smoothing: $\frac{c_i + \alpha}{v_i + \alpha + \beta}$
Featurization: Contextual Features

- Per Listing Contextual Features:
  - Listing Id, Shop Id, Categorical Id
  - Text Features (Title, Tags, Description)
  - Price, Currency Code
  - Image Features (ResNet 101 embedding)
Data Exploration

Initial Insights

- *Historical Features* - performed highest for frequently occurring listings
- *Contextual Features* - performed highest for rarely presented listing
- What’s the best way to leverage this information to create an effective model?
Proposed Ensemble Model

Data splitting (Warm and Cold)

- Split training data into two cohorts > N and < N impressions, (N=30)
- Train separate models on each warm and cold cohort
- Ensemble models (Stacking) together in order to get best possible predictions
Primary Models

Instance
- Historical Features
- Contextual Features

Switch

Historical Model
- >N

Contextual Model
Primary Models

- Warm/Historical Model
  - Trained on high-frequency data
  - Uses Historical Features - Smoothed CTR

- Cold/Contextual Model
  - Trained on low-frequency data
  - Uses Contextual Features (Title, Tags, Images, Ids, Price)
Ensemble Layer

IC = Floor(Log(Impression Count))
Table 1: Changes in AUC (%), average impression log loss ($\times 10^3$) and normalized cross entropy ($\times 10^3$) on the dataset is compared to a numerical listing id only model, which was deployed as the production model, across different variants. The historical model utilized historical smoothed CTR for the logistic regression while the content-based model uses the multimodal listing features discussed in Section 3.3. The ensemble model is trained with content-based and historical models scores along with $[\log(\text{impression count})]$ as an input to the ensemble. The impressions break threshold for this experiment is set at $k = 30$ as discussed in Section 4.2.
Learned Attentions

Normalized Ensemble Model Weights at Different Frequency positions