Modeling Uplift from Observational Time-Series in Continual Scenarios

Sanghyun Kim[1], Jungwon Choi[1], Namhee Kim[2], Jaesung Ryu[3], Juho Lee[1]

[1]Kim Jaechul Graduate School of AI, KAIST
[2]Department of Digital Analytics, Yonsei University
[3]AFI Inc.
{nannullna, jungwon.choi, juholee}@kaist.ac.kr, cocopops@yonsei.ac.kr, kade@afidev.com

Abstract

As the importance of causality in machine learning grows, we expect the model to learn the correct causal mechanism for robustness even under distribution shifts. Since most of the prior benchmarks focused on vision and language tasks, domain or temporal shifts in causal inference tasks have not been well explored. To this end, we introduce **Backend-TS** dataset for modeling uplift in continual learning scenarios. We build the dataset with CRUD data and propose continual learning tasks under temporal and domain shifts.

Introduction

Uplift modeling is a particular type of predictive causal model with broad applications in marketing, personalized medicines, and politics. *Uplift* is defined as Individual Treatment Effect (ITE), but its evaluation metric differs from the other causal tasks (Radcliffe and Surry 1999). Separating causality from spurious relationships and precise estimation of treatment effects are crucial in causal tasks. However, by modeling uplift with causality, $p(y|do(t), x)$, we ultimately target a subgroup of individuals with high uplift scores, and therefore, the model's performance is measured by cumulative uplift across the population. Identifying this subgroup cannot be answered by the propensity model, $p(y|t = 1, x)$, which merely predicts one's future behavior.

In practice, the bottlenecks of causal models are data availability, scalability, and distribution shifts. In randomized controlled trials (RCTs), an individual's treatment is randomly assigned; therefore, we can *identify* Average Treatment Effect (ATE) with the difference between the treatment and control group's average outcomes (Pearl 2010). In many cases where RCTs are infeasible, however, practitioners are given observational data. No matter how many variables one has collected, unobserved confounders may still exist. Even if one can collect more covariates, the curse of dimensionality may occur. It is problematic, particularly for causal inference with high-dimensional data, as the chance of violating the positivity assumption increases (Zhao, Small, and Ertefaie 2017; D'Amour et al. 2021). Moreover, distribution may change over time and among different domains, resulting in improper validation and, eventually, the degradation of the fitted model.

To challenge the aforementioned issues with causality in high-dimensional spaces and bridge the gap between research and practice environments, we publish **Backend-TS** dataset[1], a real-world uplift dataset from mobile game users. The task is to predict uplift to push notifications by recognizing patterns from each user's CRUD[2] history. A model must learn underlying causal mechanisms and continuously adapt to distributions varying over time and to other games; otherwise, its performance will drop sharply when the distribution changes. We also argue that distribution changes can cause severe problems in causal inference since we model future customer behaviors based on their history. To the best of our knowledge, **Backend-TS** is the first uplift dataset with time-series under domain shifts.

Background

Causal inference and its notations. Potential outcomes framework (Rubin 1974) defines causal effect as the difference between two potential outcomes $Y(1) - Y(0)$: when receiving treatment ($T = 1$) and under control ($T = 0$). The fundamental problem of causal inference (Holland 1986) states that either $Y_i(1)$ or $Y_i(0)$ is observable for each unit indexed by $i \in \{1, \ldots, n\}$, and the unobserved outcome is called *counterfactual*. To estimate ITE, or uplift, $u_i := Y_i(1) - Y_i(0)$, we model Conditional Average Treatment Effect (CATE) conditioned on features $\mathbf{X}$, *i.e.*, $u(\mathbf{X}) := \mathbb{E}[Y(1) - Y(0)|\mathbf{X}]$. Among the assumptions needed to identify CATE, two assumptions are crucial and often likely to be violated (Pearl 2010; Neal 2020): unconfoundedness, *i.e.*, $Y \perp\!\!\!\perp T|\mathbf{X}$, and positivity, *i.e.*, $P(T|\mathbf{X} = \mathbf{x}) > 0, \forall \mathbf{x} : P(\mathbf{X} = \mathbf{x}) > 0$.

Uplift modeling. Here, we introduce marketing terms following Radclifte and Simpson (2008) to illustrate the concept of uplift modeling. Individuals can be segmented into four groups along two axes: **received treatment** and **response to it**. *Sure Things* will stay (or buy a product)

[1]The dataset is available under CC BY-NC-SA 4.0 license at https://blog.thebackend.io/research/backend-ts, and the baseline code for models and dataloader is available at https://github.com/nannullna/ts4uplift

[2]CRUD refers to the four functions necessary for storage and server applications: create, read, update, and delete.

whether or not they receive treatment (*e.g.,* an advertisement), and *Lost Causes* will leave (or not buy the product) in either case. In short, the treatment has neither positive nor negative effects on both groups, *i.e.*, $u_i = Y_i(1) - Y_i(0) \approx 0$. On the other hand, *Persuadables* are likely to stay *only if* they receive the treatment, *i.e.*, $u_i > 0$, but *Sleeping Dogs* would be annoyed and eventually leave, *i.e.*, $u_i < 0$. Based upon this fundamental segmentation, the main goal is thus to identify as many *Persuadables* as possible while avoiding *Sleeping Dogs* for the treatment.

Time-series modeling. Time-series is a sequence of discrete-time data. Many previous works have dealt with regular time-series, but in this paper, we mainly focus on irregular time-series, where intervals between two consecutive data points are not the same. RNNs (Rumelhart, Hinton, and Williams 1986; Hochreiter and Schmidhuber 1997; Cho et al. 2014), TCNs (Bai, Kolter, and Koltun 2018) with dilated convolutions (Yu and Koltun 2015), and Transformers (Vaswani et al. 2017) have become popular choices for handling time-series data. However, there is no *one-size-fits-all* augmentation strategy in various types of time-series (Yue et al. 2022) except for dropout (Srivastava et al. 2014), or random masking (Devlin et al. 2018; He et al. 2022).

Continual learning. Continual Learning (CL) aims to effectively learn new tasks and adapt a model to distribution shifts over time while minimizing performance degradation in the learned scenarios, which is called *catastrophic forgetting* (McCloskey and Cohen 1989; Kirkpatrick et al. 2017). It is also infeasible in practice to fully retrain the model whenever new data are available due to training costs or the unavailability of previous data. Therefore, recent algorithms for CL aim to accumulate knowledge and reuse them in future scenarios without forgetting information (*e.g.*, iCaRL (Rebuffi et al. 2017), A-GEM (Chaudhry et al. 2019), EWC (Kirkpatrick et al. 2017), SI (Zenke, Poole, and Ganguli 2017)). Moreover, causal inference tasks require the model to capture the causal mechanism over distributional shifts, on which existing CL algorithms have not focused.

Previous Benchmarks

Benchmarks for uplift modeling. Researchers on uplift have relied on (semi-)synthetic data for testing algorithms since underlying causal mechanisms are fully specified and counterfactuals thus exist. On the other hand, as of now, the largest observational benchmark is **Criteo dataset** (Eustache et al. 2018) with 12 static features from ∼14M real-world users. Thus far, there has been little motivation to use deep learning, and therefore, related works have been restricted to smaller neural networks (# params $<$ 1K) or other machine learning algorithms. With regard to causal inference with time-series, a subset of **MIMIC II/III** (Johnson et al. 2016) has been used for causal discovery or inference. See Moraffah et al. (2021) for a comprehensive review.

Benchmarks for CL. Benchmarks in various fields and tasks with CL scenarios have been introduced, *e.g.*, object recognition in robotics (Fanello et al. 2013; Lomonaco and Maltoni 2017; She et al. 2020), classification tasks in various domains on images (Rebuffi, Bilen, and Vedaldi 2017; Lake, Salakhutdinov, and Tenenbaum 2015; He, Shen, and Cui 2021), videos (Roady et al. 2020), 3D objects (Stojanov et al. 2019), and natural language (Hussain et al. 2021; Srinivasan et al. 2022). However, domain or temporal shifts in causal inference tasks have not been well explored.

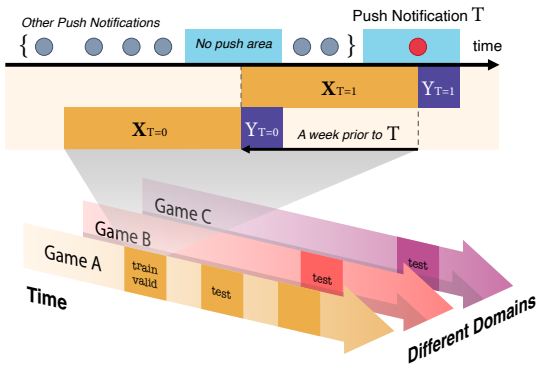

Figure 1: Illustration of **Backend-TS** dataset construction.

Backend-TS Dataset

In this section, we introduce **Backend-TS** dataset. The dataset construction method and the proposed tasks are briefly illustrated in Figure 1.

Background. We collected data from AFI Inc., a Backend-as-a-Service (BaaS) company specializing in mobile games. The company owns backend servers and provides APIs that game developers can access to quickly release their apps without backend servers of their own. One of the features is to send **a push notification** to *all users at the same time*. We wanted to build a model that targets only a subset of users with high gains from a push message. However, CRUD log data are only available to us since the company does not collect user-specific information and has no access to each game's code or internal data.

Construction. The treatment is not assigned randomly in a typical observational dataset, and the true treatment assignment mechanism is unknown. In our data, however, the treatment group only exists as the push message is given to all users simultaneously. To circumvent this problem, for a train set, we sampled a *pseudo*-control group exactly one week before the push so as to eliminate the time and weekday effect. We also introduced a concept of *no push area*, an -12∼+6 hour window around which no other pushes must exist to prevent interference from them. Note that some users exist in both groups, and utilizing those data points (*e.g.*, randomly choosing either one or using both) is up to modeling strategies. For a test set, we randomly split those overlapping users into either group to simulate RCTs.

Overview. The dataset consists of three games (A, B, and C) with a total of 16.7M lines of CRUD logs from 5,360 users. Only a handful of games met the conditions mentioned above among hundreds of games in service, most of

which either sent pushes too frequently or did not use this API at all. Each consists of a triple $(\mathbf{X}_i, T_i, Y_i)$, where $\mathbf{X}_i$ is a sequence of categorical variables along with corresponding timestamps, and $T_i, Y_i \in \{0, 1\}$ are binary indicators of the treatment and whether a gamer logged in within {three, six, twelve}[3] hours after the push message had been sent. Although the games use the same APIs provided by the company, they differ in response rates, lengths, API usage, and other factors. For example, two different games may use the same type of API calls for different purposes.

Tasks. We experimented on uplift modeling in the following proposed tasks:

- In-domain (ID): train with the game A (APR, MAY) and test on 20% random-split holdout set.
- Temporal shift (TS): train with the game A (APR, MAY) and test on the game A (JUN).
- Out-of-domain (OOD): train with the game A (APR, MAY) and test on the game B with fine-tuning (OOD w/) or on the game C without fine-tuning (OOD w/o).

Experiments

Model	Ckpt	ID	TS	OOD w/	OOD w/o
Dragon	VAL	.091/.056	.006/.003	.118/.038	.037/.023
	MAX		.112/.074	.372/.082	.123/.081
Siamese	VAL	.145/.062	-.036/-.011	.154/.057	-.057/-.030
	MAX		.249/.067	.207/.075	.036/.022
$P(Y=1)$		11.9%	12.2%	5.9%	22.4%

Table 1: Baseline results. VAL denotes the best checkpoint on the holdout set, and MAX denotes the best metric during entire training, showing the discrepancy of the performance.

Baselines. We used Dragonnet (Shi, Blei, and Veitch 2019) and Siamese network (Mouloud, Olivier, and Ghaith 2020) with 11 TCN blocks (receptive field of length 2,048, and each time-series was truncated accordingly.) and applied EWC for CL. Dragonnet is trained to directly predict a conditional mean $\mathbb{E}[Y|T, \mathbf{X}]$ as well as the propensity score, $e(\mathbf{X}) := P(T = 1|\mathbf{X})$, based on its sufficiency for adjustment (Rosenbaum and Rubin 1983). For Siamese network, a variable transformation method, $Z_i = \frac{T_i Y_i}{e(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-e(\mathbf{X}_i)}$, was used based on the fact that its conditional expectation, *i.e.*, $\mathbb{E}[Z|\mathbf{X}]$, is equal to the true uplift $u(\mathbf{X})$ (Athey and Imbens 2015). We attached an embedding layer with LayerNorm (Ba, Kiros, and Hinton 2016) which is similar to language models like BERT (Devlin et al. 2018) for categorical variables and used sinusoidal functions to encode second, hour and weekday information as follows:

$$f(t) = \left[\sin\left(\frac{2\pi t}{\max_t}\right), \cos\left(\frac{2\pi t}{\max_t}\right)\right],$$

where $\max_t$ is the maximum possible value of t, *i.e.*, 3600 seconds in an hour, 24 hours in a day, and 7 for weekday.

[3]The shorter the time interval, the greater the influence of the push, but the smaller the number of people responding. In our experiment, "three hours" was used as a target.

Evaluation. The performance of an uplift model can be evaluated by qini coefficients (QINI) (Radcliffe 2007) and area under uplift curve (AUUC) (Devriendt et al. 2020). The two metrics are basically similar, measuring cumulative incremental gains when the treatment is given only to the top individuals sorted by uplift scores predicted by the model.

Results. Table 1 shows QINIs (left) and AUUCs (right) of the best checkpoint on the holdout set (VAL) and among the entire training checkpoints (MAX) for each task. The difference between VAL and MAX can be attributed to the model capturing spurious correlations rather than the true mechanisms and the wrong validation due to distributional shifts.

- TS: The performance gap between VAL and MAX was significant, and VAL actually performed worse than random targeting (QINI & AUUC below zero). This empirically shows the existence of the temporal distribution changes.
- OOD W/: Fine-tuning with the additional data using the CL algorithm has somewhat reduced the performance gap. We conjecture that the model became more robust since it further learns common mechanisms and forgets relationships irrelevant to the true effect.
- OOD W/O: The performance dropped sharply without fine-tuning. We emphasize that the true causal model should perform equally well both in ID and TS and generalize to different games even without training, although they may potentially have a very different user base.

Conclusion and Future Work

In this paper, we introduce **Backend-TS** dataset and propose uplift tasks accordingly, combining causal inference with CL scenarios. We demonstrate that naïvely applying existing methods may fail as uplift modeling tries to predict future behaviors based on historical data. All observational datasets have inherent biases; identifying causal relationships and eliminating undesirable effects would be one of the most important follow-up research topics. We believe that learning causal mechanisms invariant over time is crucial for the way toward general-level AI and that the dataset will contribute to developing such algorithms.

Ethical Statement and Societal Impact

We did not collect any sensitive information, and all data have been fully anonymized. Do not attempt to misuse it for purposes other than research, including but not limited to, identifying individuals or games, hacking, and cracking the system. **Backend-TS** will contribute to developing robust models and algorithms that can infer correct causal mechanisms in high-dimensional spaces.

Acknowledgement

We thank AFI Inc. and anonymous game companies for allowing data to be published for research purpose. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)).

References

Athey, S.; and Imbens, G. W. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5): 1–26.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2019. Efficient Lifelong Learning with A-GEM. In *International Conference on Learning Representations*.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devriendt, F.; Van Belle, J.; Guns, T.; and Verbeke, W. 2020. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*.

D'Amour, A.; Ding, P.; Feller, A.; Lei, L.; and Sekhon, J. 2021. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654.

Eustache, D.; Artem, B.; Renaudin, C.; and Massih-Reza, A. 2018. A Large Scale Benchmark for Uplift Modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM.

Fanello, S. R.; Ciliberto, C.; Santoro, M.; Natale, L.; Metta, G.; Rosasco, L.; and Odone, F. 2013. iCub World: Friendly Robots Help Building Good Vision Data-Sets. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 700–705.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110: 107383.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Hussain, A.; Holla, N.; Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Lomonaco, V.; and Maltoni, D. 2017. CORe50: a New Dataset and Benchmark for Continuous Object Recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, 17–26.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Moraffah, R.; Sheth, P.; Karami, M.; Bhattacharya, A.; Wang, Q.; Tahir, A.; Raglin, A.; and Liu, H. 2021. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 1–45.

Mouloud, B.; Olivier, G.; and Ghaith, K. 2020. Adapting neural networks for uplift models. *arXiv preprint arXiv:2011.00041*.

Neal, B. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*.

Pearl, J. 2010. Causal inference. *Causality: objectives and assessment*, 39–58.

Radcliffe, N. 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, 14–21.

Radcliffe, N.; and Surry, P. 1999. Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*.

Radclifte, N. J.; and Simpson, R. 2008. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management*, 1(2).

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Roady, R.; Hayes, T. L.; Vaidya, H.; and Kanan, C. 2020. Stream-51: Streaming Classification and Novelty Detection From Videos. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.

She, Q.; Feng, F.; Hao, X.; Yang, Q.; Lan, C.; Lomonaco, V.; Shi, X.; Wang, Z.; Guo, Y.; Zhang, Y.; Qiao, F.; and Chan, R. H. M. 2020. OpenLORIS-Object: A Robotic Vision Dataset and Benchmark for Lifelong Deep Learning. In *2020 International Conference on Robotics and Automation (ICRA)*, 4767–4773.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Srinivasan, T.; Chang, T.-Y.; Alva, L. L. P.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. CLiMB: A Continual Learning Benchmark for Vision-and-Language Tasks. *arXiv preprint arXiv:2206.09059*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.

Stojanov, S.; Mishra, S.; Thai, N. A.; Dhanda, N.; Humayun, A.; Yu, C.; Smith, L. B.; and Rehg, J. M. 2019. Incremental Object Learning From Contiguous Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, 3987–3995. PMLR.

Zhao, Q.; Small, D. S.; and Ertefaie, A. 2017. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*.