Nested CTR-PTR Modeling for Sparse Purchase Prediction in Sponsored Search

Greg Kocher
eBay Inc.
San Jose, California, USA
gkocher@ebay.com

Abstract

Sponsored search ranking requires accurate and well-calibrated estimates of click and purchase probabilities. But these predictive tasks remain challenging because click and sale labels are sparse, and clicked-only post-click conversion modeling introduces sample selection bias. In this work, we exploit the simple but useful observation that for small probabilities, the logit of the probability is approximately equal to the log of the probability, which yields an additive logit-space decomposition of the purchase prediction task into a click-propensity term and a learned residual. We then introduce a simple, model agnostic, modular approach for overall $p(sale)$ prediction models by directly injecting $p(click)$ model outputs into a downstream $p(sale)$ model as a feature, a prior, or both, giving a flexible, additive alternative to the usual multiplicative click–conversion factorization for sparse purchase prediction. Training both click and sale models on the entire space of impression data mitigates sample selection bias while allowing the sale model to learn nonlinear interactions of click propensity with post-click behavior, rather than relying on a rigid multiplicative decomposition. This provides a practical and modular path to incorporate denser click signals into the sparse purchase prediction task, across many modeling architectures. Offline results at a major e-commerce company showed consistent gains in classification and calibration performance metrics; online A/B tests gave statistically significant gains in user and platform metrics; and a variant of the method was launched to production.

CCS Concepts

• **Information systems → Sponsored search advertising**; *Learning to rank*; Computational advertising.

Keywords

Search Ranking, Recommender Systems, Advertising, Conversion Rate Prediction

ACM Reference Format:
Greg Kocher. 2026. Nested CTR-PTR Modeling for Sparse Purchase Prediction in Sponsored Search. In *Proceedings of Workshop on AdKDD at KDD 2026 (AdKDD '26)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

AdKDD '26, Jeju, Korea

ACM ISBN 978-X-XXXX-XXXX-X/2026/08
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sponsored search sits at the intersection of search, recommendation, and advertising. Ranking policies for sponsored items must balance a three-way interaction of buyers, sellers, and the platform, and simultaneously optimize item relevance and quality, seller velocity, and platform revenue. Although many choices of ranking functions are possible, the ranking of Cost-per-Click (CPC) [4, 13] ad items typically involves estimating click-through-rate (CTR), i.e. $p(\text{click})$ [12]; while Cost-per-Acquisition (CPA) settings [9] usually require predicting purchase-through-rate (PTR), i.e. $p(\text{sale})$ [2, 5], either directly or via the multiplication of a CTR probability with a post-click conversion (CVR) probability.

PTR prediction is challenging for two related reasons. **(i) Data sparsity (DS):** purchase labels are naturally rare, since only a small fraction of item impressions lead directly to a sale. Click labels are less sparse but are a weaker signal. **(ii) Sample selection bias (SSB):** the standard method of estimating conversion conditional on click is trained on a clicked-only subset of impressions, but at serving time the model must score the entire impression population, causing an inference time mismatch. This problem motivated entire space and multi-task formulations such as ESMM [8], which models click and conversion jointly across the full impression space, and later extensions such as ESM^2 [15], which uses intermediate post-click actions to add further supervision. Subsequent work has continued in this direction: ESCM^2 [14] formalizes two pathologies in ESMM-style training, an *inherent estimation bias* (IEB), where ESMM's expected post-click CVR estimate is positively biased, and a *potential independence priority* (PIP) problem, in which ESMM is prone to modeling conversion as an unconditional CVR rather than conditioned on click, making the CTR and CVR estimates conditionally independent, and addresses them via a counterfactual risk regularizer; DCMT [18] then extends the debiasing into a counterfactual non-click subspace.

This paper studies a simple, model agnostic alternative. We use the observation that for small probabilities $\text{logit}(p) \approx \ln p$, so that the classical multiplicative click–conversion factorization is approximately equivalent to an *additive* decomposition in logit space. This motivates injecting an upstream CTR score into a downstream PTR model in three ways: as a feature, as a logit-space prior, or both. Training both models on the entire impression space avoids the most severe form of clicked-only sample selection bias, while the downstream PTR model is free to learn nonlinear corrections and interactions between click propensity and sale-specific variables such as user and item context. The approach is independent of the choice of underlying model class and applies wherever the CTR and PTR models can each produce and consume a logit score.

This design has several advantages. First, it is model agnostic, which is valuable since in many latency constrained settings, there can be stages of ranking and re-ranking using different model families with varying computational cost [6, 17]. Second, it is modular: the CTR and PTR models can use shared or disjoint feature sets and be trained following their own training pipelines with their own data, filtering, and sampling procedures. Third, it incorporates multi-task signals without requiring hybrid objective functions that need tuning of per-task weights to a combined loss function. Fourth, it avoids forcing the final purchase estimate to be an explicit product of separately estimated click and post-click conversion terms. Finally, it allows the PTR model to distinguish different regimes of click behavior—for example, curiosity driven *Exploratory Clicks* versus high purchase intent clicks—by conditioning on both the CTR signal and additional sale-relevant variables. In this way, the resulting model behaves like a flexible approximation to combined click–purchase modeling, and sidesteps the PIP problem [14], without requiring a full multi-task architecture which may not be possible for some stages of ranking which could still benefit from shared click and sale labels.

Our main empirical observation is that injecting CTR into PTR as a feature and/or logit-space prior can improve $p(\text{sale})$ prediction relative to standard baselines. Feature analyses further suggest that when CTR is included, downstream features associated with post-click conversion become more important, supporting the interpretation that the model is implicitly combining dense click supervision with sparse purchase information. Overall, the method offers a useful tradeoff: more interpretable and principled than ad-hoc stacking of upstream scores, but simpler and more modular than the end-to-end factorized multi-task frameworks (ESMM, ESM^2, ESCM^2, DCMT) it approximates.

The contributions of this work are as follows:

- We give a simple derivation, based on the approximation $\text{logit}(p) \approx \ln p$ for small probabilities, leading to an additive logit-space approximation of the classical multiplicative click–conversion factorization, $CTR \times CVR$.
- We introduce a nested CTR-PTR modeling framework in which a CTR score is injected into a sale model as a feature, as a logit-space prior, or both. The approach is model-agnostic and is a simple but effective approach for incorporating denser upstream click signals into sparser downstream $p(sale)$ modeling.
- Empirically, the method shows improvements in offline performance and statistically significant gains in online A/B test metrics, and a variant was deployed to production at eBay.
- We analyze how CTR injection changes the feature importances and feature usage patterns of the overall PTR model in intuitive ways.

2 Related Work

Entire Space and Factorized Click–Conversion Modeling. A major line of prior work addresses sparse purchase labels and clicked-only sample selection bias through entire space modeling, explicit click–conversion factorization, and multi-task learning. ESMM [8] is the canonical example: it jointly models CTR, post-click CVR, and their product over the full impression space, transferring signal from dense click labels to sparse conversion labels while avoiding training CVR only on clicked impressions. Later methods enhance this framework with additional structure. ESM^2 [15] decomposes the post-click path into intermediate actions such as add to cart and wish list events; AITM [16] uses sequential conversion-related tasks; and [3] divides conversions into hard vs. easy conversions which are treated by different model heads. Other recent work has focused on biases induced by the ESMM-style product architecture itself: ESCM^2 [14] identifies *inherent estimation bias* (IEB) in the CVR estimate and *potential independence priority* (PIP), in which the factorized structure treats conversion as independent of the click rather than conditioned on it, making the CTR and CVR estimates conditionally independent; DCMT [18] further debiases via a counterfactual mechanism. Our approach is motivated by the same problems, sparse purchase prediction and avoiding selection bias, but makes different design choices. We do not explicitly estimate CVR or enforce an exact product decomposition. Instead, the downstream PTR model predicts overall $p(\text{sale})$ directly over the impression space, while using CTR as an upstream feature, a logit-space prior, or both. This preserves the benefit of denser click supervision but avoids a constrained CVR head and allows the purchase model to learn arbitrary nonlinear, feature-dependent relationships between click propensity and sale propensity, all while staying well-calibrated with respect to the $p(\text{sale})$ task.

Multi-Task Learning, Representation Sharing, and Modular Transfer. Our method can also be viewed as a simple form of task transfer. Multi-task and representation-sharing systems in industry often use dense auxiliary tasks to improve sparse downstream objectives, both architecturally, e.g. via shared experts or gating mechanisms as in MMoE [7], and at the task-design level, where modeling additional conversion-related sub-tasks has been shown to improve purchase-intent prediction in production ad systems [3, 11]. Click and sale are correlated, and in our setting they cover the full range of the engagement spectrum: click is the densest but weakest available label, while sale is the most sparse but also strongest signal and is the modeling target. Intermediate post-click actions such as add-to-cart can contribute additional signal between these extremes, but incorporating clicks and sales may already capture the bulk of the dense-but-weak to sparse-but-strong intent spectrum.

Positioning of the Present Work. The prior literature on click–conversion modeling has mostly emphasized *exact probabilistic structure*: explicitly model click and conversion together through multiplicative factorization or multi-task learning to reduce bias and deal with sparsity. Our method instead uses the rare-event approximation $\text{logit}(p) \approx \ln p$ to convert this multiplicative structure into an approximate additive logit-space form (Section 3.3). The result is a simple injection of an upstream CTR score into a downstream sale model—as a feature, a logit-space prior, or both—that is interpretable, model-agnostic, and captures much of the benefit of joint click–conversion modeling without requiring a custom multi-task architecture. This positioning is practical in sponsored search settings where the overall goal is purchase prediction but can benefit from click signals, and click and sale models are often developed and maintained independently.

3 Methods

3.1 Background

Our goal is to estimate the overall purchase probability when showing an item impression i in a context featurized as x. That is, we want an accurate $p(\text{sale} \mid x)$. This quantity could be directly predicted over the entire space of item impressions, which is simple and avoids sample selection bias but omits valuable intermediate signals like clicks. Because sales are rarer than clicks, the direct sale prediction method suffers from worse label sparsity and class imbalance than the already imbalanced click prediction task.

3.2 Notation

Let the observed impression-level dataset of size N be

$$\mathcal{D} = \{(x_i, y_i \to z_i)\}_{i=1}^{N}$$

where $x \in \mathcal{X}$ denotes impression-level features, $y \in \{0,1\}$ is the click label, and $z \in \{0,1\}$ is the sale label. The notation $y \to z$ indicates the sequential dependence between click and sale: for our use case, every sale is assumed to come from a click on an item on a Search Results Page (SRP). This assumption may not be valid in all e-commerce settings, depending on the possible user shopping journey trajectories. When other sequences are possible, alternative factorizations of the shopping journey can be used as in [15] but $click \to sale$ suffices in our case.

With the standard click–conversion decomposition, this overall purchase-through-rate, p_{PTR} (i.e. sale probability), is factored as:

$$\underbrace{p(z=1 \mid x)}_{p_{\text{PTR}}} = \underbrace{p(y=1 \mid x)}_{p_{\text{CTR}}} \times \underbrace{p(z=1 \mid y=1, x)}_{p_{\text{CVR}}} \tag{1}$$

Here p_{CTR} is the click-through probability and p_{CVR} is the post-click sale probability. Standard post-click CVR methods model the second term directly using click-only data and combine it multiplicatively with CTR. In contrast, we aim to predict $p(\text{sale} \mid x)$ directly while still leveraging the information contained in the CTR term.

Let $\hat{p}_{\text{CTR}}(x)$ be the output of an upstream CTR model, and let

$$s_{\text{CTR}}(x) \equiv \text{logit}\big(\hat{p}_{\text{CTR}}(x)\big)$$

denote its logit. We feed this CTR score into a downstream PTR model in three ways: as a feature, as an additive logit-space prior, or both. This is broadly applicable across many model families: many classifiers naturally operate in logit space, so a logit-space prior can be added to the model's logit score for many common model classes like logistic regression, gradient-boosted decision trees (GBDT), and neural networks with a sigmoid output of a predicted logit.

3.3 Additive Logit-Space Decomposition

We introduce three variants below, all motivated by the same simple observation: for small probabilities, the multiplicative click–conversion factorization in Eq. (1) is approximately equivalent to an *additive* decomposition in logit space.

Taking logs of Eq. (1) gives an exact log-probability space identity:

$$\ln p_{\text{PTR}}(x) = \ln p_{\text{CTR}}(x) + \ln p_{\text{CVR}}(x) \tag{2}$$

For small probabilities, $\text{logit}(p) = \ln p - \ln(1-p) \approx \ln p$ since $\ln(1-p) \approx 0$. Click and sale probabilities in sponsored search are often on the order of 10^{-3} to 10^{-2} or smaller, so this rare-event approximation is tight in practice (see Appendix A). Substituting $\text{logit}(\cdot) \approx \ln(\cdot)$ in Eq. (2) gives

$$\text{logit}\big(p_{\text{PTR}}(x)\big) \approx \underbrace{s_{\text{CTR}}(x)}_{\text{CTR prior}} + \underbrace{g(x)}_{\text{learned residual}} \tag{3}$$

where $g(x)$ absorbs the post-click conversion term $\ln p_{\text{CVR}}(x)$ plus the small mismatch between log and logit parameterizations. A full derivation is given in Appendix A.

Equation (3) suggests a natural model design: rather than enforcing an explicit probabilistic factorization $p_{\text{PTR}} = p_{\text{CTR}} \cdot p_{\text{CVR}}$, we work directly in logit space and let a downstream PTR model learn $g(x)$, allowing the overall model to capture click information but still remain well-calibrated toward the sale target. The three variants below differ in how the CTR score $s_{\text{CTR}}(x)$ enters the PTR model. The prior variant relies on this approximation, which justifies adding $s_{\text{CTR}}(x)$ as a fixed logit-space offset, whereas the feature variant does not require it and instead passes $s_{\text{CTR}}(x)$ to a learned function that can model non-additive corrections even when click rates are higher. The combined variant uses both.

3.4 CTR as a Prior

In the first variant, the CTR score is added as a fixed offset on the logit of the PTR model, acting as a logit-space prior. The sale model is then learned as a residual correction on top of this click prior:

$$\underbrace{\text{logit}\big(p(z=1 \mid x)\big)}_{\text{sale logit}} = \underbrace{s_{\text{CTR}}(x)}_{\text{CTR prior}} + \underbrace{g_{\text{PTR}}(x)}_{\text{learned correction}} \tag{4}$$

This has a factorization-like interpretation in logit space: the CTR model provides the initial estimate, and the PTR model learns the adjustment needed to map click propensity into overall sale propensity. Conceptually, the learned correction acts similarly to a post-click conversion model (see Section 5). Also, note that the idea of adding a logit-space prior offset is common in practice, e.g. as a means of accounting for position bias [10].

3.5 CTR as a Feature

In the second variant, the CTR score is added as an explicit input feature, along with other features, to the downstream PTR model:

$$\underbrace{\text{logit}\big(p(z=1 \mid x)\big)}_{\text{sale logit}} = f_{\text{PTR}}\Big(x, \underbrace{s_{\text{CTR}}(x)}_{\text{CTR feature}}\Big) \tag{5}$$

Relative to the decomposition in Eq. (1), this replaces explicit multiplication by a learned nonlinear function, and relative to Eq. (4) the model can learn richer relationships between click propensity and other features, rather than just an additive offset. For example, the model can learn about *Exploratory Clicks* and distinguish when high click probability corresponds to genuine buying intent versus exploratory or low-intent behavior.

3.6 CTR as both a Prior and a Feature

In the third variant, the CTR output is used both as a feature and as the logit-space prior:

$$\underbrace{\text{logit}\big(p(z=1 \mid x)\big)}_{\text{sale logit}} = \underbrace{s_{\text{CTR}}(x)}_{\text{CTR prior}} + h_{\text{PTR}}\Big(x, \underbrace{s_{\text{CTR}}(x)}_{\text{CTR feature}}\Big) \tag{6}$$

This combines the two roles of the CTR signal. The logit-space prior gives a click-informed initialization, while the feature pathway allows the downstream model to learn context-dependent interactions between click propensity and other sale-specific variables.

4 Empirical Results

The proposed approach compared favorably to baseline alternatives in offline evaluation (Section 4.1) according to standard classification and calibration metrics, so the approach was further tested in online A/B tests on live eBay traffic (Section 4.2).

4.1 Offline Results

We assess offline performance on a held-out test set with >170M impressions, >10M clicks, and >400K sales. Table 1 summarizes the lift in several classification and calibration metrics, for the proposed approach vs. several baselines. All values are reported as relative percentage change with respect to the single-task $p(\text{sale})$ baseline model. The variants incorporating CTR score as either a prior, a feature, or both, improve over the production baseline according to many metrics like ROC-AUC, PR-AUC, and Relative Information Gain (RIG). Although the gains are modest, they are consistent across many metrics. Further, since the CTR model is itself a model that adds to the overall capacity, we compare against another baseline with equivalent overall model complexity (P(sale)–Large), and we see that our approach outperforms this baseline too. Finally, we compare against the standard multiplicative CTR × CVR factorization, and an alternative multi-task implementation (MTL–Joint) which is fit by optimizing a hybrid weighted click and sale objective. The Nested CTR-PTR variants outperform these baselines as well in most metrics. Note that Calibration Error, i.e. a binned 2-norm of prediction errors, is inherently small for all models, so very small absolute differences translate to large percentage differences. Notably, the multiplicative baseline has ROC-AUC and PR-AUC which move in opposite directions (+0.13% vs. -0.66%): it improves bulk ranking slightly while ranking actual sales worse than the direct single-task baseline. Our variants avoid this failure mode: small ROC-AUC gains but larger PR-AUC gains are consistent with cutting false positives from high-CTR, low-intent impressions (*Exploratory Clicks*) at the top of the ranking—where PR-AUC is sensitive and ROC-AUC, dominated by the abundant true negatives, is not.

We do not include direct comparisons against ESM², ESCM², or DCMT because these methods require additional supervision like intermediate post-click actions, counterfactual labels, or causal structure assumptions, which are less directly applicable in our setting. For example, ESM² targets settings where conversions are preceded by intermediate post-click actions such as add-to-cart or wishlist, and decomposes the conversion path through those signals; however, in our setting, the vast majority of sales follow directly from a click without intermediate actions, so the additional supervision ESM² relies on is not as helpful in our use case.

4.2 Online A/B Test Results

Based on the results of Section 4.1 above, we test Nested CTR-PTR variants in online A/B tests on several weeks of live session-randomized traffic. Table 2 summarizes key metrics. All values are reported as a relative percentage change with respect to the production baseline variant. For confidentiality, we anonymize exact metric names but we note that the *CTR as both Prior and Feature* variant in Table 2 has statistically significant improvements in user metrics related to purchase rates and platform metrics related to sales volume, and directionally positive improvements in several ad revenue metrics. We also note that although we ran multiple A/B tests with *CTR as a Feature* and each time it brought statistically significant gains in platform and user metrics, due to real-world testing constraints it was bundled together with other simultaneous improvements, so we do not report it here since it is not a pure ablation. However, we note that given the positive results we observed, a variant of Nested CTR-PTR was launched to production.

5 Discussion

Interpretability. Beyond the empirical performance gains from Nested CTR-PTR, it also leads to intuitive changes to the overall model structure. When comparing feature importances of the Nested CTR-PTR variants and baseline models, we see that for the *CTR as a Feature* case, the CTR feature becomes the second most important feature in the model; while in the *CTR as a Prior* case, we see that the highest importance features become those specifically associated with post-click CVR prediction. Intuitively, using the CTR as both a prior and a feature reduces the feature importance of the CTR model since some of its signal is captured already in the prior offset. Further, we see that the overall learned model is consistent with the concept of *Exploratory Clicks*, where an eye-catching item may attract attention in clicks but be unlikely to be purchased: a plot of the Accumulated Local Effects (ALE) [1] of the CTR feature within the PTR model is nearly monotonically increasing for much of the domain of the CTR score but once the predicted CTR reaches a certain threshold, the overall PTR model has a noticeable drop in its dependence on CTR. See Appendix B.

Relation to Entire Space Modeling. Next, we note that in practice, one potential downside of Nested CTR-PTR is that because it replaces the rigid multiplicative decomposition in favor of a more flexible approximation, it does *not* strictly enforce that for a given impression, $p_{\text{CTR}}(x) \geq p_{\text{PTR}}(x)$. This may be undesirable for certain applications, however, we note that since our intended task is sale prediction alone, this is not inherently a problem, and also, empirically, less than 0.005% of impressions showed this inversion phenomenon. Also, our approach has several other desirable properties from the Entire Space Modeling literature: for example, when using CTR as a feature, it sidesteps the PIP problem by allowing the overall model to learn rich dependencies between CTR and CVR tasks, rather than treating them independently and multiplying them to get a final PTR score.

Tradeoffs. The offline gains in Table 1 are similar in magnitude across the three variants, suggesting that a main driver of the improvement could be whether the dense click signal is injected at all, rather than the specific mechanism used to inject it. The method is thus a lightweight way to give a single-task $p(\text{sale})$ model much of the benefit of multi-task or task-transfer approaches in a simple, model agnostic way. This also allows for choosing between variants based on practical engineering tradeoffs, such as complexity and

Table 1: Offline performance on a held-out test set. Relative changes are with respect to the baseline PTR model.

Model	ROC-AUC	PR-AUC	RIG	Log Loss	Calibration Error
P(sale)–Baseline	0.00%	0.00%	0.00%	0.00%	0.00%
P(sale)–Large	+0.09%	+0.91%	+0.58%	-0.11%	+26.06%
Multiplicative (CTR × CVR)	+0.13%	-0.66%	+0.55%	-0.11%	-58.51%
MTL–Joint	+0.16%	+0.82%	+0.88%	-0.16%	+84.57%
CTR Prior Only	+0.27%	+1.57%	+1.54%	-0.30%	-11.17%
CTR Feature Only	+0.26%	+2.38%	+1.62%	-0.31%	-8.51%
CTR Prior + Feature	+0.29%	+2.28%	+1.80%	-0.34%	-43.62%

Table 2: Online A/B test results showing relative changes with respect to the baseline ranking policy.

Model	Revenue Metric 1	User Metric 1	Platform Metric 1
CTR Prior + Feature	+0.29% (p=.20)	+0.33% (p=.04)	+0.48% (p=.03)

latency: the additive logit form of the prior-only variant lets the CTR and overall PTR models be evaluated in parallel and summed, whereas the feature-only variant may be simpler conceptually and in implementation but could add latency if the CTR score is used directly as an input and both scores need to be computed in serial. However, for some model architectures, this variant may admit partial parallelization, e.g. as in a late-fusion design where a separate CTR path is combined near the output, or in some settings if a CTR score was already computed upstream, so reusing it adds little additional latency. Finally, although in principle the combined prior and feature variant may add the strongest click signal, in practice, having the click signal from just one pathway (either prior or feature alone) is simpler and achieves comparable performance.

6 Conclusion

We introduce a Nested CTR-PTR approach as a simple and effective means for incorporating denser click signals into sparse purchase prediction modeling. The method leads to an interpretable approximation to the standard PTR = CTR × CVR factorization and leads to intuitive feature usage patterns. It allows for modular, model agnostic development of CTR and PTR models, and has favorable properties from the Entire Space Modeling literature. Finally, it showed statistically significant gains and was launched to production at a major e-commerce company.

Declaration on Generative AI

The author used ChatGPT 5.4/5.5 and Claude Opus 4.6/4.7 to check grammar and spelling and to suggest refinements for some sentences. After using these tools, the author reviewed and edited the content as needed. All ideas and work are the author's own, and the author takes full responsibility for the content in this work.

References

[1] Daniel W. Apley and Jingyu Zhu. 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 4 (2020), 1059–1086. doi:10.1111/rssb.12377

[2] Olivier Chapelle. 2014. Modeling Delayed Feedback in Display Advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1097–1105. doi:10.1145/2623330.2623634

[3] Yuval Dishi, Ophir Friedler, Yonatan Karni, Natalia Silberstein, and Yulia Stolin. 2025. Practical Multi-Task Learning for Rare Conversions in Ad Tech. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. ACM, New York, NY, USA, 1042–1045. doi:10.1145/3705328.3748110

[4] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. *American Economic Review* 97, 1 (2007), 242–259. doi:10.1257/aer.97.1.242

[5] Zhen Ge, Wei Zhou, Jesse Lute, and Adam Ilardi. 2021. Relevance Constrained Re-ranking in Sponsored Listing Recommendations. http://papers.adkdd.org/2021/papers/adkdd21-ge-relevance.pdf AdKDD '21 workshop paper.

[6] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising (ADKDD'14)*. ACM, New York, NY, USA, Article 5, 9 pages. doi:10.1145/2648584.2648589

[7] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)*. ACM, New York, NY, USA, 1930–1939. doi:10.1145/3219819.3220007

[8] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 1137–1140. doi:10.1145/3209978.3210104

[9] Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. 2008. Dynamic Cost-Per-Action Mechanisms and Applications to Online Advertising. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 179–188. doi:10.1145/1367497.1367522

[10] Phuong Ha Nguyen, Djordje Gligorijevic, Hung Nguyen, Sheng Shen, Zhenke Xi, and Abraham Bagherjeiran. 2026. Mitigating Position Bias in Click Predictor Models: A Novel Downsampling Approach for Enhanced Accuracy and Efficiency. In *Computational Advertising (Communications in Computer and Information Science, Vol. 2796)*. Springer Nature Switzerland, Cham, 42–56. doi:10.1007/978-3-032-16358-5_4 AdKDD 2025.

[11] Junwei Pan, Yizhi Mao, Alfonso Lobos Ruiz, Yu Sun, and Aaron Flores. 2019. Predicting Different Types of Conversions with Multi-Task Learning in Online Advertising. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*. ACM, New York, NY, USA, 2689–2697. doi:10.1145/3292500.3330783

[12] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 521–530. doi:10.1145/1242572.1242643

[13] Hal R. Varian. 2007. Position Auctions. *International Journal of Industrial Organization* 25, 6 (2007), 1163–1178. doi:10.1016/j.ijindorg.2006.10.002

[14] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. ESCM^2: Entire Space Counterfactual Multi-Task Model for Post-Click Conversion Rate Estimation. In *Proceedings of the 45th*

International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). ACM, New York, NY, USA, 363–372. doi:10.1145/3477495.3531972

[15] Hong Wen, Jing Zhang, Yuan Wang, Fuyu Lv, Wentian Bao, Quan Lin, and Keping Yang. 2020. Entire Space Multi-Task Modeling via Post-Click Behavior Decomposition for Conversion Rate Prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, New York, NY, USA, 2377–2386. doi:10.1145/3397271.3401443

[16] Dongbo Xi, Zhen Chen, Peng Yan, Yinger Zhang, Yongchun Zhu, Fuzhen Zhuang, and Yu Chen. 2021. Modeling the Sequential Dependence among Audience Multi-step Conversions with Multi-task Learning in Targeted Display Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*. ACM, New York, NY, USA, 3745–3755. doi:10.1145/3447548.3467071

[17] Zhixuan Zhang, Yuheng Huang, Dan Ou, Sen Li, Longbin Li, Qingwen Liu, and Xiaoyi Zeng. 2023. Rethinking the Role of Pre-ranking in Large-scale E-Commerce Searching System. arXiv:2305.13647 [cs.IR] doi:10.48550/arXiv.2305.13647

[18] Feng Zhu, Mingjie Zhong, Xinxing Yang, Longfei Li, Lu Yu, Tiehua Zhang, Jun Zhou, Chaochao Chen, Fei Wu, Guanfeng Liu, and Yan Wang. 2023. DCMT: A Direct Entire-Space Causal Multi-Task Framework for Post-Click Conversion Estimation. In *Proceedings of the 2023 IEEE 39th International Conference on Data Engineering* (Anaheim, CA, USA) *(ICDE '23)*. IEEE, Piscataway, NJ, USA, 3113–3125. doi:10.1109/ICDE55515.2023.00239

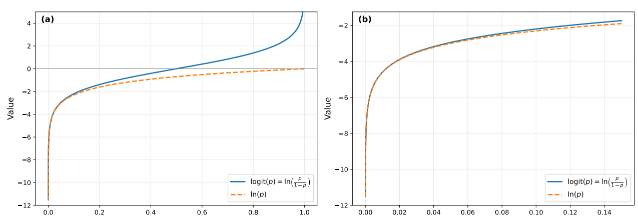

Figure 1: Comparison of $\mathrm{logit}(p)$ and $\ln(p)$. Panel (a) shows the full range $p \in (0,1)$, where the two functions are close for small p but diverge as p increases. Panel (b) shows the low probability regime up to $p = .15$, showing the closeness of the approximation at small probabilities.

Table 3: Comparison of $\mathrm{logit}(p)$ and $\ln(p)$ for small values of p. Values are rounded to 3 decimal places.

p	$\mathrm{logit}(p)$	$\ln(p)$	Abs. Diff.	% Diff.
0.001	-6.907	-6.908	0.001	0.014
0.005	-5.293	-5.298	0.005	0.095
0.010	-4.595	-4.605	0.010	0.219
0.030	-3.476	-3.507	0.030	0.876
0.050	-2.944	-2.996	0.051	1.742
0.100	-2.197	-2.303	0.105	4.795

A Relation Between Probability-Space and Logit-Space Factorizations

We derive the additive logit-space decomposition of Section 3.3. Starting from the standard click–conversion factorization of Eq. (1) and taking logs:

$$p_{\mathrm{PTR}}(x) = p_{\mathrm{CTR}}(x) \cdot p_{\mathrm{CVR}}(x) \tag{7}$$

$$\ln p_{\mathrm{PTR}}(x) = \ln p_{\mathrm{CTR}}(x) + \ln p_{\mathrm{CVR}}(x) \tag{8}$$

Use $\mathrm{logit}(p) = \ln p - \ln(1-p)$ and substitute in Eq. (8):

$$\begin{aligned}\mathrm{logit}\big(p_{\mathrm{PTR}}(x)\big) &= \ln p_{\mathrm{PTR}}(x) - \ln\big(1 - p_{\mathrm{PTR}}(x)\big) \\ &= \ln p_{\mathrm{CTR}}(x) + \ln p_{\mathrm{CVR}}(x) - \ln\big(1 - p_{\mathrm{PTR}}(x)\big)\end{aligned} \tag{9}$$

Write the CTR prior as $s_{\mathrm{CTR}}(x) \equiv \mathrm{logit}\big(p_{\mathrm{CTR}}(x)\big)$. Then

$$\ln p_{\mathrm{CTR}}(x) = s_{\mathrm{CTR}}(x) + \ln\big(1 - p_{\mathrm{CTR}}(x)\big)$$

and substituting into Eq. (9) gives:

$$\begin{aligned}\mathrm{logit}\big(p_{\mathrm{PTR}}(x)\big) &= s_{\mathrm{CTR}}(x) + \ln p_{\mathrm{CVR}}(x) \\ &\quad + \ln\big(1 - p_{\mathrm{CTR}}(x)\big) - \ln\big(1 - p_{\mathrm{PTR}}(x)\big)\end{aligned} \tag{10}$$

The prior-based form of Eq. (4) can therefore be seen as a case of

$$\mathrm{logit}\big(p_{\mathrm{PTR}}(x)\big) = s_{\mathrm{CTR}}(x) + g(x) \tag{11}$$

where $g(x)$ absorbs both the post-click conversion contribution $\ln p_{\mathrm{CVR}}(x)$ and the log-vs-logit mismatch $\ln(1-p_{\mathrm{CTR}}) - \ln(1-p_{\mathrm{PTR}})$. In the rare event setting $p \ll 1$ so $\ln(1-p) \approx 0$, so $\mathrm{logit}(p) \approx \ln p$ and the log-vs-logit mismatch shrinks:

$$\mathrm{logit}\big(p_{\mathrm{PTR}}(x)\big) \approx \ln p_{\mathrm{CTR}}(x) + \ln p_{\mathrm{CVR}}(x) \approx s_{\mathrm{CTR}}(x) + \ln p_{\mathrm{CVR}}(x) \tag{12}$$

This recovers the form in Eq. (3) and explains why an additive logit-space model behaves similarly to the multiplicative factorization when the target event is sufficiently rare. Table 3 and Figure 1 show the $\mathrm{logit}(p) \approx \ln p$ approximation error of small probabilities at the scale of p_{PTR} to p_{CTR}.

Further, the approximation error also has a simple closed form. Since $\mathrm{logit}(p) - \ln p = -\ln(1-p)$, a Taylor expansion gives

$$\mathrm{logit}(p) - \ln p \;=\; -\ln(1-p) \;=\; \sum_{n=1}^{\infty} \frac{p^n}{n} \;=\; p + \frac{p^2}{2} + O(p^3),$$

so the absolute error is $p + O(p^2)$. This quantity is exactly the *Abs. Diff.* column of Table 3.

B CTR Feature Usage Patterns

Figure 2 shows the Accumulated Local Effects (ALE) of the CTR feature within the PTR model, for the *CTR as a Feature* variant. The dependence increases for most of the CTR range but at the highest CTR values, where impressions are sparse and the confidence band widens, the effect flattens and dips slightly. This pattern is consistent with an *Exploratory Clicks* regime, although given the sparsity and wide confidence intervals at the high end, this relationship should be considered suggestive but not conclusive.

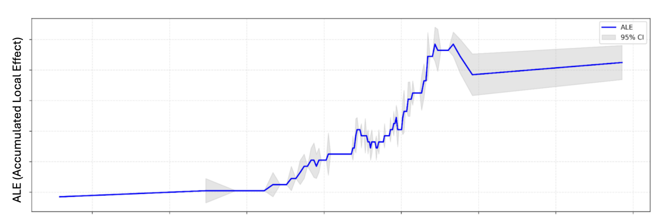

Figure 2: Accumulated Local Effects (ALE) of the CTR feature within the PTR model (*CTR as a Feature* variant).