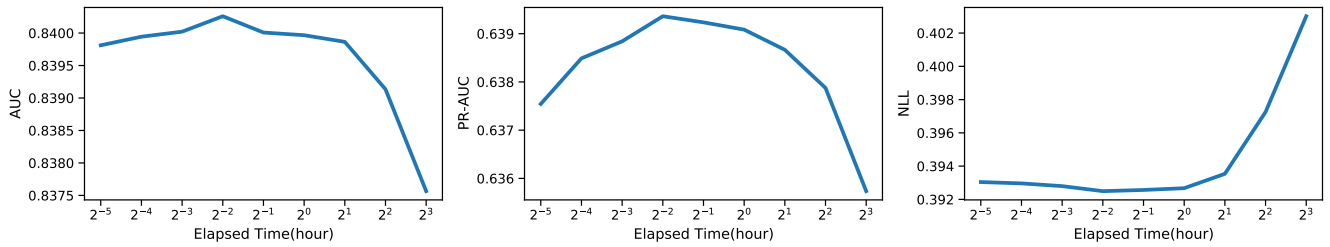

Figure 2: Experiments on the effect of elapsed time on performance. We control the elapsed time by a parameter c, which is the value on the x axis.

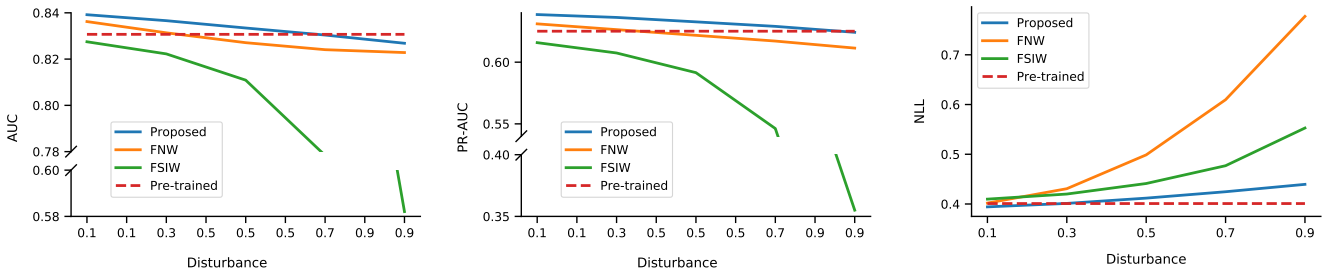

Figure 3: The experiment on resistance to disturbation. x axis is the disturbation strength which controls the portion of positive samples to be flipped.

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, distinc-

tio aliquid explicabo iure eius exercitationem nam at inventore, laboriosam minima in qui beatae odio deleniti obcaecati numquam aliquid delectus corporis?