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.

THE 37TH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE

Challenges in Causal Models

Data Availability — Limited to synthetic dataset

Scalability — Unconfoundedness positivity trade-off

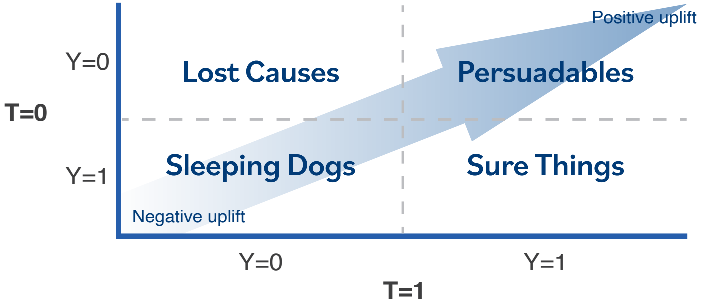
Distribution Shifts — Different domain and test time

Causality in high-dimensional spaces

A gap between research and practice

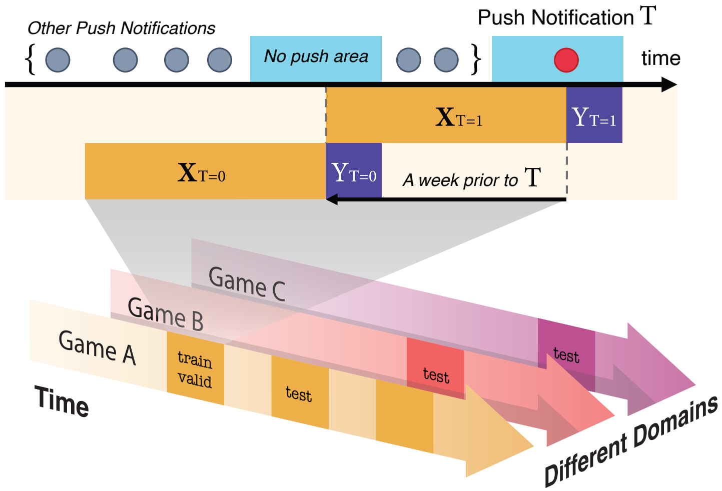
Uplift Modeling

Uplift modeling aims to identify a subgroup of individuals with high uplift scores (or Individual Treatment Effect, ITE).



Background

- Data was collected from AFI Inc., a Backend-as-a-Service (BaaS) company specializing in **mobile games**.
- The company provides APIs for game developers to release apps without the need for backend servers.
- The goal was to build a model that targets only a subset of users with high gains from **a push message**.
- The data used for benchmark is **CRUD log data**, as the company does not collect user-specific information or have access to game's code or internal data.



Backend-TS Dataset

16.7 million lines of CRUD log data from 5,360 users in three mobile games

- Each data point consists of a triple (X, t, y).
- **pseudo-control group**: the control group data was sampled exactly one week prior to the push message.
- **no push area**: an -12~+6 hour window around which no other pushes must exist.

Proposed Tasks

	Different Time	Different Game	Fine-tuning
ID (in-domain)	✗	✗	✗
TS (temporal shift)	✓	✗	✗
OOD (out-of-domain) **w/**	✓	✓	✓
OOD (out-of-domain) **w/o**	✓	✓	✗

Baselines

We used **Dragonnet** (Shi, Blei, and Veitch, 2019) and **Siamese network** (Mouloud, Olivier, and Ghaith, 2020) with **11 TCN blocks** and applied **EWC** for CL. Time/week information is embeded with sinosoidal functions, and API call type (discrete) is embeded with an embedding layer.

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%

The table shows **QINI**s (left) and **AUUC**s (right) of the best checkpoint on the holdout set (VAL) and among the entire training checkpoints (MAX) for each task.

- **TS**: The performance gap between VAL and MAX was significant, and VAL actually performed worse than random targeting. 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.
- **OOD w/o**: The performance dropped sharply without fine-tuning. We emphasize that the true causal model should perform equally well and generalize to different games even without training, although they may potentially have a very different user base.

nannullna@kaist.ac.kr
github.com/nannullna/ts4uplift

Download dataset