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.

AAAI-23 Continual Causality Bridge

February 8th, 2023

Introduction

Modeling Uplift from Observational Time-Series in Continual Scenarios

- Modeling uplift: (simple) causal inference
- Observational time-series: a novel real-world dataset "Backend-TS"
- Continual scenarios: continual learning scenarios

Challenges in Causality

Data Availability

Scalability

Distribution Shifts

- Limited to synthetic dataset
- RCTs are expensive and often impossible.

- Unconfoundedness-positivity trade-off
- Causality in high-dimensional spaces

- Generalizability to different (unseen) domains
- Train time != test time (temporal difference)

Uplift Modeling

- Models the uplift (or ITE, Individual Treatment Effect) of each user as follows:

$$u_i \; = \; E[Y_i(1) - Y_i(0)]$$

- Due to the fundamental problem of causal inference, we instead model CATE (Conditional Average Treatment Effect) as follows:

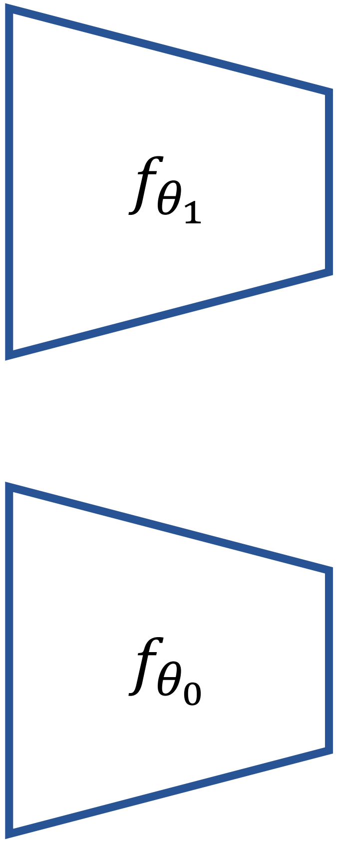
$$u(X) \; = \; E[Y(1) - Y(0)|X]$$

- Ultimately targets a subgroup of users with high uplifts from the treatment (e.g., push message, advertisement, drug)

A Naïve Implementation

X	t	y
X_1	1	0
X_2	1	1
X_3	0	0
X_4	0	1
…		
X_{n-1}	1	1
X_n	0	1

X	t	y(1)	y(0)
X_1	1	0	N/A
X_2	1	1	N/A
X_3	0	N/A	0
X_4	0	N/A	1
…			
X_{n-1}	1	1	N/A
X_n	0	N/A	1



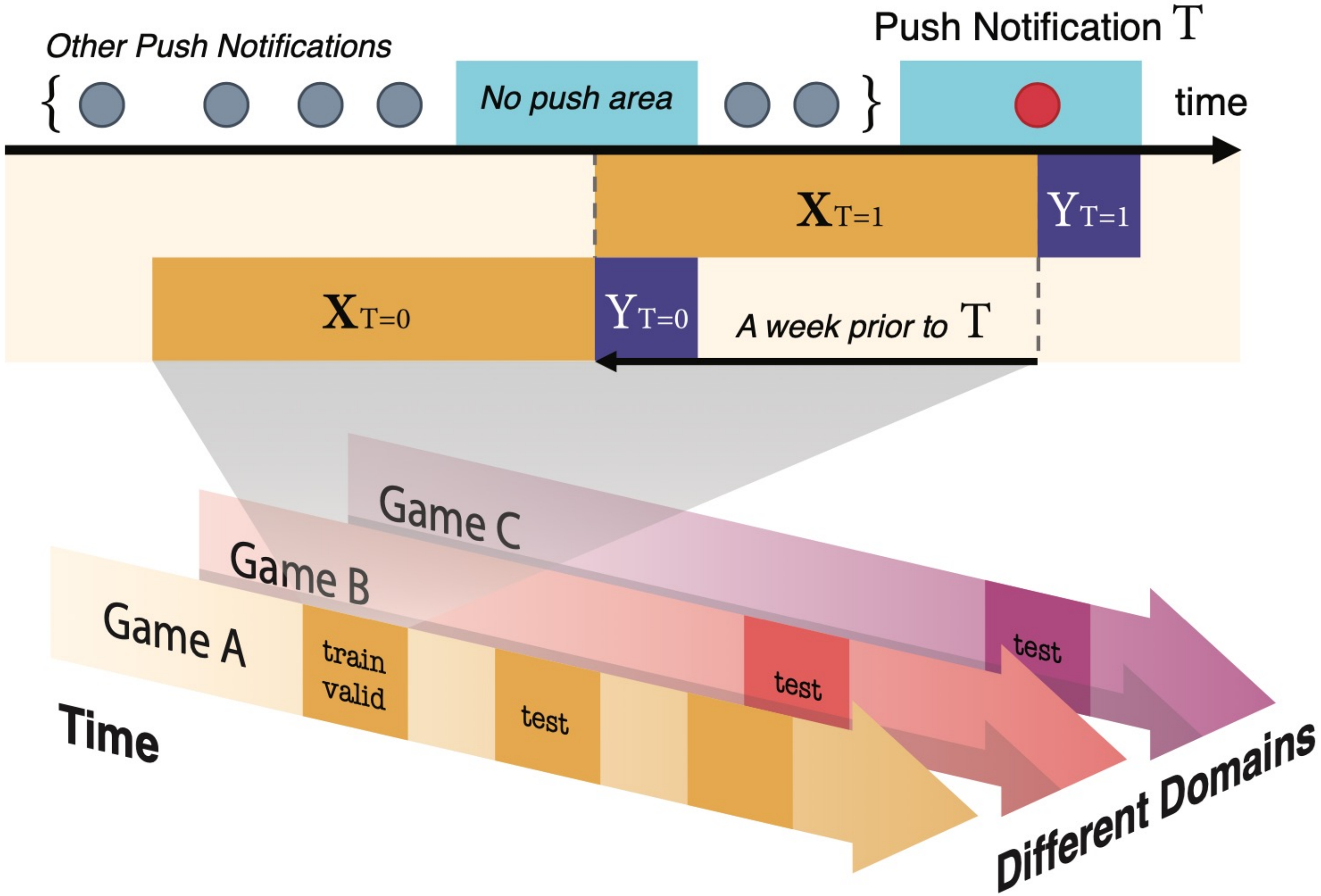
$\Pr(Y = 1 | T = 1)$

$\Pr(Y = 1 | T = 0)$

Dataset Construction

- CRUD log
  - CRUD: Create, Read, Update, and Delete
  - Transaction logs are stored in data warehouses.
  - The company provides common APIs but does not have access to internal data.
- Pseudo-control group
  - The control group does not exist in the raw data.
  - Sample a pseudo-control group when no push exists a week (168 hrs) before the push message for the treatment group.
- No push area
  - An -12~+6 hour window around which no other pushes must exist.
  - To prevent interference from other push messages.

Dataset Illustration



Dataset Overview

- 16.7 million lines from 5,360 users of three mobile games (A, B and C) currently in service
- A triple (X, t, y), where
  - X: datetime information (millisecond)
  - t: treatment/control group (push message)
  - y: user login within 3/6/12 hours from the push message
- URL: https://github.com/nannullna/ts4uplift

Proposed Tasks

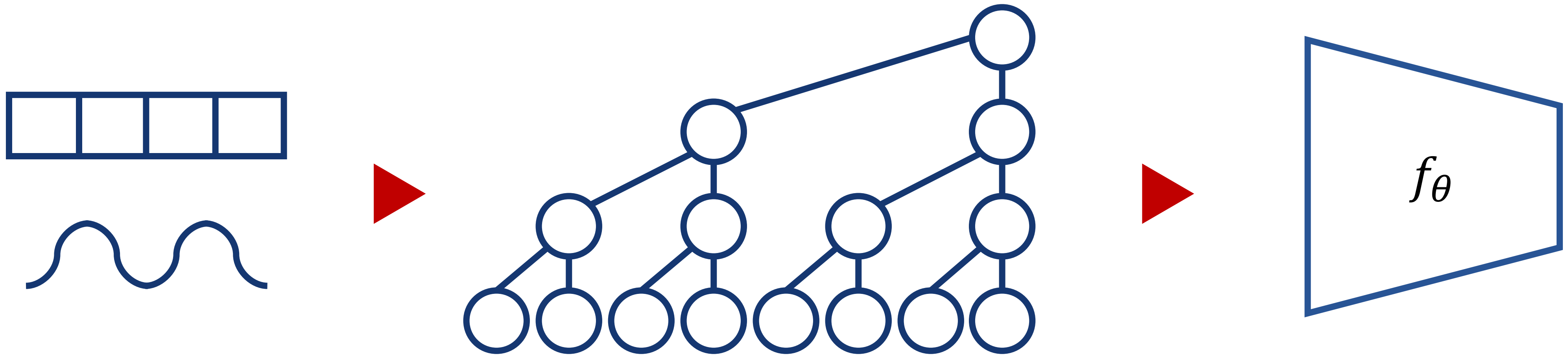
	Different Time	Different Game	Fine-tuning
ID (in-domain)	✗	✗	✗
TS (temporal shift)	✓	✗	✗
OOD (out-of-domain) **w/**	✓	✓	✓
OOD (out-of-domain) **w/o**	✓	✓	✗

Task	Train set	Valid set	Test set
ID	Game A APR + MAY	Game A APR + MAY (20% split)	-
TS	Game A APR + MAY	Game A APR + MAY (20% split)	Game A JUN
OOD w/	Game A APR + MAY & Game B JUN	Game B JUN (20% split)	Game B JUL
OOD w/o	Game A APR + MAY	Game A APR + MAY (20% split)	Game C JUL

Baseline

- TCN
  - 11 dilated 1D convolution blocks
  - Receptive field (max length of inputs) of 2,048
  - Additional embedding layer & sinusoidal functions to embed categoricals
- Dragonnet (Shi et al., 2019)
  - Regularization on the propensity score
- Siameses Network (SMITE) (Mouloud et al., 2020)
  - Z variable transformation (Athey, 2015)

Baseline Illustration



Embedding Layer

TCN Backbone

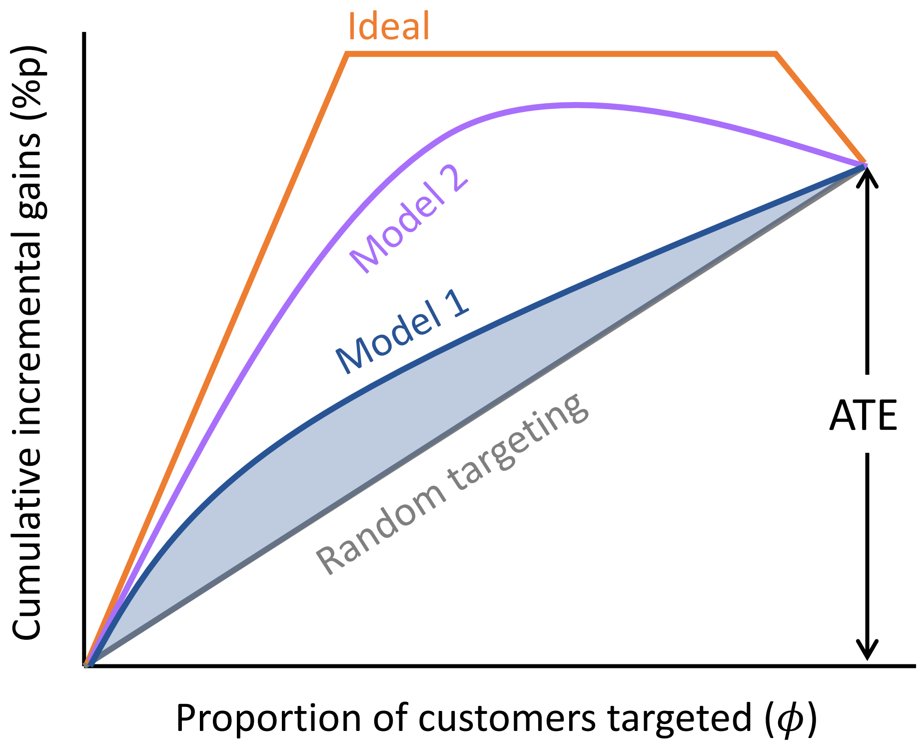
Dragonnet/Siamese Network

Evaluation

- Qini coefficient
  : a normalized area (shaded) between the qini curve and the random targeting line (ATE).
- Alternatively, AUUC (area under uplift curve)
- The Qini Curve

$$Qini\ curve(\phi) = \frac{n_{t,y=1}(\phi)}{N_t} - \frac{n_{c,y=1}(\phi)}{N_c}$$

- In the right figure, Model 2 performs better than Model 1.



Results

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%

- 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.

Results

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%

- 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.

Results

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

- 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.

Acknowledgement

We thank AFI Inc. and anonymous game companies
for allowing data to be published for research purpose.