Winning 24-hour Modeling Competitions

dexgroves.com/talks

Declan Groves

13 September 2016

But why

- It's fun!
- Quick POC
- Take-home modeling exercises

Strategy

```
design etl
```

Strategy

```
design etl
engineer response
```

Strategy

```
design etl
engineer response
while awake:
```

Strategy

```
design etl
engineer response
while awake:
    engineer features
```

Strategy

```
design etl
engineer response
while awake:
    engineer features
    remove features
```

Strategy

```
design etl
engineer response
while awake:
    engineer features
    remove features
    xgboost
```

Strategy

```
design etl
engineer response
while awake:
    engineer features
    remove features
    xgboost
    validate actions
```

Strategy

```
design etl
engineer response
while awake:
    engineer features
    remove features
    xgboost
    validate actions
optimize hyperparameters (maybe go to bed)
```

Strategy

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

Response Engineering

- Transform y

Response Engineering

- Transform y
- High performance gain per time investment

Response Engineering

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

Feature Engineering

- Transform X

Feature Engineering

- Transform X
- Reverse generative process by thinking

Feature Engineering

- Transform X
- Reverse generative process by thinking
- . . . or just throw stuff at wall

Example Feature Engineering Targets

- Dates
  - EG: unix datetime of accident
  - Weekend, time-of-day, season, . . .
- High cardinality factors
  - EG: Make-model-modelyear
  - {Another talk}

Feature Pruning

- Random/unstable predictors do harm

Feature Pruning

- Random/unstable predictors do harm
- Low influence

Feature Pruning

- Random/unstable predictors do harm
- Low influence
- Unexpectedly high influence

Feature Pruning

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

Validation

- Feedback loops are dangerous

Validation

- Feedback loops are dangerous
- Every iteration, credibility is lost

Validation Hierarchy

1. Holdout {fits quickly, overfits quickly}
2. Cross-validation {fits slowly, overfits slowly}
3. Leaderboard {overfit at your peril}

Model Speedrunning

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

Thanks for listening!

- dexgroves.com/talks