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
- $\approx$ 220 data people
- Quarterly kaggles
- 24/48hr time limit

Strat

```
design etl
while time remains:    ___
    engineer features     |
    remove features       |-- Maximize time spent here
    fit a gbm             |
    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 Engineering Examples

- Dates $\rightarrow$ time since something
- Counts $\rightarrow$ fractions
- Ordinals $\rightarrow$ ordinals vs group averages

Feature Pruning

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

Collaborating

```
add_dex_features <- function(df) {
  ...
  df
}
```

```
source("add_dex_features.R")
...
df <- add_dex_features(df)
df <- add_cdc_features(df)
df <- add_jeremy_features(df)
df <- add_jesse_features(df)
...
```

Validation

- Want a fast, reliable feedback loop
- Going quickly around loop is good
- Every iteration, credibility is lost

Validation Hierarchy

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

Model Speedrunning

- Fewer trees, higher shrinkage (`eta`)
- Column subsampling
  - `colsample_bytree`
  - `colsample_bylevel`

Thanks for listening!

- github.com/dexgroves/talks