Winning 24-hour Modeling Competitions github.com/dexgroves/talks

Declan Groves

13 September 2016

► Allstate

- ► Allstate
- ightharpoonup pprox 220 data people

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

- Allstate
- ightharpoonup pprox 220 data people
- Quarterly kaggles
- ▶ 24/48hr time limit

Strat

Response Engineering

- High performance gain per time investment
- Examples:
- ▶ Target a ∆
- Target a percentage
- ▶ Target a transformation, log(y) etc

▶ Transformed X captures signal better than vanilla X

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

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

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

Feature Engineering Examples

- lackbox Dates o time since something
- ▶ Counts → fractions
- lackbox Ordinals ightarrow 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
}</pre>
```

```
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)
...</pre>
```

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