Generalizations and extensions

- Using target variable in different tasks. Regression, multiclass
- Domains with many-to-many relations
- Timeseries
- Encoding interactions and numerical features

Regression and multiclass

- More statistics for regression tasks. Percentiles, std, distribution bins.
- Introducing new information for one vs all classifiers in multi class tasks

Many-to-many relations

- Cross product of entities
- Statistics from vectors

User_id	APPS	Target
10	APP1; APP2; APP3	0
11	APP4; APP1	1
12	APP2	1
100	APP3; APP9	0

LONG REPRESENTATION

User_id	APP_id	Target
10	APP1	0
10	APP2	0
10	APP3	0
11	APP4	1
11	APP1	1

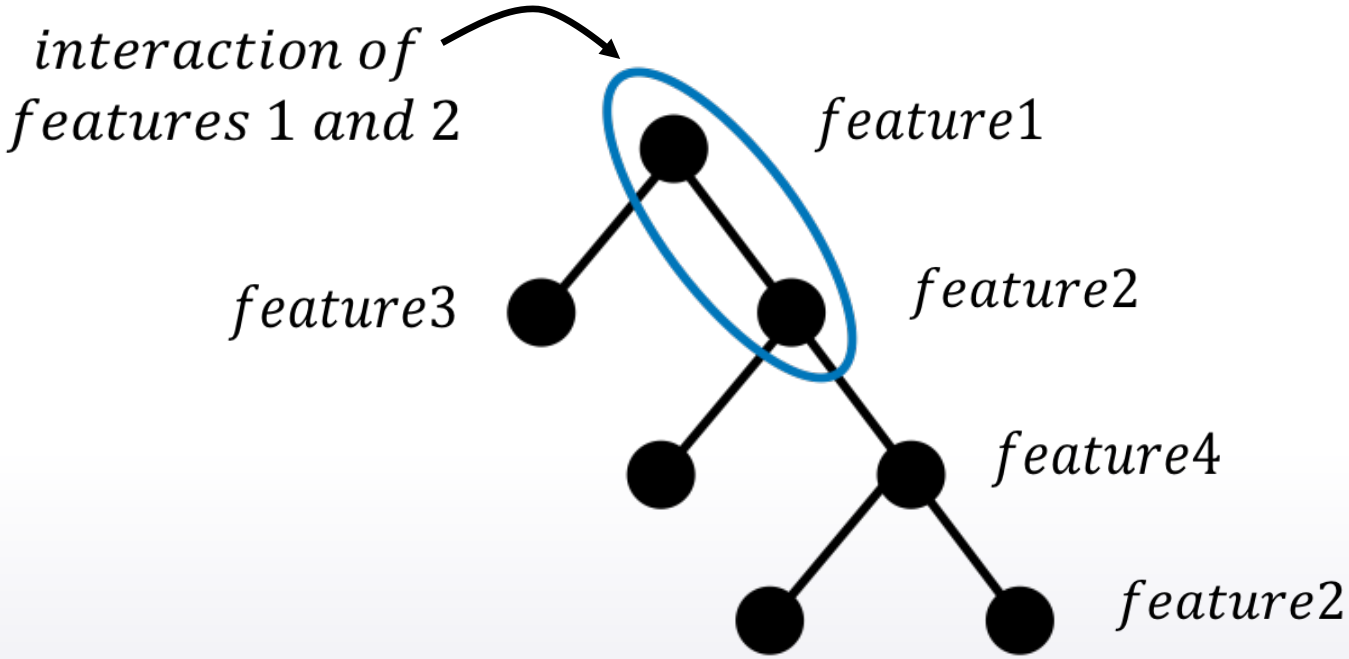
Time series

- Time structure allows us to make a lot of complicated features.
- Rolling statistics of target variable

Day	User	Spend	Amount	Prev_user	Prev_spend_avg
1	101	FOOD	2.0	0.0	0.0
1	101	GAS	4.0	0.0	0.0
1	102	FOOD	3.0	0.0	0.0
2	101	GAS	4.0	6.0	4.0
2	101	TV	8.0	6.0	0.0
2	102	FOOD	2.0	3.0	2.5

Interactions and numerical features

- Analyzing fitted model
- Binning numeric and selecting interactions



Amazon.com

Amazon.com - Employee Access Challenge Competition

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
cat_boost1.csv	a few seconds ago	0 seconds	0 seconds	0.91581

Complete

Jump to your position on the leaderboard ▾

Your most recent submission

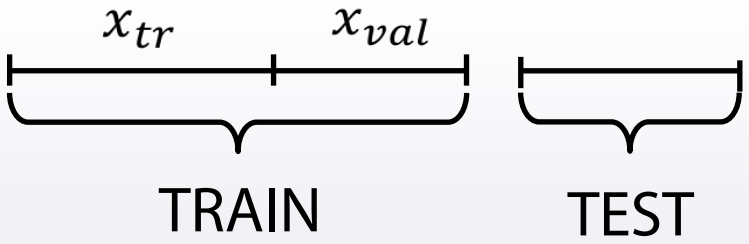
Name	Submitted	Wait time	Execution time	Score
lgb1.csv	just now	0 seconds	0 seconds	0.87209

Complete

Jump to your position on the leaderboard ▾

Correct validation reminder

- Local experiments:
  - Estimate encodings on X_tr
  - Map them to X_tr and X_val
  - Regularize on X_tr
  - Validate model on X_tr/ X_val split
- Submission:
  - Estimate encodings on whole Train data
  - Map them to Train and Test
  - Regularize on Train
  - Fit on Train



End

- Main advantages:
  - Compact transformation of categorical variables
  - Powerful basis for feature engineering
- Disadvantages:
  - Need careful validation, there a lot of ways to overfit
  - Significant improvements only on specific datasets