Online Conversion Rate Prediction via Multi-Interval Screening and Synthesizing under Delayed Feedback

Qiming Liu[1,2], Xiang Ao[1,2,3]*, Yuyao Guo[1,2], Qing He[1,2]*

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, CAS, Beijing 100049, China
[3]Institute of Intelligent Computing Technology, Suzhou, CAS
{liuqiming21s, aoxiang, guoyuyao21s, heqing}@ict.ac.cn

Abstract

Due to the widespread adoption of the cost-per-action (CPA) display strategy that demands a real-time conversion rate prediction (CVR), delayed feedback is becoming one of the major challenges in online advertising. As the true labels of a significant quantity of samples are only available after long delays, the observed training data are usually biased, harming the performance of models. Recent studies show integrating models with varying waiting windows to observe true labels is beneficial, but the aggregation framework remains far from reaching a consensus. In this work, we propose the Multi-Interval Screening and Synthesizing model (MISS for short) for online CVR prediction. We first design a multi-interval screening model with various output heads to produce accurate and distinctive estimates. Then a light-weight synthesizing model with an assembled training pipeline is applied to thoroughly exploit the knowledge and relationship among heads, obtaining reliable predictions. Extensive experiments on two real-world advertising datasets validate the effectiveness of our model.

Introduction

In the advertising market, advertisers purchase ads via real-time bidding platforms with various paying options such as cost-per-click (CPC) and cost-per-action (CPA) (Guo et al. 2023; Hojjat et al. 2017; Zhang, Yuan, and Wang 2014; Liu et al. 2023). Cost-per-action, which enables advisers to bid on pre-defined conversions, e.g., purchase or registration, has become the primary objective due to its strong connections to the final return and resistance to notorious frauds (Chapelle, Manavoglu, and Rosales 2014; Goldfarb and Tucker 2011). Therefore, a precise estimation of conversion rate (CVR) becomes a critical demand for all advertising platforms. Especially, online systems require a streaming serving paradigm that continuously predicts and learns with the latest data (He et al. 2016; Guo et al. 2022).

As a consequence, the delayed feedback problem is becoming one of the imperative challenges. Concretely, after an ad is clicked, it takes a delay ranging from several seconds to a few days to receive the corresponding conversion. For example, Figure 1 exhibits the feedback proportion w.r.t the delayed time on the Criteo dataset. Only about 60 percents of conversions happen in the first day after clicks, while the longest delay between the click and conversion could reach 30 days. Online models that wait for 30 days to train would suffer severely from data staleness. On the other hand, a shorter wait time would introduce fake negatives that are temporarily observed as negative but may convert later, which affect the label accuracy of training.

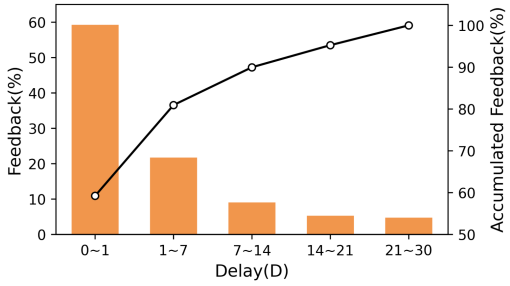

Figure 1: Conversion distribution (bar) and accumulated conversion distribution (line) of Criteo dataset.

In previous studies, various strategies were put out to address the delayed feedback issue. Early strategies estimate the anticipated conversion delay via jointly trained models for accurate CVR prediction (Chapelle 2014; Yoshikawa and Imai 2018), but they require lots of offline data and are not suited for streaming deployment. As for online learning, custom approaches choose a proper wait interval, known as the waiting window, to wait and observe samples (Yang et al. 2021), achieving a trade-off between data freshness and label accuracy. Several methods utilize no or one short waiting window and rely on importance sampling (Bottou et al. 2013) to adjust the weights of samples, making fake negatives less influential. Despite their success, these methods still suffer from highly biased training data with Missing-Not-At-Random problems, resulting in sub-optimal performance (Chen et al. 2022; Gao and Yang 2022).

The delayed feedback provides samples with an extra dimension of waiting time. Simply determining a waiting interval and screening out one distinctive data slice is inadequate. Recently, merging models with different waiting windows becomes one alternative solution for the delayed feedback issue (Gao and Yang 2022; Hou et al. 2021;

*Correspondence to Xiang Ao and Qing He.

Li et al. 2021). Multi-task approaches categorize conversions into various bins and model them separately (Gao and Yang 2022; Wang et al. 2020). Several methods predict CVR in various waiting windows to help update the final estimation by conditional probability (Gu et al. 2021; Hou et al. 2021). Waiting windows stand for unique observation views to balance label accuracy and data freshness. Therefore, these models based on difference waiting windows can give distinctive predictions and are naturally suitable for ensemble methods. (Li et al. 2021) comes up with an objective-oriented way to aggregate models, but the aggregation framework and effective fusion strategy remains far from reaching a consensus.

In this paper, we propose an approach named MISS (short for the Multi-Interval Screening and Synthesizing models) for online CVR prediction. MISS provides a universal framework for measuring CVR at unique views and aggregating them effectively. First, a multi-head screening model is devised to estimate CVR under various waiting windows, with different optimization tasks for each head. Additionally, a global weighting method is used to increase the accuracy of heads while preserving their individuality. The relationships among the heads are then investigated using a light-weight synthesizing model with normalization. The synthesizing model gives dynamic weights to aggregate predictions and is trained on an assembled pipeline comprising the most recent positive samples and real negatives. Through this way, our model simultaneously replicates the ideal unbiased distribution and enables the data freshness. Experiments on two widely-used benchmarks show MISS significantly outperforms existing methods.

The main contributions are summarized as follows:

- We underline the significance of incorporating models with various waiting windows to screen out abundant information in the online CVR task, which offers unique perspectives and forms accurate predictions.
- We propose MISS, which offers a general aggregation framework that exploits information behind various waiting windows. MISS also decreases the bias of integrated models while maintaining their uniqueness.
- We conduct extensive experiments on two real-world CVR prediction datasets and demonstrate the state-of-the-art performance of our method. Notably, MISS significantly outperforms other multi-task approaches.

Related Work

Delayed Feedback Models

Conversion modeling has been thoroughly studied in the literature for its high value in online advertising (Badanidiyuru et al. 2021; Choi et al. 2020; Lee et al. 2012). Under the hypothesis of an exponential delay distribution, Chapelle (Chapelle 2014) presented DFM, which contains two jointly-trained models for estimating the CVR and the delay time, respectively. While an exponential distribution is not necessarily practical in real scenarios, DFM was later evolved into a non-parametric delayed feedback model NoDeF (Yoshikawa and Imai 2018) without any assumptions about parametric distributions. Lately, studies like GDFM (Yang and Zhan 2022) focused on specific scenarios with user behaviors and use such auxiliary sequence information to assist training (Su et al. 2021). Multi-task models (Gao and Yang 2022; Hou et al. 2021; Huangfu et al. 2022; Wang et al. 2020) can make use of conversions by categorizing them into different bins based on their delay time. (Li et al. 2021) developed a multi-head framework FTP to model conversions in various delay settings and aggregate head outputs by imitating an ideal CVR model. Existing multi-task models employ head outputs directly as predictions or auxiliary knowledge, whereas our MISS utilizes a synthesizing method to further explore the data pipeline characteristics of each head.

Due to their capacity to infer real data distribution from observed biased data, unbiased estimate models gained prominence in online CVR prediction (Gu et al. 2021; Ktena et al. 2019). Existing methods could create a unique data pipeline with a well-designed importance weighting (Bottou et al. 2013) formula to regulate the weight of each sample, yielding an unbiased estimate of conversions (Yasui et al. 2020). To name a couple, the FNW (Ktena et al. 2019) method would label all new samples as negative and duplicate delayed positive samples with a corrected label when conversions arrive. To minimize fake negatives, ES-DFM (Yang et al. 2021) establishes a waiting window and only duplicates positive samples that do not receive conversion within the waiting period. Recently, (Chen et al. 2022) proposed DEFUSE, which further separates observed samples into four groups with different importance weights. While the majority of unbiased approaches concentrate on sampling strategy designs and add additional models to precisely determine every sample weight, MISS applies a simple, low-cost way instead to reduce global bias.

Preliminary

Data Pipeline

In this work, we focus on the online CVR prediction problem. At time τ, the ground-truth dataset can be formulated as:

$$\hat{\mathcal{D}}_\tau = \{(c_i, v_i, x_i, \hat{y}_i)\}_{i=1}^{N_\tau}, \tag{1}$$

where a sample contains c_i the timestamp when it is clicked, v_i the timestamp when a conversion action happens, x_i the feature, and $\hat{y}_i \in (0, 1)$ the ground-truth label indicating whether a conversion takes place in the end. $\mathcal{D}_\tau$ contains all samples that were clicked before time τ. Notice that the sample never has a conversion would be given $v_i \equiv \infty$, and only the conversion that happens in the maximum attribution time d_{max} would be regarded as valid. Thus we can define the label $\hat{y}_i$ as:

$$\hat{y}_i = \begin{cases} 1, & v_i \le c_i + d_{max}, \\ 0, & v_i > c_i + d_{max}. \end{cases} \tag{2}$$

However, due to the delayed feedback, it is unable to directly obtain the full ground-truth dataset. Models have to wait for at most a maximum attribution time d_{max} to see the real label of a sample after it gets clicked, the effect of data staleness is intolerable in online services. Most online models choose to wait for a shorter waiting window $d < d_{max}$ to

determine the label of each sample. We can define the training dataset they observed as:

$$\begin{aligned}\mathcal{D}_{\tau,d} &= \{(c_i, v_i, x_i, y_i)\}_{i=1}^{N_\tau}|_{c_i<\tau-d} \\ &= \{(c_i, v_i, x_i, y_i)\}_{i=1}^{N_{\tau-d}}.\end{aligned} \tag{3}$$

$\mathcal{D}_{\tau,d}$ limits the training data to a subset logged before $\tau-d$ to make sure label y of every sample is always available at time τ, which would be:

$$y_i = \begin{cases} 1, & v_i \le c_i + d, \\ 0, & v_i > c_i + d. \end{cases} \tag{4}$$

The observed real positive samples and real negative samples at time τ is:

$$\mathcal{P}_\tau = \hat{\mathcal{D}}_\tau|_{v_i \le \tau}, \tag{5}$$

$$\mathcal{N}_\tau = (\hat{\mathcal{D}}_\tau \setminus \mathcal{P}_\tau)|_{c_i \le \tau - d_{max}}. \tag{6}$$

Note that it takes a maximum attribution time d_{max} to identify real negative samples. Some real positive samples with a delay longer than d in $\mathcal{P}_\tau$ can be observed at time τ but were already falsely labeled as negative in $D_{\tau,d}$. To narrow the gap between the training dataset $D_{\tau,d}$ and the ground-truth dataset $\hat{\mathcal{D}}_\tau$, duplication mechanisms are applied to ingest those observed delayed positive samples with correct labels into training pipelines again. The adjusted training dataset would be defined as:

$$\mathcal{D}^+_{\tau,d} = \mathcal{P}_\tau|_{c_i+d<v_i} \cup \mathcal{D}_{\tau,d}. \tag{7}$$

Importance Sampling

The goal of the online CVR prediction problem is to continuously learn a function f with parameter θ that minimizes the following ideal loss:

$$\mathcal{L}_{\text{ideal}} = \sum_{(x_i,\hat{y}_i)\in\hat{\mathcal{D}}_\tau} \ell(\hat{y}_i, f_\theta(x_i)) = \mathbb{E}_{(x,y)\sim\hat{p}(x,y)}\ell(y, f_\theta(x)), \tag{8}$$

where ℓ denotes the binary cross-entropy loss, $\hat{p}$ is the ideal data distribution of $\hat{\mathcal{D}}_\tau$.

The altered training dataset $\mathcal{D}^+_{\tau,d}$ lacks freshness and contains fake negatives in comparison to the ideal dataset. Models trained on such biased data distributions can, nevertheless, approximate the distribution of ground-truth data via importance sampling (Bottou et al. 2013). We define q as the data distribution of the training dataset $\mathcal{D}^+_{\tau,d}$. Following previous work (Ktena et al. 2019; Yang et al. 2021), we assume $\hat{p}(x) \approx q(x)$ and derive the ideal loss as :

$$\begin{aligned}\mathcal{L}_{ideal} &= \mathbb{E}_{(x,y)\sim\hat{p}(x,y)}\ell(y, f_\theta(x)) \\ &= \int \hat{p}(x)dx \int \hat{p}(y|x)\ell(y, f_\theta(x))dy \\ &\approx \int q(x)dx \int q(y|x)\frac{\hat{p}(y|x)}{q(y|x)}\ell(y, f_\theta(x))dy \\ &\approx \mathbb{E}_{(x,y)\sim q(x,y)}\frac{\hat{p}(y|x)}{q(y|x)}\ell(y, f_\theta(x)) \\ &\approx \sum_{(x_i,y_i)\in\mathcal{D}^+_{\tau,d}} w(x_i, y_i)\ell(y_i, f_\theta(x_i)).\end{aligned} \tag{9}$$

By controlling the weight term $w(x, y)$ in Eq. (9), the bias of training on $\mathcal{D}^+_{\tau,d}$ could be reduced. Existing approaches utilize the output of their CVR model and extra models to calculate accurate weight $w(x, y)$ for each sample. In contrast, we present a light-weight technique to lessen the bias of heads globally while maintaining their distinction.

Methodology

In this part, we present our approach MISS to address the delay feedback issue. First, we introduce the multi-interval screening model in MISS, which includes shared neural network layers and a predefined number of unique output heads trained on data pipelines with various waiting windows. Then, adopting an assembled training pipeline, we demonstrate the synthesizing aggregation strategy to thoroughly use the knowledge of heads. Lastly, a low-cost method is adopted to globally enhance the weights of real positive samples, lowering the prediction bias brought on by the delayed feedback. Figure 2 illustrates the design of MISS.

Multi-Interval Screening Modeling

Recall that, the delay feedback problem leads to the change of sample labels and makes it difficult to learn the real distribution of online data. In recent years, it has been observed that integrating models with multiple waiting windows helps in addressing CVR prediction tasks with delayed feedback (Hou et al. 2021; Li et al. 2021). While shorter waiting windows would guarantee models to capture the most recent information, models with longer waiting windows are likely to have samples with accurate labels. Suppose the training dataset is $\mathcal{D}^+_{\tau,d}$, the length of d depends on the trade-off between label accuracy and data freshness, both of which are important to the performance of models.

We design the multi-interval screening model to balance the needs. The model consists of shared bottom layers, including an embedding layer and hidden layers, as well as numerous output heads that independently predict the probability of conversion. Concretely, we allocate different waiting windows $d_1, d_2, ..., d_N$ for the output heads $h_1, h_2, ..., h_N$ on the top of model. We assume that $d_{max} \ge d_1 > d_2 > ... > d_N > 0$. Each head h_i is training on its own data pipeline $\mathcal{D}^+_{\tau,d_i}$. The loss function for the multi-interval screening model would be:

$$\begin{aligned}\mathcal{L}_{\text{heads}} &= \sum_{1\le i\le N}\sum_{(x_j,y_j)\in\mathcal{D}^+_{\tau,d_i}} \ell(y_j, h_i(s(x_j))) \\ &= \sum_{1\le i\le N}\sum_{(x_j,y_j)\in\mathcal{D}^+_{\tau,d_i}} \ell(y_j, y_{h_i})),\end{aligned} \tag{10}$$

where s is the shared layers. The training gradients from each pipeline would only update the parameters of the corresponding output head and shared layers.

We employ the adjusted dataset with duplication techniques instead of the naive dataset $\mathcal{D}_{\tau,d_i}$ to train heads, thereby reducing the discrepancy between training data and real data. Additionally, real positives repeatedly update the weights of shared layers, reducing the impact of fake negatives. Heads with adjusted training data would produce more

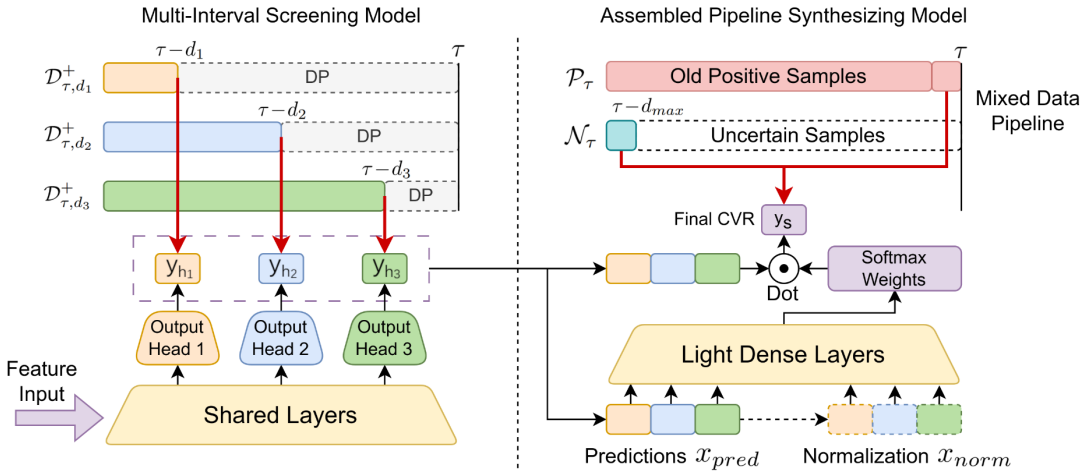

Figure 2: An illustration of MISS, including the multi-interval screening model, the assembled pipeline synthesizing model, and their distinct training data pipelines. The screening model has shared bottom layers and multiple heads training on various pipelines. Their predictions would be concatenated as the input of the synthesizing model and aggregate the final estimate. DP stands for delayed positive samples.

robust predictions. This is one of the major differences between our method and previous studies.

Assembled Pipeline Aggregation

The multi-head architecture and adjusted data help improve the accuracy of each prediction. The multiple waiting windows, however, still ensure that the heads would learn varied knowledge. Each head represents a unique degree of trade-off between introducing fake negatives and sacrificing data freshness, and thus has a distinct contribution to the final prediction. For instance, if the distribution of the data did not vary over time, the head with the longest waiting window would have the best prediction for adding fewer fake negatives. Otherwise, only the heads with short waiting windows could detect the trend in time and update their layers when a large number of rapid conversions arrived all at once. This example also indicates the importance to analyze the relationship of predictions, e.g., the maximum value. Because a large value predicted by a head with a short waiting window may suggest the arrival of instant conversions.

The diverse properties of heads make them naturally suitable for ensemble methods like bagging and stacking (Breiman 2004; Wolpert 1992), which concentrate on combining the output of a few models to produce more reliable estimations. Here, we utilize a light-weight model to generate dynamic weights for each head and perform aggregation. After the prediction of the multi-interval screening model, we concatenate the outputs from heads as the input x_{pred} for the synthesizing model. To compare the predictions from various heads and provide more information, we also apply normalization to generate extra input x_{norm}.

$$x_{pred} = [y_{h_1}, y_{h_2}, ..., y_{h_N}]_{concat}, \tag{11}$$

$$x_{norm} = [x_{pred}]_{norm}, \tag{12}$$

$$x = [x_{pred}, x_{norm}]_{concat}. \tag{13}$$

Instead of middle results from the embedding layer or hidden layers, we directly use the head predictions as input, which decreases model complexity without losing valuable information. We validate the contribution of head predictions and middle results input in ablation study. We apply dense layers with small sizes, as well as a softmax activation function to generate a set of dynamic weights $w = [w_1, w_2, ..., w_N]$, and then produce the final estimate y_s.

$$y_s = \sum_{i=1}^{N} w_i \cdot y_{h_i}. \tag{14}$$

The synthesizing model requires a reliable training pipeline. Previous methods choose $\mathcal{D}_{\tau,d_{max}}$, guaranteeing label accuracy at the cost of training on old samples from timestamp $\tau - d_{max}$. Our method, however, develops an assembled data pipeline M_τ including the latest positive samples in P_τ and real negatives in N_τ. Negative samples in N_τ are the same with $\mathcal{D}_{\tau,d_{max}}$ as it requires a maximum attribution time d_{max} to find them. On the contrary, positive samples get confirmed once they receive conversions. Therefore, we use the latest converted samples from P_τ to substitute the old positive samples from $\mathcal{D}_{\tau,d_{max}}$. Positive samples with a delay time d and a click timestamp at $\tau - d_{max}$ would be replaced by positive samples with the same delay time but were clicked at $\tau - d$ and converted at τ. Intuitively, the data freshness would be improved due to the ingestion of new samples. To further illustrate the effect of updating positive samples, we calculate the accumulated Kullback-Leibler divergence between the conversion distribution of ideal positive samples from D_τ, and distribution of old and latest positives from $D_{\tau,d_{max}}$ and P_τ respectively on Criteo. The results are shown in Figure 3. Obviously, the latest positive samples from P_τ have a closer distribution to the ideal one, leading to better performance of the synthesizing model.

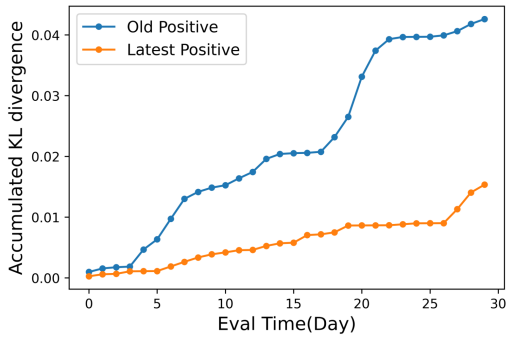

Figure 3: Accumulated KL divergence of positive samples in $\mathcal{D}_{\tau,d_{max}}$ and M_τ.

Global Positive Weighting

The synthesizing model above produces robust predictions by a dynamic weighting aggregation strategy, but there is still a risk of underestimating CVR. Note that, each head h_i from the multi-interval screening model is training on a pipeline $\mathcal{D}^+_{\tau,d_i}$. if $d_i < d_{max}$, extra fake negatives would decrease the global prediction values of heads. The weighted ensemble results from heads that underestimates CVR still suffer from the influence of fake negatives. To tackle such a problem, we utilize the importance sampling (Bottou et al. 2013) approach to globally amplify the weights of all positive samples. Following Eq. (9), the weights for positive samples from the pipeline for head h_i would be:

$$w(x, y, d) = \frac{\hat{p}(y = 1|x)}{q(y = 1|x)}, \tag{15}$$

where d is the delay time of the positive sample. $\mathcal{D}^+_{\tau,d_i}$ inserted duplicated positives to cover all the positives, we could normalize $\hat{p}(y = 1)$ to get $q(y = 1|x)$.

$$q(y = 1|x) = \frac{\hat{p}(y = 1|x)}{1 + \hat{p}(y = 1|x)\hat{p}(d > d_i|y = 1, x)}. \tag{16}$$

So the positive weights could be reformulated as:

$$\begin{aligned} w(x, y, d) &= 1 + \hat{p}(y = 1|x)\hat{p}(d > d_i|y = 1, x) \\ &= 1 + \alpha \frac{\sum_{(x_j,y_j,d_j)\in\mathcal{P}_\tau} \mathbb{1}(d_j > d)}{|\mathcal{P}_\tau|}. \end{aligned} \tag{17}$$

Recall that, positive samples from $\mathcal{P}_\tau$ make up a delay distribution relatively close to the ideal one, so we count the number of delayed positives among $\mathcal{P}_\tau$ to simulate $\hat{p}(d > d_i|y = 1, x)$. Moreover, as the ideal CVR prediction $\hat{p}(y = 1|x)$ could be difficult to calculate, we utilize a predefined hyperparameter $\alpha \in [0, 1]$ to replace it, reflecting the global CVR. A higher α would lead to a larger degree of amplification to the weights of positive samples. We use the same α for every head in our experiment for simplicity and convenience.

A short discussion. There are several differences between earlier studies and MISS in terms of weighted training. Previous methods focus on obtaining precise sample weights to enhance model performance. They train auxiliary models to assist in computing weight terms like $\hat{p}(d > d_i|y = 1, x)$ for each sample, respectively. However, excluding the extra cost of models, predicting those terms is as difficult as the CVR prediction itself. Contrarily, MISS only aims to lessen the global bias as our method relies on the following synthesizing model to increase accuracy. No extra models for weights calculation are used in MISS. Besides, the static setting of weights(e.g., the weights for positive samples from the same batch would be the same) avoids excessively affecting the ranking ability of heads, maintaining their distinction for subsequent aggregation.

Experiments

In this section, we first provide an overview of the design and implementation of experiments, and then validate our proposed model on two representative public advertising datasets, responding to the following research questions:

- **RQ1:** How does MISS perform in conversion rate prediction tasks, compared to the state-of-the-art models?
- **RQ2:** How does the global positive weighting help decrease the global bias?
- **RQ3:** How do adjusted datasets and the synthesizing model affect the performance of MISS respectively?

Datasets

The statistics of the two datasets are given in Table 1. We follow the original maximal attribution window setting of datasets in experiments.

Criteo Conversion Logs Criteo[1] is a widely used dataset for CVR prediction task (Chen et al. 2022). It contains 60 days data, with a 30 days attribution window d_{max}.

Tencent Advertising Algorithm Competition 2017 Tencent dataset[2] includes 9 days of data with a 5 days attribution window d_{max}. The dataset contains 22 million samples.

Experimental Settings

Online Simulation The online streaming process requires models to keep predicting CVR of new samples and then training on them. Following previous work (Chen et al. 2022; Yang et al. 2021), we separate the dataset into pretraining part and streaming part. Methods are allowed to use all the data from the former part to complete pretraining as they need. Then, models keep getting evaluated and updated hour by hour on the streaming part. The online training data only contains information available at the current timestamps. We adopted three distinctive evaluation metrics to evaluate the model performance, the area under the ROC curve (**AUC**), the negative log likelihood (**NLL**), and the precision-recall curve (**PR-AUC**).

Compared Baselines
Oracle: An ideal model training on dataset $\hat{D}_\tau$ with the ground-truth label, representing the upper bound.
Pretrain: A model training on pretraining data, representing the lower bound.

[1]https://labs.criteo.com/2013/12/conversion-logs-dataset/
[2]https://algo.qq.com/?lang=en

Dataset	# features	# conversions	# Samples	# average CVR	# log period	# attribution period
Criteo Dataset	17	3,619,793	15,898,863	0.2277	60 days	30 days
Tencent Dataset	19	624,411	22,601,402	0.0276	9 days	5 days

Table 1: Statistics of Criteo and Tencent dataset.

Datasets	Criteo Dataset			Tencent Dataset		
Metrics	AUC	NLL	PR-AUC	AUC	NLL	PR-AUC
Pretrain	0.0%(0.833)	0.0%(0.407)	0.0%(0.628)	0.0%(0.775)	0.0%(0.108)	0.0%(0.089)
Vanilla	27.5%	33.9%	26.5%	70.0%	60.3%	75.1%
FNW (Ktena et al. 2019)	46.0%	49.2%	20.1%	74.3%	70.7%	61.1%
FNC (Ktena et al. 2019)	44.9%	22.8%	7.4%	73.0%	67.2%	55.7%
ES-DFM (Yang et al. 2021)	64.6%	<u>74.0%</u>	66.4%	70.2%	65.5%	56.6%
DEFUSE (Chen et al. 2022)	<u>66.9%</u>	-19.3%	<u>66.8%</u>	70.0%	-70.7%	58.4%
MTDFM (Huangfu et al. 2022)	65.8%	56.8%	65.4%	75.6%	68.4%	70.8%
FTP (Li et al. 2021)	59.5%	50.8%	54.8%	<u>79.9%</u>	<u>72.4%</u>	<u>72.8%</u>
MISS	**83.7%***	**83.9%***	**78.1%***	**86.0%***	**82.8%***	**88.2%***
Oracle	100%(0.851)	100%(0.382)	100%(0.656)	100%(0.818)	100%(0.102)	100%(0.111)

Table 2: Performance comparisons of the proposed model with baseline models on AUC, NLL, and PR-AUC metrics. The Pretrain method and the Oracle method respectively correspond to 0% and 100%, their absolute performance is in parentheses. The best value in one column is displayed by the bold value, and the second is indicated by the underlined value. * indicates statistical significance improvement compared to the best baseline measured by t-test at p-value of 0.05.

Vanilla: A model training with a finetuned waiting window.
FNW (Ktena et al. 2019): A model training with a duplication mechanism and the fake negative weighted loss.
FNC (Ktena et al. 2019): A model training with a duplication mechanism and the fake negative calibration.
ES-DFM (Yang et al. 2021): A model training with a duplication mechanism using the ES-DFM loss.
DEFUSE (Chen et al. 2022): A model training with a duplication mechanism using the DEFUSE loss.
MTDFM (Huangfu et al. 2022): A two-task model training with a duplication mechanism.
FTP (Li et al. 2021): A model training with a multi-task learning mechanism and aggregation strategy.

We choose DEFUSE instead of Bi-DEFUSE (Chen et al. 2022) as the former achieved much better results on Criteo in the original paper. We also apply detailed comparison with existing multi-task approaches like FTP.

Parameter Settings We implement the MISS in Tensorflow and the source code will be available on GitHub[3]. A DNN model with a fixed hidden size (128,128) is used as the base model for all the methods. Each hidden layer is followed by the Leaky ReLU activation function (Maas et al. 2013). The synthesizing model for MISS only has one hidden layer with size [32]. L2 regularization is set to 10^{-6} on Criteo Dataset and 10^{-7} on Tencent Dataset. The models are updated by the Adam optimizer (Kingma and Ba 2014). For a fair comparison, we apply the grid search strategy to tune the best learning rate among {0.0001, 0.0005, 0.001}, and tune the waiting window for previous models in accordance with the original papers. MISS and FTP apply the same waiting windows, [1D, 7D, 14D, 21D, 30D] on Criteo and [1H, 6H, 24H, 48H, 120H] on Tencent.

[3]https://github.com/NealWalker/MISS

Main Experiments (RQ1)

We execute 5 random runs on Criteo and Tencent to illustrate the overall performance of MISS and baselines. The averaged results with significance test are given in Table 2. Following (Gu et al. 2021; Yang et al. 2021; Yang and Zhan 2022), we report the relative improvement to the performance gap between the Pretrain model and the Oracle model. We have the following findings upon their performance comparison:

- Our MISS method significantly performs better than the baselines on both datasets. In particular, MISS outperforms the strongest baselines w.r.t. relative-AUC by 16.8%, 6.1% on Criteo and Tencent datasets, respectively. Similar improvements are observed at NLL and PR-AUC metrics. Our strategy yields remarkable improvements because it applies multi-head design to learn distinctive distributions and provides a new, robust synthesizing model to aggregate exploitable information. Besides, compared with ES-DFM and DEFUSE which require a whole extra auxiliary model with a heavy embedding layer, few extra parameters introduced by MISS are totally affordable.
- Vanilla can perform better than expected if its waiting window is relatively long, e.g., 10 days, suggesting the value of maintaining a long waiting window to observe data, which is covered in MISS. By introducing importance sampling or calibration, FNW and FNC reach better performance. Such advantage is enlarged in ES-DFM by auxiliary models that calculate importance weights. MTDFM applies extra head for prediction calibration, achieving better AUC. DEFUSE comes up with a hidden model to precisely determine importance weight and reaches the best AUC and PR-AUC among baselines on

Metrics	AUC (↑)	NLL (↓)	PR-AUC (↑)
MISS	0.8477	0.3856	0.6495
MISS_O	0.8451	0.3944	0.6435
MISS_L	0.8390	0.3950	0.6382
MISS_A	0.8469	0.3860	0.6489
MISS_R	0.8459	0.3886	0.6430
MISS_H	0.8477	0.3857	0.6496

Table 3: Ablation study of MISS on Criteo.

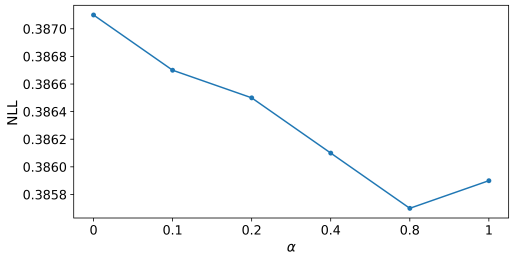

Figure 4: MISS trained with various α on Criteo

Criteo. However, sophisticated terms contained in its calculation cause a relatively high NLL.

- Tencent dataset represents the real industrial scenarios with massive data but scant feedback (an average conversion rate of 2.76%). Its low CVR and long-tail distribution weaken the strengths of weighting, leading to mediocre results for importance sampling methods. FTP and MISS, on the other hand, have lengthy waiting windows to model the distribution without adding many fake negatives. They do not rely on importance sampling and are resistant to scenarios with different CVR and distribution. Lastly, while the physical meaning of heads of FTP is the CVR restricted to various delay time windows, the meaning in MISS is the ground-truth CVR achieved by various sampling and weighting strategies, and the strategies are determined by various delay time. As a result, the FTP heads trained at samples from shorter waiting windows would inevitably underestimate the CVR, because these heads only observe part of the real positive samples. The synthesizing model with a specially designed pipeline and reliable heads helps MISS produce predictions pretty close to the ground-truth CVR and outperform FTP.

Ablation Study (RQ2)

In this section, we focus on ablation studies to validate the effect of adjusted datasets and the synthesizing model. We evaluate the following formulations of MISS on Criteo:
MISS_O: MISS without duplication mechanism.
MISS_L: MISS without the synthesizing model and outputting last head predictions.
MISS_A: MISS without the synthesizing model and output average head predictions.
MISS_R: MISS with a synthesizing model training on $\hat{D}_{\tau,d_{max}}$ instead.
MISS_H: MISS taking both head predictions and middle results as inputs for the synthesizing model.

The results are shown in Table 3. The removal of the duplication mechanism leads to a significant decrease at performance, indicating the importance to adjust the training data of heads when they are directly used for predictions. MISS_A outperforms MISS_L by a great margin for its naive aggregation. Such advantages are further developed by the dynamic aggregation strategy of the proposed model. MISS_R obtains mediocre results at PR-AUC and NLL. A synthesizing model trained on $\hat{D}_{\tau,d_{max}}$ gives dominant weights to the head with the longest waiting window for their similarity on training data, ignoring other heads. In contrast, our synthesizing model with an assembled pipeline could comprehensively exploit the value of every head. Finally, we include the middle results from the hidden layer of the model as additional inputs for the synthesizing model. The similar results suggest that the head predictions alone are sufficient and extra information is not necessary.

Global Positive Weighting (RQ3)

In global positive weighting, a hyper-parameter α controls the degree of positive weighting. Here we evaluate how the value of α influence the bias of our predictions. We use the NLL metric to illustrate the bias for its sensitivity to the absolute value of the prediction. The results of other metrics are omitted for similar trends. According to the results in Figure 4, with the increase of α, the NLL of MISS continues decreasing, suggesting predictions with higher accuracy and confidence are made. Notably, α replaces the ideal CVR prediction value in Eq. (17), but a value higher than the global average CVR could actually reach better performance. One explanation is that as we do not decrease the weights of negatives, excessive weights for positive samples could achieve a similar effect.

Conclusion

In this paper, we concentrated on the online CVR prediction task and proposed the MISS approach to deal with the issue of delayed feedback. We underline the value of integrating observations of various waiting windows and design a general framework to synthesize predictions by investigating their relationships on assembled unbiased data. MISS also decreases the bias of models by a universal weighting strategy with an assembled training pipeline. Experiments on two real-world datasets demonstrate the significance of our method.

Acknowledgements

The research work is supported by National Key R&D Plan No. 2022YFC3303302, the National Natural Science Foundation of China under Grant No.61976204, and the CAAI Huawei MindSpore Open Fund. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS and the Beijing Nova Program.

References

Badanidiyuru, A.; Evdokimov, A.; Krishnan, V.; Li, P.; Vonnegut, W.; and Wang, J. 2021. Handling many conversions per click in modeling delayed feedback. *arXiv preprint arXiv:2101.02284*.

Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14(11).

Breiman, L. 2004. Bagging predictors. *Machine Learning*, 24: 123–140.

Chapelle, O. 2014. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1097–1105.

Chapelle, O.; Manavoglu, E.; and Rosales, R. 2014. Simple and Scalable Response Prediction for Display Advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5: 1 – 34.

Chen, Y.; Jin, J.; Zhao, H.; Wang, P.; Liu, G.; Xu, J.; and Zheng, B. 2022. Asymptotically Unbiased Estimation for Delayed Feedback Modeling via Label Correction. In *Proceedings of the ACM Web Conference 2022*, 369–379.

Choi, Y.; Kwon, M.; Park, Y.; Oh, J.; and Kim, S. 2020. Delayed Feedback Model with Negative Binomial Regression for Multiple Conversions.

Gao, H.; and Yang, Y. 2022. Multi-Head Online Learning for Delayed Feedback Modeling. *arXiv preprint arXiv:2205.12406*.

Goldfarb, A.; and Tucker, C. 2011. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3): 389–404.

Gu, S.; Sheng, X.-R.; Fan, Y.; Zhou, G.; and Zhu, X. 2021. Real Negatives Matter: Continuous Training with Real Negatives for Delayed Feedback Modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2890–2898.

Guo, Y.; Ao, X.; Liu, Q.; and He, Q. 2023. Leveraging Post-Click User Behaviors for Calibrated Conversion Rate Prediction Under Delayed Feedback in Online Advertising. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Guo, Y.; Li, H.; Ao, X.; Lu, M.; Liu, D.; Xiao, L.; Jiang, J.; and He, Q. 2022. Calibrated Conversion Rate Prediction via Knowledge Distillation under Delayed Feedback in Online Advertising. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3983–3987.

He, X.; Zhang, H.; Kan, M.-Y.; and Chua, T.-S. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 549–558.

Hojjat, S. A.; Turner, J. G.; Cetintas, S.; and Yang, J. 2017. A Unified Framework for the Scheduling of Guaranteed Targeted Display Advertising Under Reach and Frequency Requirements. *Oper. Res.*, 65: 289–313.

Hou, Y.; Zhao, G.; Liu, C.; Zu, Z.; and Zhu, X. 2021. Conversion Prediction with Delayed Feedback: A Multi-task Learning Approach. In *2021 IEEE International Conference on Data Mining (ICDM)*, 191–199.

Huangfu, Z.; Zhang, G.-D.; Wu, Z.; Wu, Q.; Zhang, Z.; Gu, L.; Zhou, J.; and Gu, J. 2022. A Multi-Task Learning Approach for Delayed Feedback Modeling. In *Companion Proceedings of the Web Conference 2022*, WWW '22, 116–120.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Ktena, S. I.; Tejani, A.; Theis, L.; Myana, P. K.; Dilipkumar, D.; Huszár, F.; Yoo, S.; and Shi, W. 2019. Addressing delayed feedback for continuous training with neural networks in CTR prediction. In *Proceedings of the 13th ACM conference on recommender systems*, 187–195.

Lee, K.-c.; Orten, B.; Dasdan, A.; and Li, W. 2012. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 768–776.

Li, H.; Pan, F.; Ao, X.; Yang, Z.; Lu, M.; Pan, J.; Liu, D.; Xiao, L.; and He, Q. 2021. Follow the Prophet: Accurate Online Conversion Rate Prediction in the Face of Delayed Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Liu, Q.; Li, H.; Ao, X.; Guo, Y.; Dong, Z.; Zhang, R.; Chen, Q.; Tong, J.; and He, Q. 2023. Online Conversion Rate Prediction via Neural Satellite Networks in Delayed Feedback Advertising. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1406–1415.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3. Citeseer.

Su, Y.; Zhang, L.; Dai, Q.; Zhang, B.; Yan, J.; Wang, D.; Bao, Y.; Xu, S.; He, Y.; and Yan, W. 2021. An attention-based model for conversion rate prediction with delayed feedback via post-click calibration. In *Proceedings of the 29th IJCAI*, 3522–3528.

Wang, Y.; Zhang, J.; Da, Q.; and Zeng, A. 2020. Delayed feedback modeling for the entire space conversion rate prediction. *arXiv preprint arXiv:2011.11826*.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks*, 5: 241–259.

Yang, J.-Q.; Li, X.; Han, S.; Zhuang, T.; Zhan, D.-C.; Zeng, X.; and Tong, B. 2021. Capturing delayed feedback in conversion rate prediction via elapsed-time sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4582–4589.

Yang, J.-Q.; and Zhan, D.-C. 2022. Generalized Delayed Feedback Model with Post-Click Information in Recommender Systems. In *NeurIPS 2022*.

Yasui, S.; Morishita, G.; Komei, F.; and Shibata, M. 2020. A feedback shift correction in predicting conversion rates under delayed feedback. In *Proceedings of The Web Conference 2020*, 2740–2746.

Yoshikawa, Y.; and Imai, Y. 2018. A nonparametric delayed feedback model for conversion rate prediction. *arXiv preprint arXiv:1802.00255*.

Zhang, W.; Yuan, S.; and Wang, J. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1077–1086.