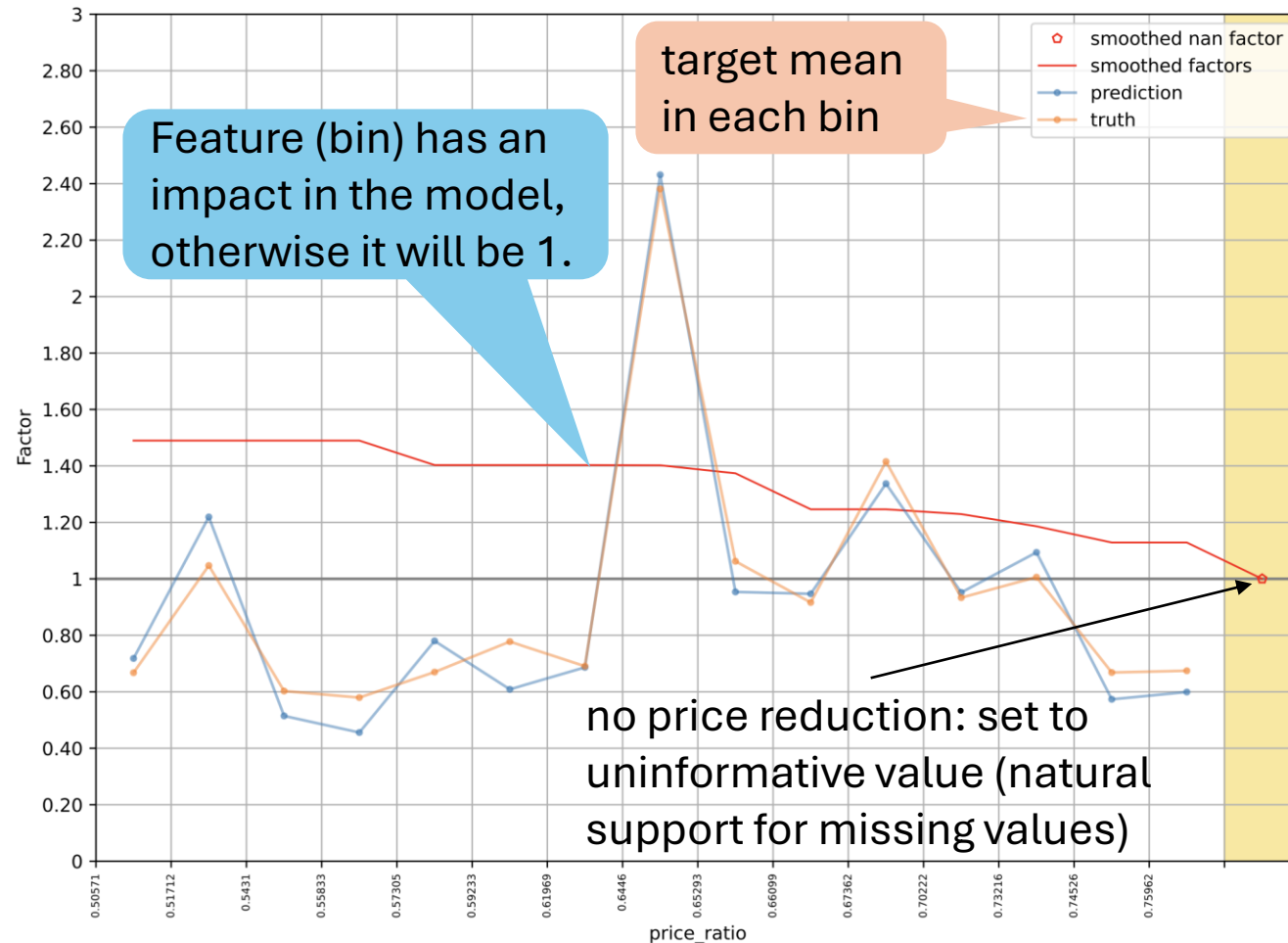


Cyclic Boosting

- binning → local optimization (low-bias method, capturing rare effects)
- smoothing over different bins (reduction of variance)
- forward-stagewise fitting (cyclic coordinate descent)
- some similarity to Generalized Additive Models (e.g., link function, interaction terms)

→ individual explainability

multiplicative mode → factors



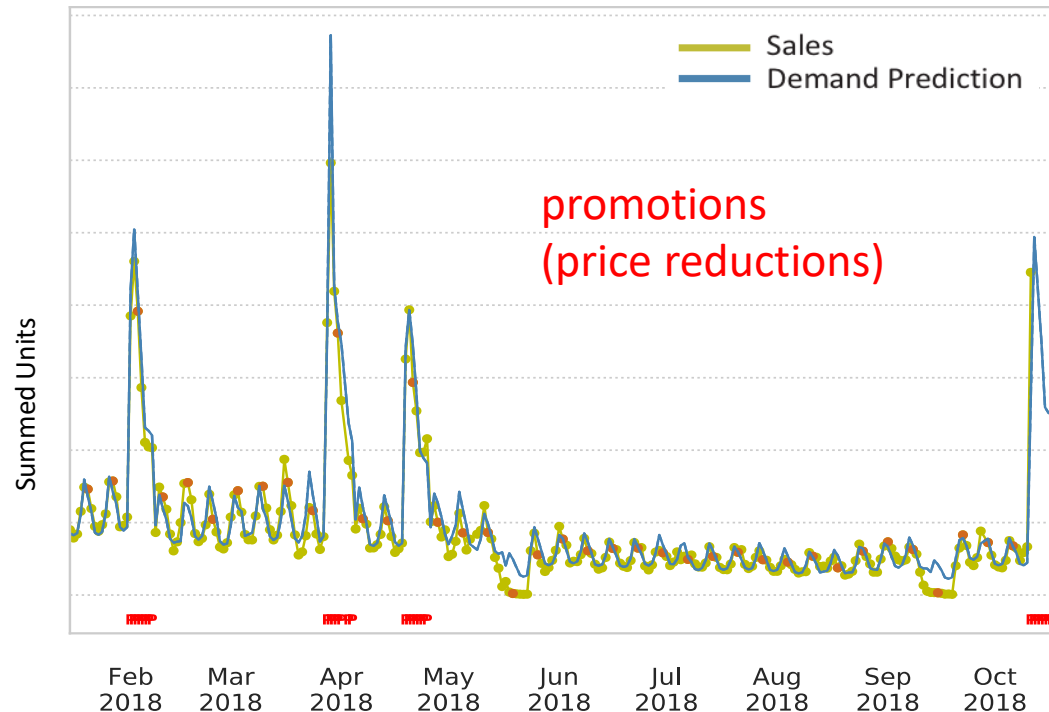
Retail Demand Forecasting & Replenishment

many individual time series to consider

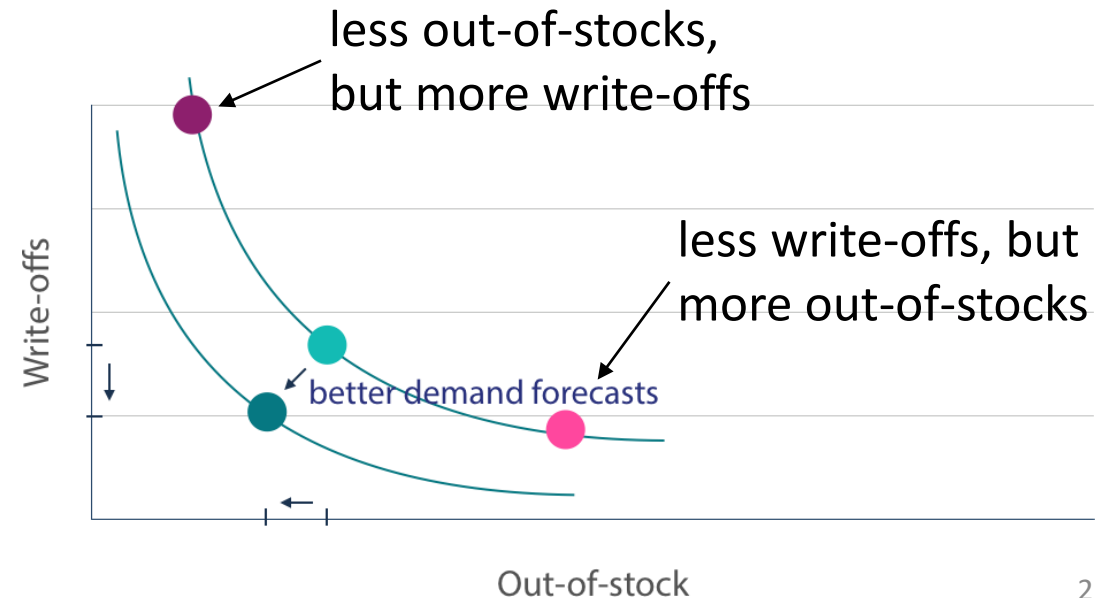
typical retail grocery chain:

- products (items): ~20k
- locations (stores): ~500
- daily/hourly aggregated sales

categorical features important:
products and locations → high cardinality



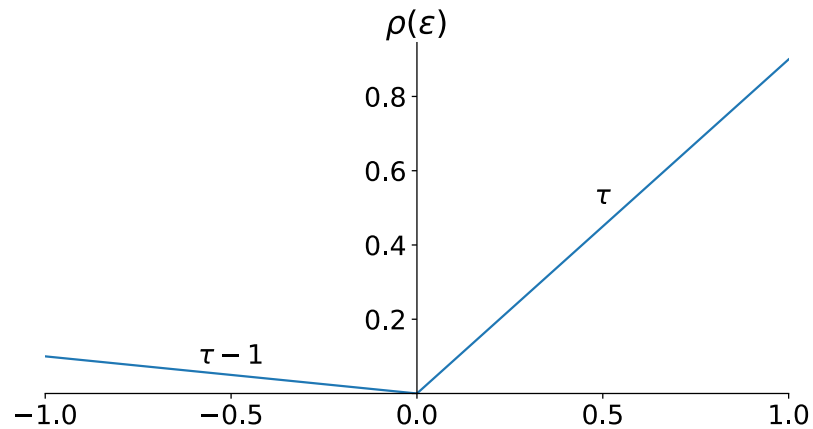
order optimization → choose demand quantile:



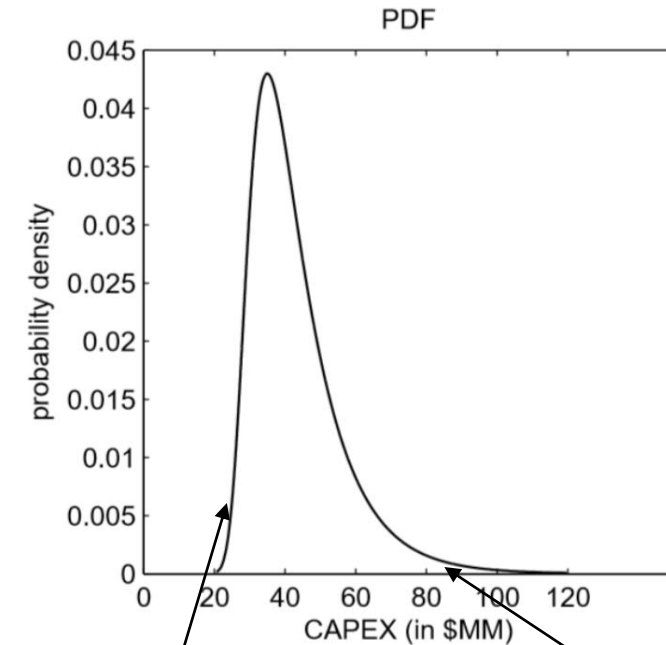
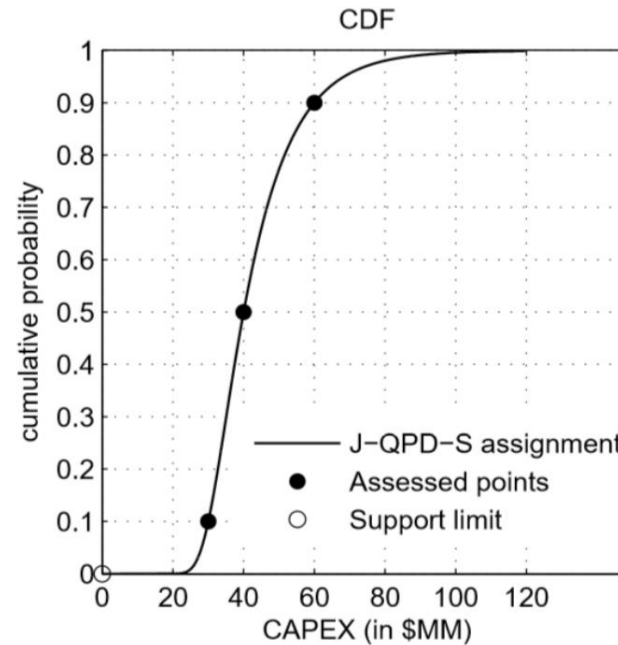
Prediction of Probability Distributions

use in quantile-parameterized distributions:

pinball loss for quantile predictions:



$$(1 - \tau) \sum_{y_i < \hat{q}_i} (\hat{q}_i - y_i) + \tau \sum_{y_i \geq \hat{q}_i} (y_i - \hat{q}_i)$$



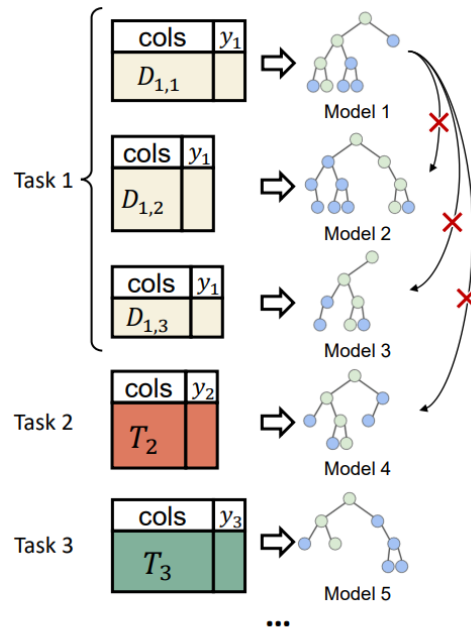
J-QPD

choose low demand
quantile for ordering:
high understock risk

choose high demand
quantile for ordering:
high overstock risk

Idea of Tabular Foundation Models

foundation models prevalent in vision and language (unstructured, homogenous data)
but not (yet) in structured/tabular, **heterogenous** data



Existing works:

- **one** model, **one** dataset;
- not transferable across datasets
- if transferable, needs finetuning on [source](#) each dataset

goal:

- pre-training across data sets and even different tasks
- finetuning on small data sets
- benefit from world knowledge in LLMs, for example in terms of data imputation

Concept Model: tabGPT

to overcome data integration challenge:

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice		
1	60	RL	69.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	MD	Normal	203500	
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	MD	Normal	181500	
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	MD	Normal	223500	
4	20	RL	68.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	MD	Abnormal	140000	
5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	MD	Normal	250000	
...	
1455	1456	60	RL	62.0	7817	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	8	2007	MD	Normal	179000
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MoPrv	NaN	0	2	2019	MD	Normal	210000
1457	1458	20	RL	66.0	9842	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	GdPrv	Shed	2500	5	2010	MD	Normal	200000
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	4	2010	MD	Normal	142125
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2008	MD	Normal	147500

1460 rows x 21 columns

convert to prompts for LLM calls

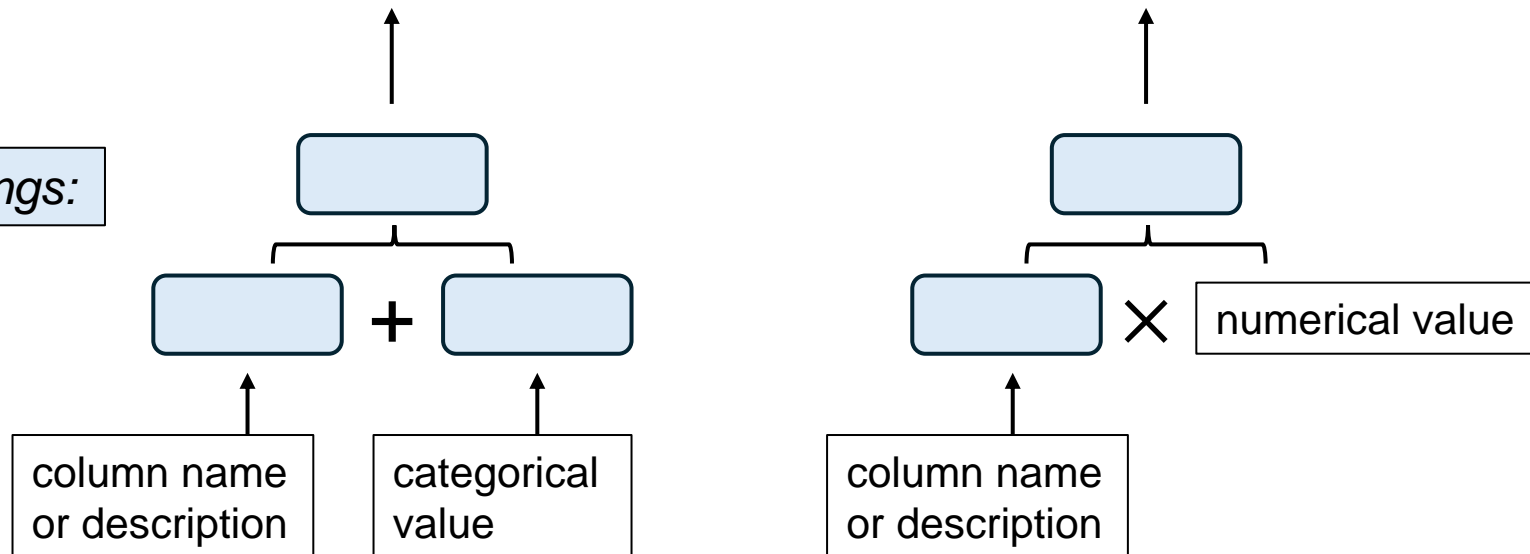
```
tensor([[[[ 0.2406, -0.0340, -0.5141, ..., 0.0476, -0.1114, -0.0198],  
[ 0.9594, 0.9598, 0.4653, ..., 1.1557, 1.3493, 1.0192],  
[ 0.6332, 0.3364, 0.8191, ..., 0.7699, 1.2227, 0.7292],  
...,  
[ 1.1688, 0.4768, 0.5724, ..., 0.7802, 0.9589, 1.1461],  
[ 0.8999, 0.5200, 0.7979, ..., 0.9777, 0.7836, 0.8079],  
[ 0.9854, 0.2225, 0.9218, ..., 0.9033, 0.9173, 0.8929]]]])
```

extract embeddings

MLP head

transformer encoder

embeddings:



Transformer for Numerical Data

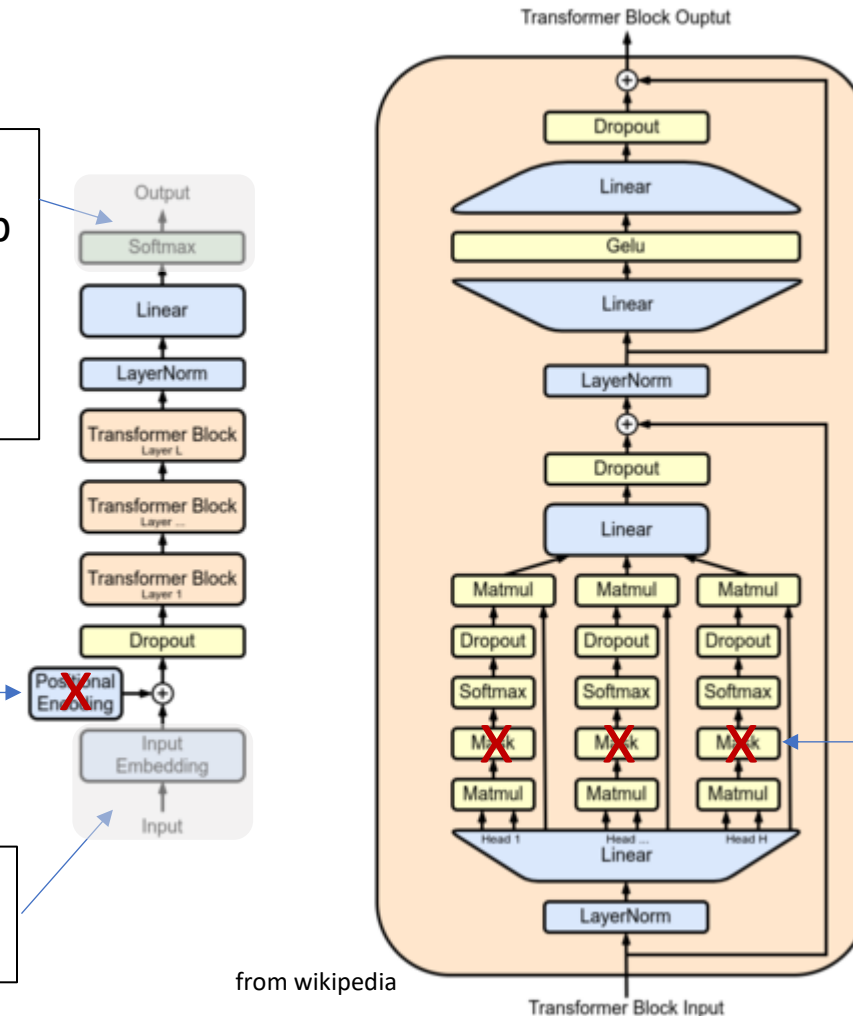
adaptions to GPT architecture:

replace language model head and loss function to reflect tabular prediction setup

also use average pooling over sequence instead of last element for prediction

drop positional encoding
(permutation invariance)

input previously generated
column embeddings directly



from wikipedia

drop causal masking
(permutation invariance)