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

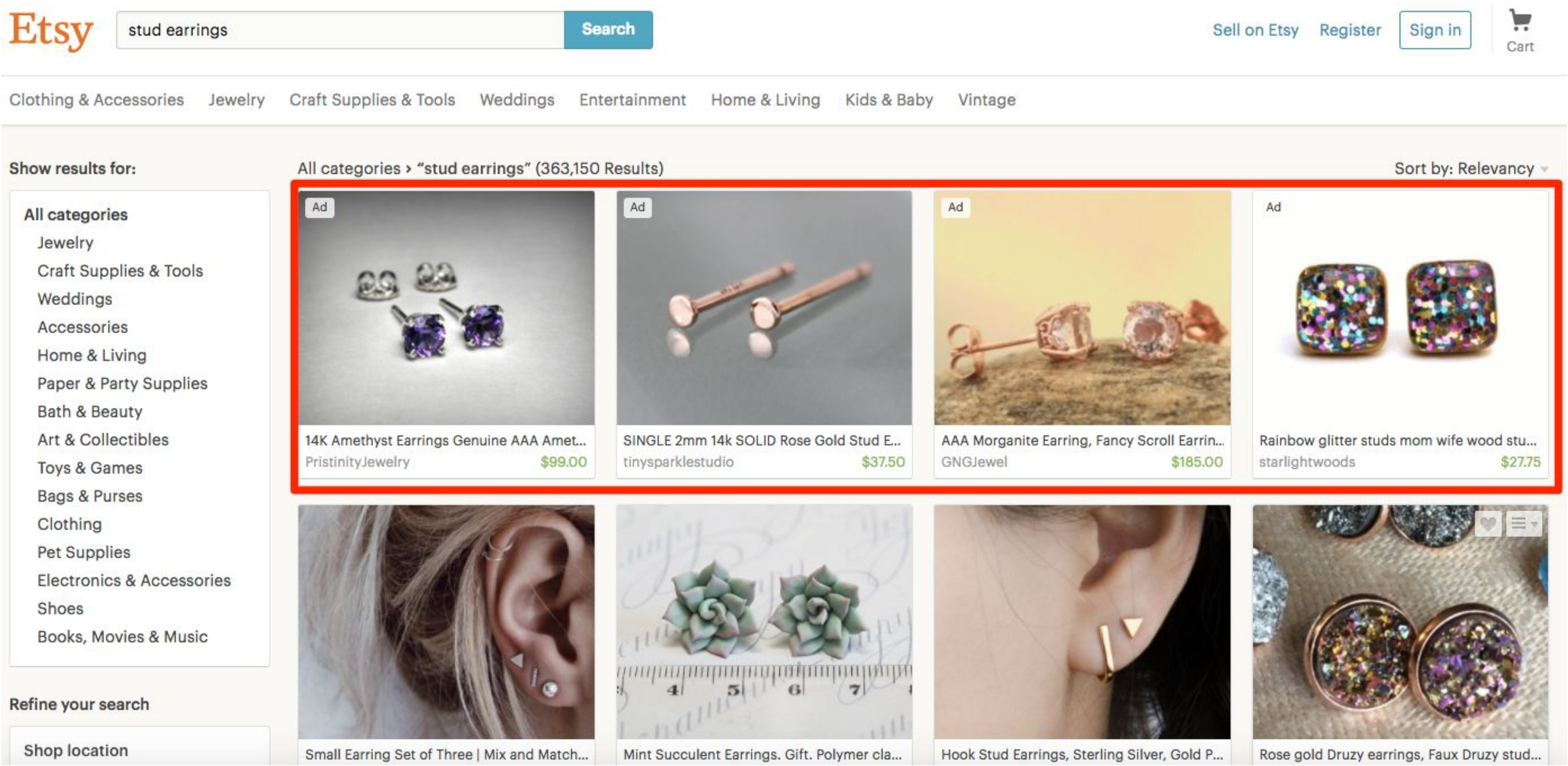
Background

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]
- $Score_i = Bid(L_i, Q_j) * ExpectedCTR(L_i, Q_j)$

Promoted Listings: Second Price Auction

Bridal Earrings Vintage, Wedding Earr..
Bid = 0.25
CTR = 0.158
Score = 0.0395
CPC = 0.13

Initial Stud Earrings A-Z, Personalized..
Bid = 0.95
CTR = 0.0202
Score = 0.01919
CPC = 0.94

Pava Crystal Ball Stud Earrings - Cryst..
Bid = 0.70
CTR = 0.0271
Score = 0.01897
CPC = 0.62

Vintage 18k Yellow Gold South Sea Pe.
Bid = 0.45
CTR = 0.0313
Score = 0.0168
CPC = 0.41

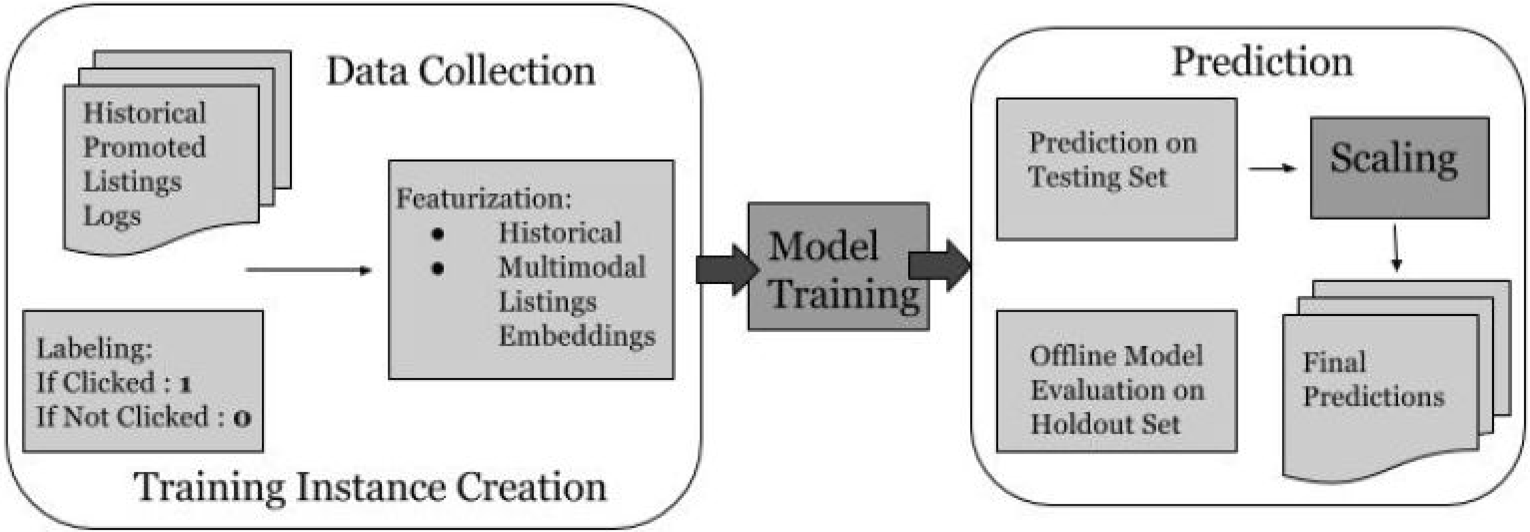
Sellers pay minimum bid required to keep their position

$$Score_i = Bid(L_i, Q_j) * ExpectedCTR(L_i, Q_j)$$

$$CPC_i = Score_{r+1} / Score * Bid_i$$

CTR Prediction Overview

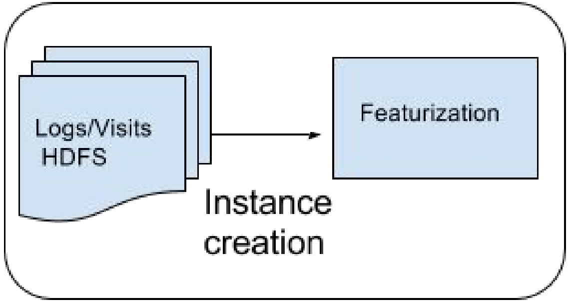
Promoted Listings: System Overview



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

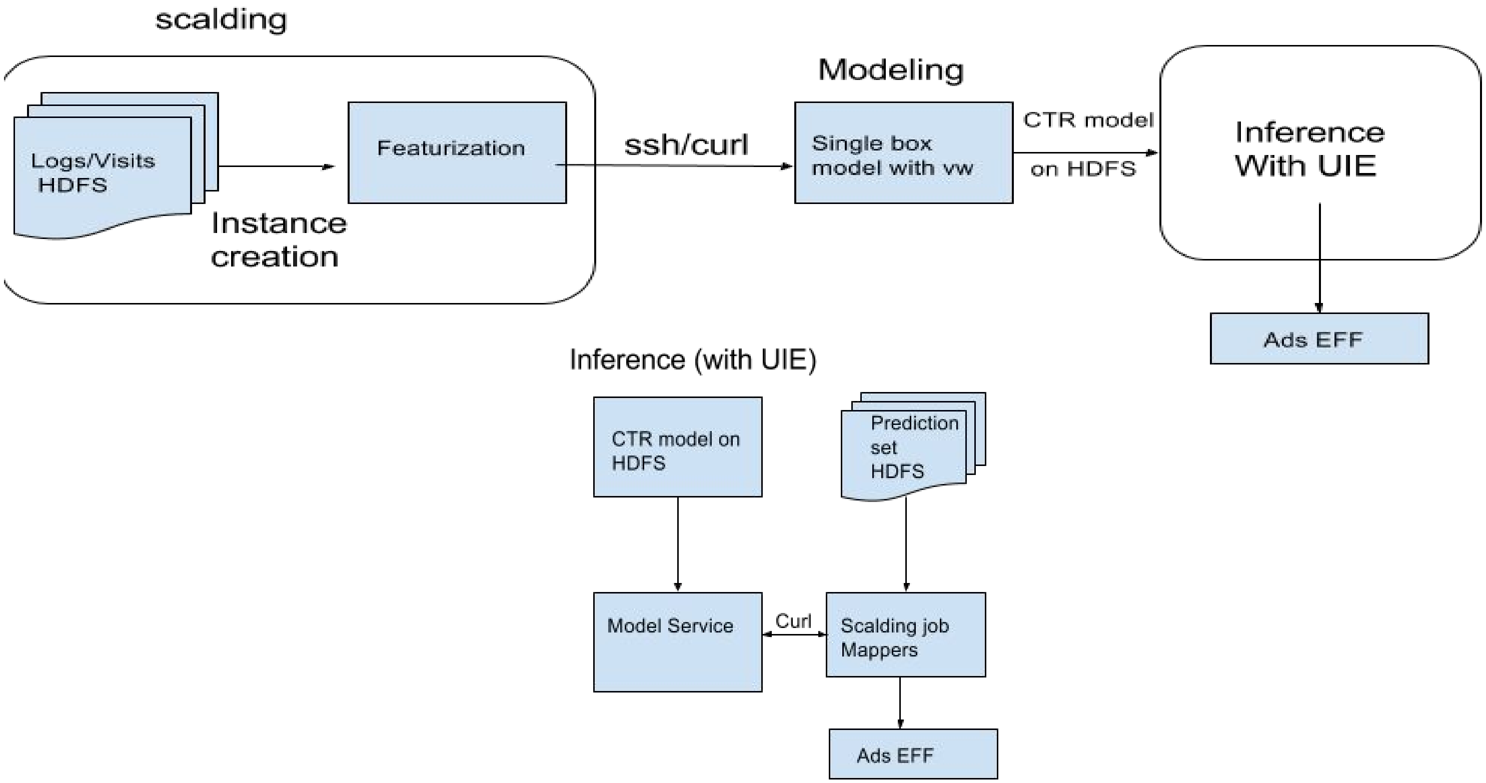
- $P(Y|X) = p(\text{click} \mid ad_i) \sim$ Logistic Regression
- Single Box training via Vowpal Wabbit
- FTRL-Proximal Algorithm to learn weights

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left(\mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2} \sum_{s=1}^{t} \sigma_s \|\mathbf{w} - \mathbf{w}_s\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 \right)$$

http://hunch.net/~vw/
H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. Ad Click Prediction: A View from the Trenches.

Inference

CTR Prediction: Inference



CTR Prediction: Scaling

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

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

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

$$S_i = (P_i - \mu_t) / (\sigma_t/\sigma_{t-1}) + \mu_{t-1}$$

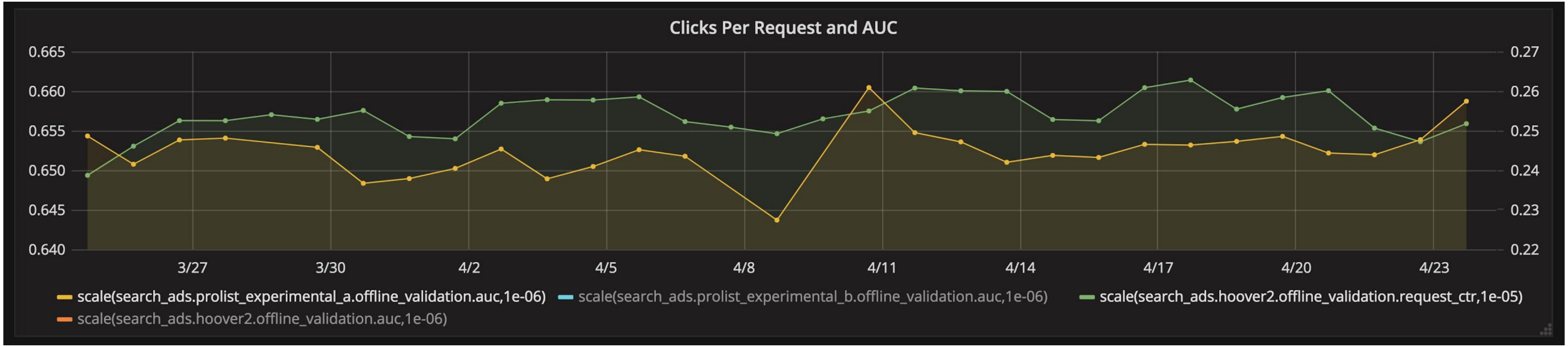
Evaluation

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 : $floor(log(impressions))$
    - 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)

Models & Performance

Data Exploration

Initial Insights

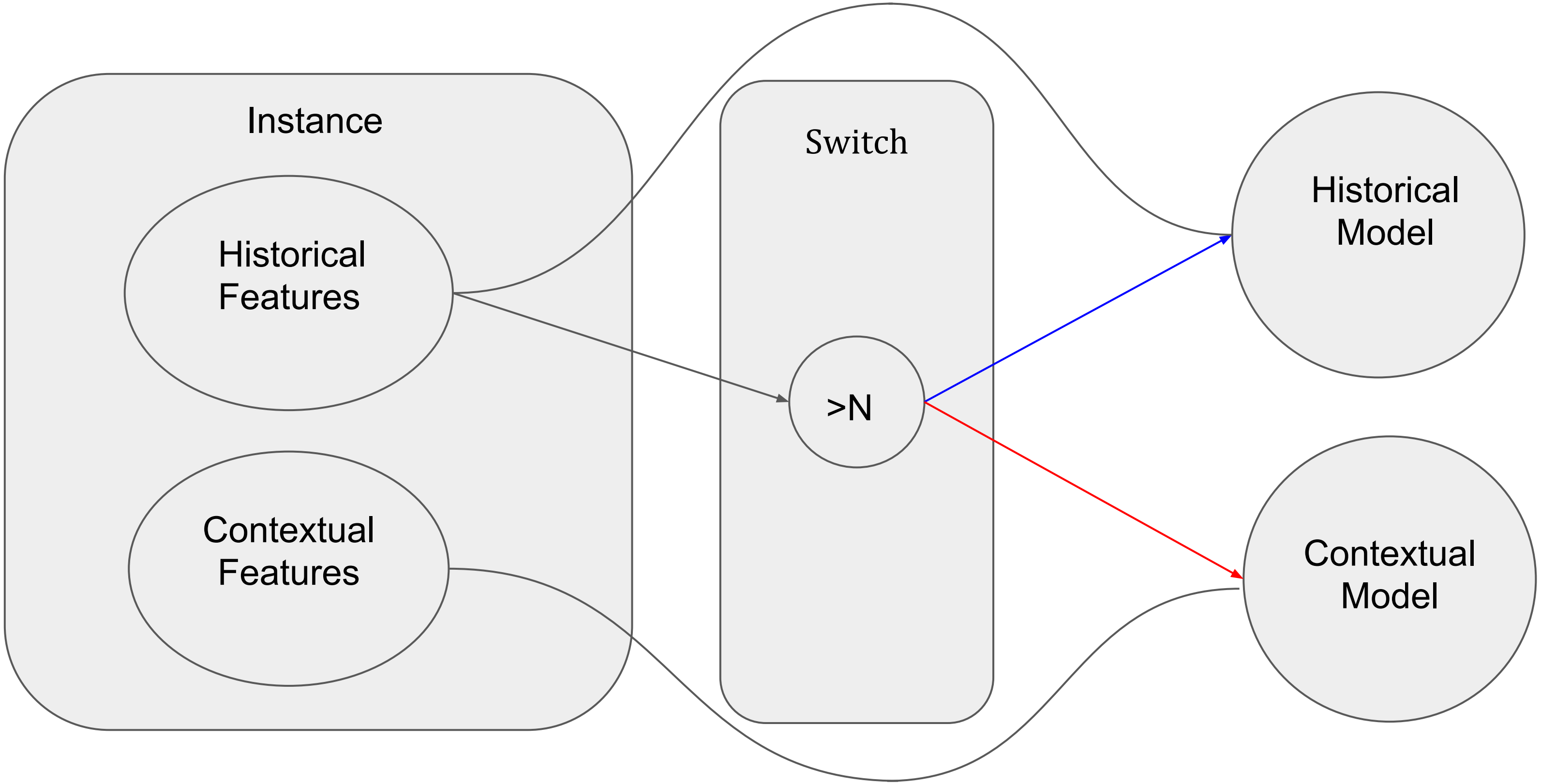
- *Historical Features* - performed highest for frequently occurring listings
- *Contextual Features* - performed highest for rarely presented listing
- What's the best way to leverage this information to create an effective model?

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

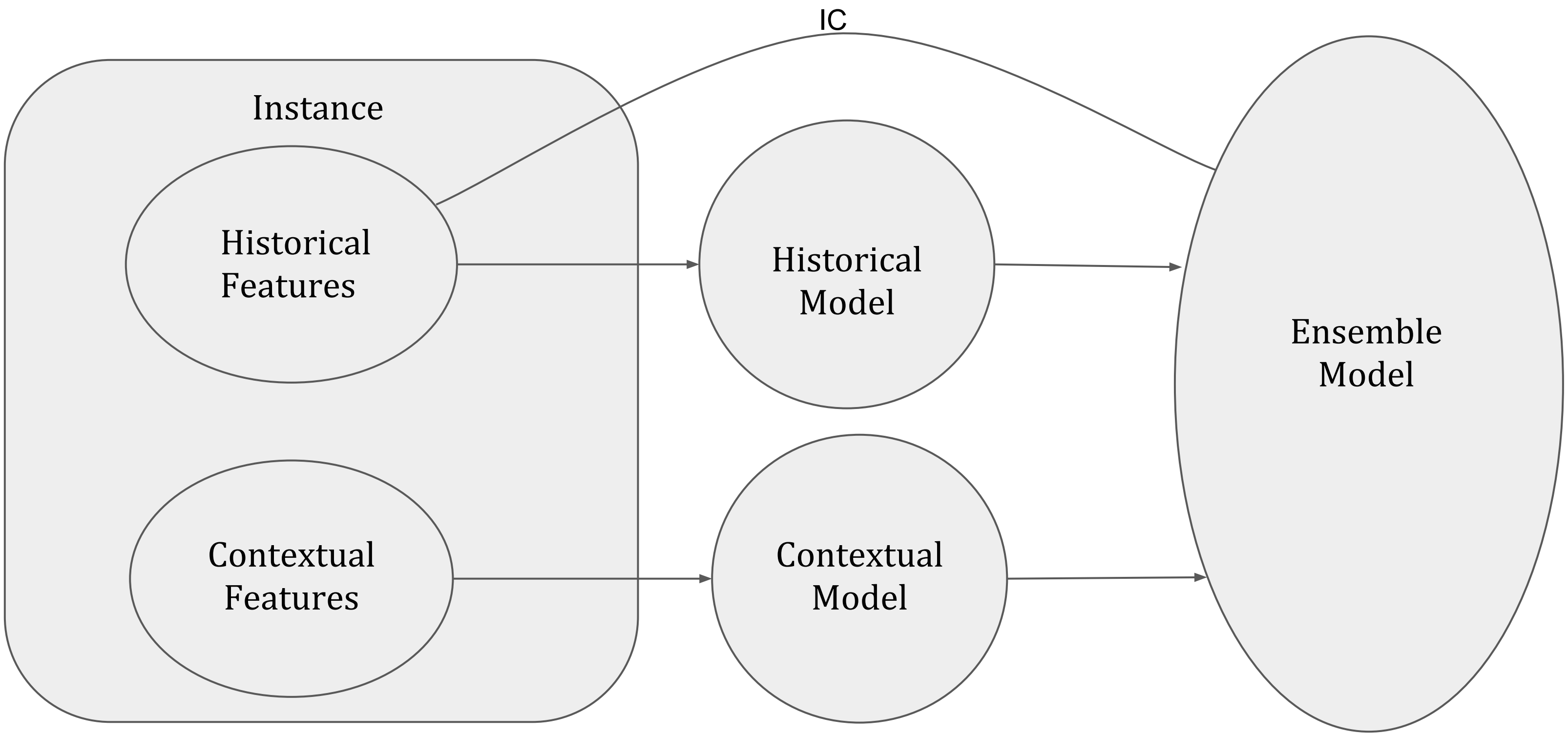


Primary Models

- 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))

Results

	Historical			Content-based			Ensemble		
	mixed	cold	warm	mixed	cold	warm	mixed	cold	warm
AUC (%)									
	$+1.56$	$+1.89$	$+1.55$	-1.57	$+6.39$	-1.74	$\mathbf{+1.95}$	$\mathbf{+8.34}$	$\mathbf{+1.83}$
Log Loss ($\times 10^3$)									
	-0.016	-0.048	-0.018	$+0.311$	-0.194	$+0.335$	$\mathbf{-0.092}$	$\mathbf{-0.332}$	$\mathbf{-0.087}$
Normalized Entropy ($\times 10^3$)									
	-0.29	-1.23	-0.31	$+5.67$	-5.00	$+6.01$	$\mathbf{-1.68}$	$\mathbf{-8.55}$	$\mathbf{-1.56}$

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 $\lfloor \log(impression_{count}) \rfloor$ as an input to the ensemble. The impressions break threshold for this experiment is set at $k = 30$ as discussed in Section 4.2.

Questions

Learned Attentions