

Winning 24-hour Modeling Competitions

dexgroves.com/talks

Declan Groves

13 September 2016

Strategy

```
design etl
engineer response
while awake:
    engineer features    ---|
    remove features     |-- Maximize time spent here
    xgboost              |
    validate actions    ---|
optimize hyperparameters / go to bed
```

Response Engineering

- ▶ High performance gain per time investment
- ▶ Example: target a percentage

Feature Engineering

- ▶ Transformed X captures signal better than vanilla X

Feature Engineering

- ▶ Transformed X captures signal better than vanilla X
- ▶ Reverse generative process by thinking

Feature Engineering

- ▶ Transformed X captures signal better than vanilla X
- ▶ Reverse generative process by thinking
- ▶ ...or just throw stuff at wall

Example Feature Engineering Targets

- ▶ Dates
- ▶ High cardinality factors

Feature Pruning

- ▶ Random or unstable predictors do harm
- ▶ Low influence
- ▶ Unexpectedly high influence
- ▶ Counterintuitive trends

Validation

- ▶ Want a reliable feedback loop
- ▶ Want a fast feedback loop
- ▶ Every iteration, credibility is lost

Validation Hierarchy

- ▶ Holdout {fits quickly, overfits quickly}
- ▶ Cross-validation {fits slowly, overfits slowly}
- ▶ Leaderboard

Model Speedrunning

- ▶ Sparsity (if it makes sense)
- ▶ Fewer trees, greater learning rate (η)
- ▶ Early stopping
- ▶ Column subsampling
 - ▶ `colsample_bytree`
 - ▶ `colsample_bylevel`

Thanks for listening!

- ▶ dexgroves.com/talks