Winning 24-hour Modeling Competitions

github.com/dexgroves/talks

Declan Groves

13 September 2016

Intro

- Allstate

Intro

- Allstate
- $\approx$ 220 data people

Intro

- Allstate
- $\approx$ 220 data people
- Quarterly kaggles

Intro

- Allstate
- ≈ 220 data people
- Quarterly kaggles
- 24/48hr time limit

Strat

```
design etl
engineer response
while time remains:    ___
    engineer features    |
    remove features      |-- Maximize time spent here
    xgboost              |
    validate actions  ___|
optimize hyperparameters
```

Response Engineering

- High performance gain per time investment
- Examples:
    - Target a Δ
    - Target a percentage
    - Target a transformation, $\log(y)$ etc

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

Feature Engineering

- Combine covariates, custom interactions
- Map covariates to a new space
- Treat high-cardinality variables

Feature Pruning

- Random or unstable predictors do harm
- Low influence
- Unexpectedly high influence
- Counterintuitive trends (marginal effects)

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 (`eta`)
- Early stopping
- Column subsampling
  - `colsample_bytree`
  - `colsample_bylevel`

Thanks for listening!

- github.com/dexgroves/talks