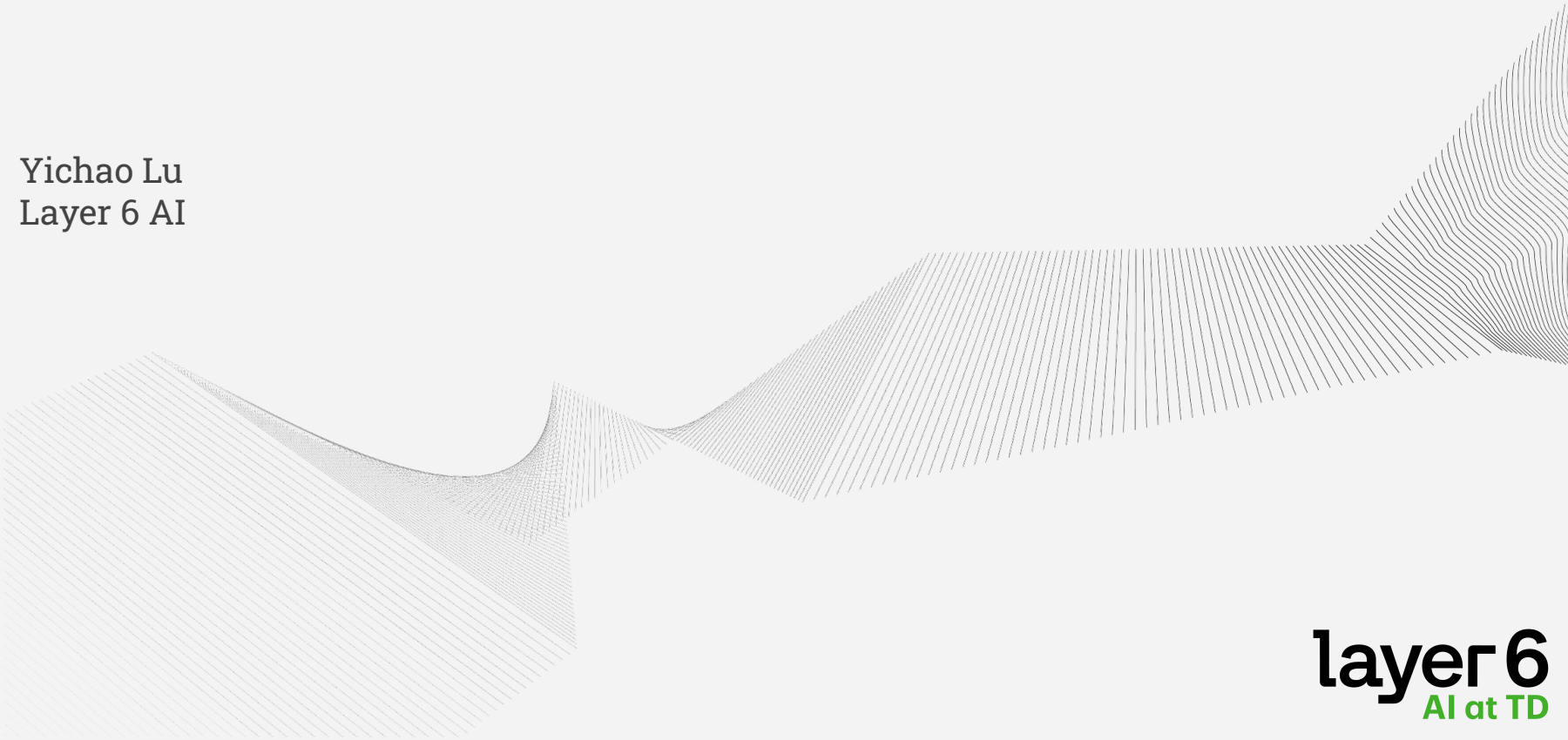
An Efficient Two-stage Gradient Boosting Framework for Short-term Traffic State Estimation

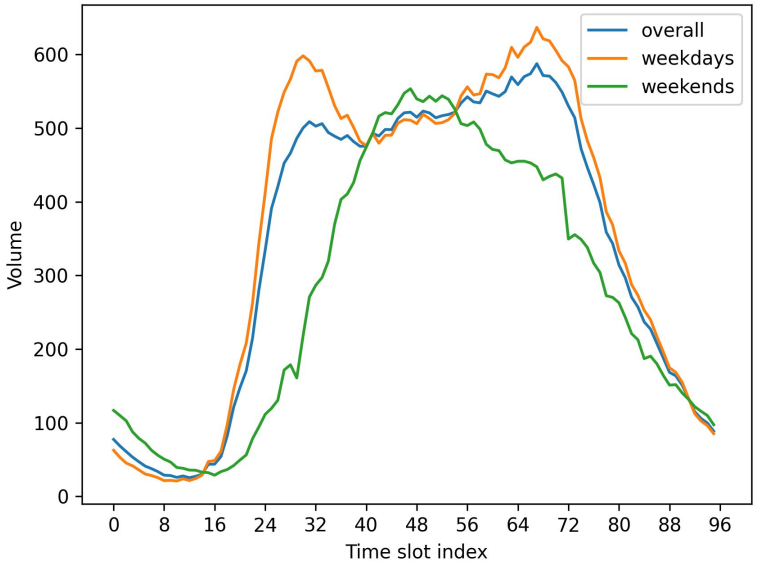
Yichao Lu
Layer 6 AI



layer 6
AI at TD

Why Two-stage?

- We observe strong seasonality and time trends in traffic flows.



- Dissecting the task of traffic state estimation into two stages (1) harnesses the information within auxiliary labels, and (2) facilitates the second stage model to capture the time patterns in traffic flows.

layer 6
AI at TD

Why Gradient Boosting?

- Robust and relatively insensitive to hyper-parameters.
- No data preprocessing required - works with categorical and numerical values as is.
- Handles missing values without imputation.
- Highly efficient and scalable.

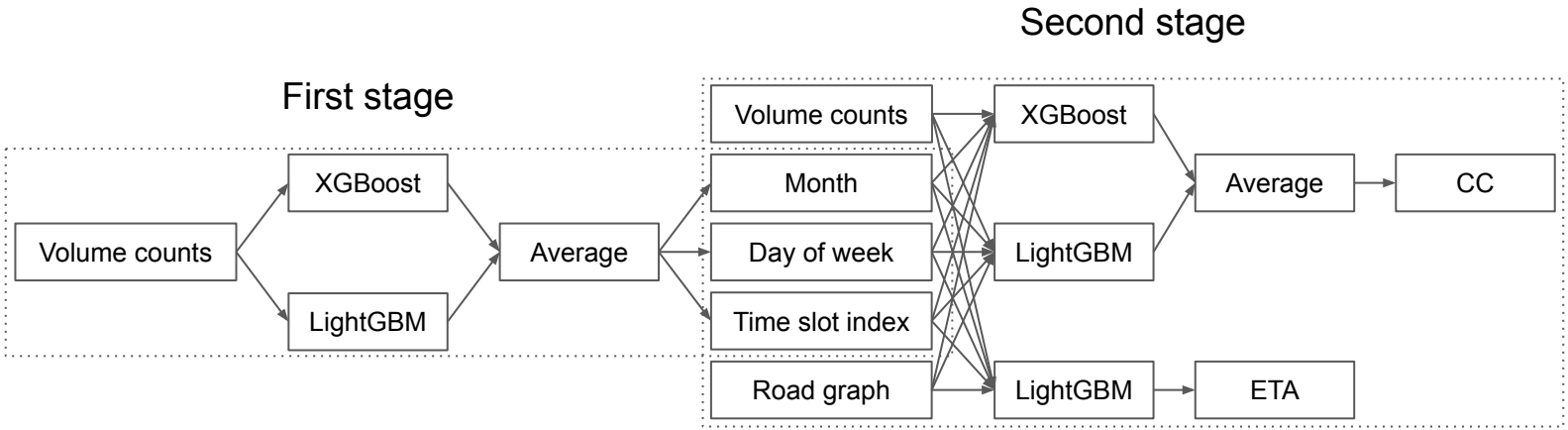
First Stage

- We use the volume counts for the nodes in the entire road graph as features, and predict the month, day of the week, and time slot index based on the sparse loop counter data.

- We use three separate models for the first stage, as gradient boosting decision trees do not directly support learning multiple targets.

- All three prediction tasks are modelled as a regression problem, where we optimize the model using L2 loss. The final prediction for the first stage is based on an ensemble of XGBoost and LightGBM models, trained separately for each city.

A Two-stage Gradient Boosting Framework



Second Stage

- For the core challenge, we engineered a number of features capturing the road network characteristics and traffic dynamics. The final prediction for the core challenge is made by an ensemble of XGBoost and LightGBM models, trained separately for each city.

- For the extended challenge, we only use LightGBM since XGBoost does not support the direct optimization of the L1 loss.

Target Encoding (Core Competition)

$$TE_{cc}([Categories]) = \frac{Count([Categories]) * mean_{cc}([Categories]) + w * mean_{cc}}{Count([Categories]) + w}$$

Target Encoding (Extended Competition)

$$TE_{eta}([Categories]) = \frac{1}{N}\sum_{i=1}^{N} ETA_i([Categories])$$

$$Smoothed\_TE_{eta}(t) = \frac{TE_{eta}(t) + \sum_{i=1}^{4}\left[TE_{eta}(t-i) + TE_{eta}(t+i)\right] * (i+1)^4}{1 + \sum_{i=1}^{4}(i+1)^4 * 2}$$

Feature Importance (Core Competition)

#	Description	Score
1	TE of cc green given the time slot index and whether the date is a weekend	674.99
2	TE of cc red given the time slot index	432.00
3	TE of cc green given the time slot index	301.92
4	The importance of the highway within the road network	181.96
5	TE of cc red given the time slot index and whether the date is a weekend	166.05
6	TE of cc yellow given the time slot index and whether the date is a weekend	135.27
7	TE of cc red given whether the date is a weekend	134.05
8	Whether the date is a weekend	118.69
9	TE of cc yellow given the time slot index	114.36
10	TE of cc yellow given whether the date is a weekend	64.03
11	The time slot index	55.28
12	Month	52.22
13	TE of cc green given the time slot index and day of the week	46.05
14	TE of cc green given day of the week	45.41
15	TE of cc red given day of the week	41.98
16	Day of the week	41.39
17	Whether the road graph edge can only be used in one direction by vehicles	39.22
18	Numerical mapping of the OSM highway class	34.11
19	TE of cc red given the time slot index and day of the week	33.58
20	TE of cc red	31.34
21	TE of cc green given whether the date is a weekend	29.81
22	Whether the road graph edge runs in a tunnel	27.98
23	TE of cc yellow given day of the week	24.75
24	Edge speed (km per hour)	21.54
25	Maximum legal speed limit	20.31
26	TE of cc yellow	19.80
27	TE of cc green	18.83
28	Number of node hops to get to the closes vehicle counter in the graph	17.44
29	Number of traffic lanes on the road graph edge	14.22
30	Edge length in meters	13.69
31	Unique identifier of the source node	12.98
32	Unique identifier of the sink node	12.82
33	Unique identifier of the road graph edge	12.27
34	Number of incoming edges for the sink node	12.04
35	Number of outgoing edges for the source node	11.97
36	Number of incoming edges for the source node	11.78
37	Number of outgoing edges for the sink node	11.77
38	TE of cc yellow given the time slot index and day of the week	10.74
39	The volume count for the target node 15 minutes in the past	8.32
40	The volume count for the source node 15 minutes in the past	8.24
41	The volume count for the target node 45 minutes in the past	8.06
42	The volume count for the target node 60 minutes in the past	8.06
43	The volume count for the source node 60 minutes in the past	8.02
44	The volume count for the source node 45 minutes in the past	7.90
45	The volume count for the target node 30 minutes in the past	7.58
46	The volume count for the source node 30 minutes in the past	7.55

layer 6
AI at TD

Feature Importance (Extended Competition)

#	Description	Score
1	Smoothed TE given the time slot index and day of the week	228374728.6
2	Unique identifier of the supersegment	76275475.3
3	Smoothed TE given the time slot index	44789850.5
4	The time slot index	7016971.0
5	TE given the time slot index and day of the week	5319797.0
6	Month	3703889.9
7	TE given day of the week	3033482.3
8	TE given the time slot index	2704734.8
9	Number of nodes in the supersegment	2693478.1
10	TE for the supersegment	1884798.9
11	Day of the week	1871912.3
12	TE given whether the date is a weekend	1080182.0
13	Whether the date is a weekend	935208.7
14	Smoothed TE given the time slot index and whether the date is a weekend	804695.4
15	TE given the time slot index and whether the date is a weekend	510518.6

Implementation Details

XGBoost

- max_depth: 5
- eta: 0.01
- subsample: 0.5
- colsample_bytree: 0.9
- colsample_bylevel: 0.9
- tree_method: gpu_hist
- early_stopping_rounds: 1000

LightGBM

- learning_rate: 0.1
- early_stopping_rounds: 1000

Experiments

Core Competition

Approach	Score	Time
MLP	0.85685871201332	3 hours 17 minutes
GNN	0.85204471385916	19 hours 34 minutes
Single-stage	0.85483927384105	**1 hour 35 minutes**
Two-stage	**0.85041532913844**	1 hour 54 minutes

Extended Competition

Approach	Score	Time
MLP	61.39940295221	2 hours 52 minutes
GNN	61.24305781924	11 hours 19 minutes
Single-stage	61.31830081748	**28 minutes**
Two-stage	**61.22274017334**	47 minutes

Competition Leaderboard

Core Competition

Rank	Team	Score
1	ustc-gobbler	0.843107938766 48
2	Bolt	0.849667906761 17
3	**oahciy (ours)**	**0.850415329138 44**
4	GongLab	0.856030921141 31
5	AP_DE	0.873508433500 93

Extended Competition

Rank	Team	Score
1	ustc-gobbler	58.4997215271
2	TSE	59.782447814941
3	**oahciy (ours)**	**61.22274017334**
4	Bolt	61.254610697428
5	discovery	62.296744028727

Discussion

- In a real-world production system the date and the time for the prediction of interest are very easy to obtain. Therefore real-world production systems should directly use the second stage model to avoid the propagation of errors from the first stage to the second stage.

- Handling the missing data using techniques such as Principal Component Analysis (PCA) may help boost performance.

- Adding the engineered features to a (graph) neural network based pipeline may also help improve the performance.

Thank You

layer 6
AI at TD