Dataset	# Users	# Items	# categorical feature	# continuous feature	# conversions	# Samples	# average CVR	# log period
Criteo Dataset	-	5443	9	8	3619801	15898883	0.2269	60 days
Taobao Dataset	0.25 billion	0.8 billion	12	10	0.32 billion	9.8 billion	0.03273	14 days

Table 1: Statistics of Criteo and Taobao Dataset.

Method	Criteo Dataset						Taobao Dataset					
	AUC	PR-AUC	NLL	*R-AUC*	*R-PR-AUC*	*R-NLL*	AUC	PR-AUC	NLL	*R-AUC*	*R-PR-AUC*	*R-NLL*
Pre-trained	0.8307	0.6251	0.4009	-0.9212	-0.2058	0.2139	0.8731	0.6525	0.1156	-1.0374	-0.5217	-0.2419
Vanilla	0.8376	0.6288	0.4047	0.0000	0.0000	0.0000	0.8842	0.6645	0.1141	0.0000	0.0000	0.0000
Oracle*	0.8450	0.6469	0.3868	1.0000	1.0000	1.0000	0.8949	0.6875	0.1079	1.0000	1.0000	1.0000
DFM	0.8132	0.5784	1.2599	-3.2581	-2.7833	-47.645	0.8702	0.6471	0.1271	-1.3084	-0.7565	-2.0968
FSIW	0.8290	0.6189	0.4099	-1.1432	-0.5479	-0.2891	0.8735	0.6591	0.1149	-0.9971	-0.2348	-0.1290
FNC	0.8373	0.6267	0.4382	-0.0393	-0.1147	-1.8646	0.8851	0.6669	0.1142	0.0841	0.1043	-0.0161
FNW	0.8373	0.6313	0.4033	-0.0308	0.1400	0.0773	0.8845	0.6672	0.1137	0.0280	0.1174	0.0645
ES-DFM	**0.8402***	**0.6393***	**0.3924***	**0.3560**	**0.5799**	**0.6831**	**0.8895***	**0.6762***	**0.1112***	**0.4953**	**0.5087**	**0.4677**

Table 2: Performance comparisons of proposed model with baseline models on AUC, PR-AUC and NLL metrics. The bold value marks the best one in one column, while the underlined value corresponds to the best one among all baselines. Here, * indicates statistical significance improvement compared to the best baseline measured by t-test at p-value of 0.05. *R-AUC*, *R-PR-AUC* and *R-NLL* are relative metrics indicating the improvements within the delayed feedback gap.

Choice of $p(e|x)$

The sampling elapsed time distribution $p(e|x)$ can be designed based on expert knowledge and the aforementioned bias analysis. For example, users need more time to consider when buying high-priced products, thus a long waiting time is required. However, the public dataset is anonymized, where information like price-level is unavailable. To verify the effectiveness of introducing $p(e|x)$ in the streaming settings, we perform a simplified implementation of $p(e|x)$. More precisely, we set $p(e = c|x) = 1$ where c is a constant, which means $p(e|x)$ degenerates to a Dirac distribution. This brings us two following advantages. First, we can strike the balance between obtaining accurate feedback information and keeping model freshness with a single parameter c. Second, we conducted experiments with different c in the public dataset, and the experimental results show that choosing the best c can significantly improve performance. The c is also tuned on the private dataset and we report the best result which is achieved using $c = 1$.

Standard Streaming Experiments: RQ1

From Table 2, we can see that our proposed method improves the performance significantly against all the baselines and achieves the state-of-the-art performance. Moreover, some further observations can be made. First, the performance of DFM and FSIW is worse than the vanilla baseline on both the public and Taobao Dataset. This is because DFM is difficult to converge, thus failing to achieve a good performance in streaming CVR prediction, and FSIW does not allow the data correction once a conversion took place afterwards, which is important for delayed feedback. Second, in most cases, FNC and FNW perform better than the vanilla baseline. Specially, FNW outperforms the baseline in both PR-AUC and NLL, which is consistent with the results reported in **?**. Third, existing methods show little superior performance in terms of AUC, while our method outperform the best baseline by 0.26% and 0.44% AUC scores on the Criteo and Taobao Dataset, respectively. As reported in **?**, DIN improves AUC scores by 1.13% and the improvement of online CTR is 10.0%, which means a small improvement in offline AUC is likely to lead to a significant increase in online CTR. In our practice, for cutting-edge CVR prediction models, even 0.1% of AUC improvement is substantial and achieves significant online promotion.

We further analyze the maximum benefit that can be achieved by resolving the delayed feedback problem. The maximum benefit is defined as the performance gap between the oracle model and baseline. Therefore, the goal of any method that tackling delayed feedback problem is to narrow this gap. We report three relative metrics within the performance gap, i.e, Relative-AUC(R-AUC), Relative-PR-AUC(R-PR-AUC) and Relative-NLL(R-NLL). As shown in Table 2, our method can narrow the delayed feedback gap significantly comparing to other methods, and the absolute improvement is larger when the delayed feedback gap is larger.

Influence of Elapsed Time: RQ2

To verify the performance of different choices of elapsed time, we have conducted experiments using different values of c on the Criteo dataset. As shown in Figure 2, the best c on the Criteo dataset is around 15 minutes, where about 35% conversions can be observed. Moreover, larger or smaller c will reduce the performance. The performance decreases slowly on smaller c, which indicates that the bias introduced by the importance weighting model is small. The performance decreases faster on larger c, which indicates that the data freshness matters more when c increase, and a c larger

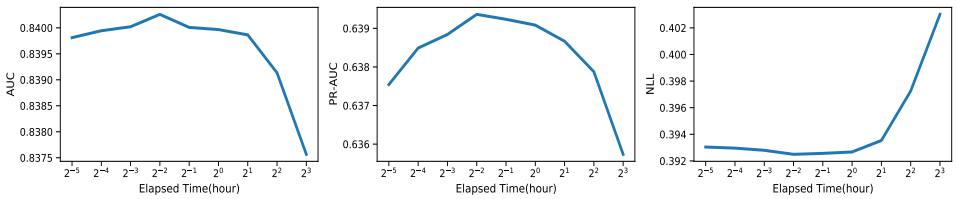

Figure 2: Experiments on the effect of elapsed time on performance. We control the elapsed time by a parameter c, which is the value on the x axis.

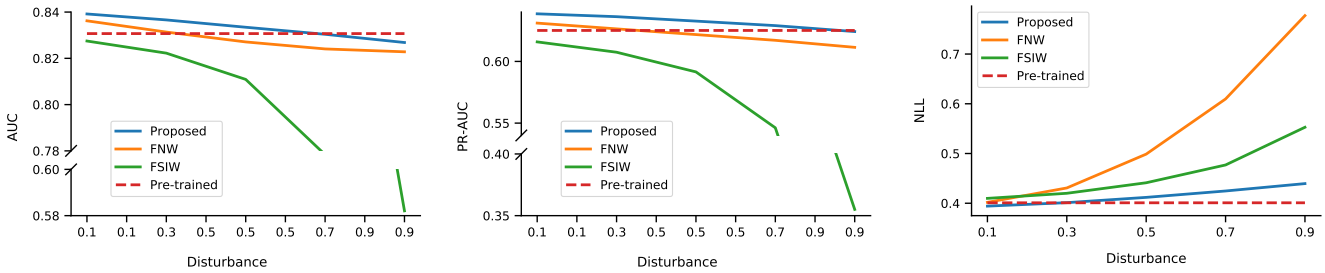

Figure 3: The experiment on resistance to disturbation. x axis is the disturbation strength which controls the portion of positive samples to be flipped.

than 1 hour will significantly harm the performance.

Experiment on Robustness: RQ3

In delayed feedback setting, the same sample may be labeled as negative or positive. It is closely related to learning with noisy labels(**?**), where some of the labels are randomly flipped. We hypothesis that a method dealing with delayed feedback problem should not only correct incorrect labels, but also reduce the negative effect of the incorrect labels before they can be corrected or the correction fails (for example, if the weighting model deviate a lot, the bias will be large and correction will fail). Thus we conducted a robustness experiment. We randomly select d portion of all the positive samples in streaming dataset, then swap it's label(and click time and pay time) with a random selected negative one. Note that we do not disturb on the pre-training dataset, so the initial CVR model and the pre-trained importance weighting models are not disturbed. We conducted experiments with different disturbance strength d, the results are shown in Figure 3. We can see that our method is more resistant to disturbance comparing to FNW and FSIW, and the performance gap is larger when disturbance increases (especially on NLL). We give an intuitive analysis about the weak robustness of FNW and FSIW in the Supplementary Material‡.

Online Evaluation: RQ4

We conducted an A/B test in our online evaluation framework. We observed a steady performance improvement, AUC increases by 0.3% within a 7 days window compared with the best baseline, CVR increases by 0.7%, GMV(Gross Merchandise Volume) increases by 1.8%, where GMV is computed by the transaction number of items multiplied by the price of each item. The online A/B testing results align with our offline streaming evaluation and show the effectiveness of ES-DFM in industrial systems.

Conclusion

The trade-off between the label accuracy and model freshness in streaming training setting has never been considered, which is an active decision of the method rather than a passive feature in offline setting. In this paper, we propose elapsed-time distribution to balance the label accuracy and model freshness to address the delayed feedback problem in the streaming CVR prediction. We optimize the expectation of true conversion distribution via importance sampling under the elapsed-time sampling distribution. Moreover, we propose a rigorous streaming training and testing experimental protocol, which aligns with real industrial applications better. Finally, extensive experiments show the superiority of our approach.

Ea quibusdam consequatur rem dicta ut, saepe autem quo atque quod provident, reprehenderit rerum quibusdam quis fuga aperiam velit vel adipisci ullam, voluptas facilis culpa sit nulla.Ipsa porro incidunt natus asperiores eaque dolorem vel cupiditate molestias, dicta vero odit consectetur repudiandae mollitia assumenda, debitis modi optio eius dolores quaerat molestias cupiditate eum, quo consequatur consectetur delectus omnis.Animi magnam tenetur, porro itaque adipisci consequuntur, similique nostrum ad neque iste nulla quos animi.Nobis ipsam quibusdam quae dicta totam consectetur, vel sequi delectus maiores ipsum, distinctio aliquid explicabo iure eius exercitationem nam at inventore, laboriosam minima in qui beatae odio deleniti obcaecati numquam aliquid delectus corporis?