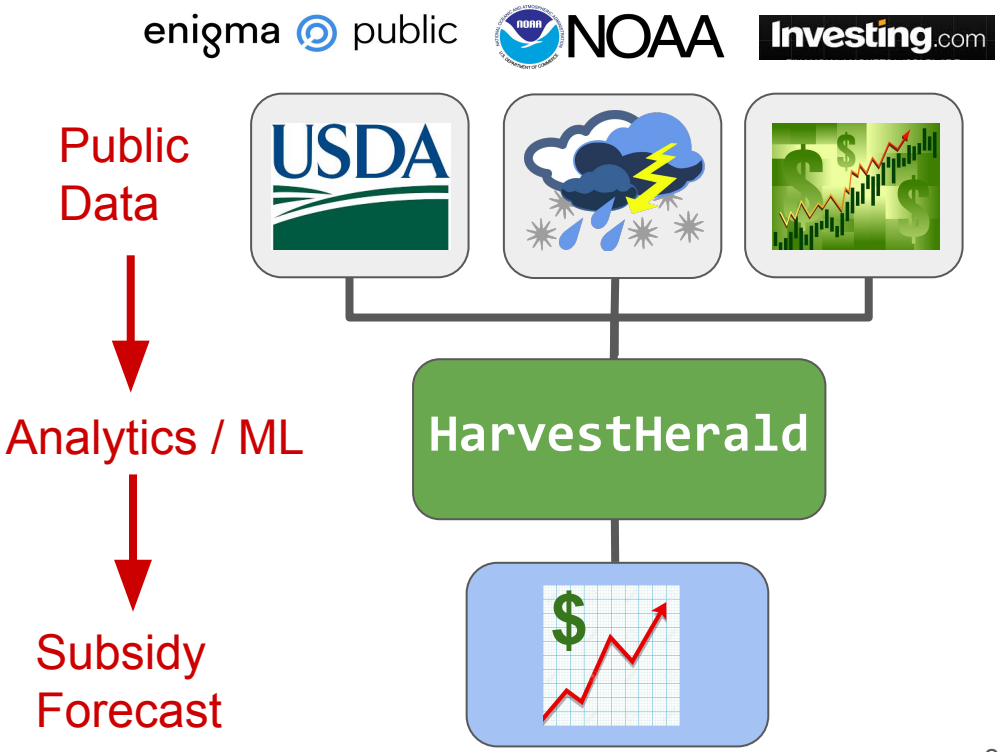
Using Public Data to Forecast USDA Farm Subsidies

Derek Araujo • Insight Data Science • Fall 2017

Crop Subsidies: ~ $25 B annually

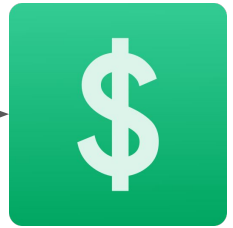
US Gov:
- Budgeting

Insurers:
- Subsidy = risk mitigation



Data Challenges:

- Unknown time of loss →
  - Unknown sale price, vol.
- Per county: subsidy is sparse/noise-dominated

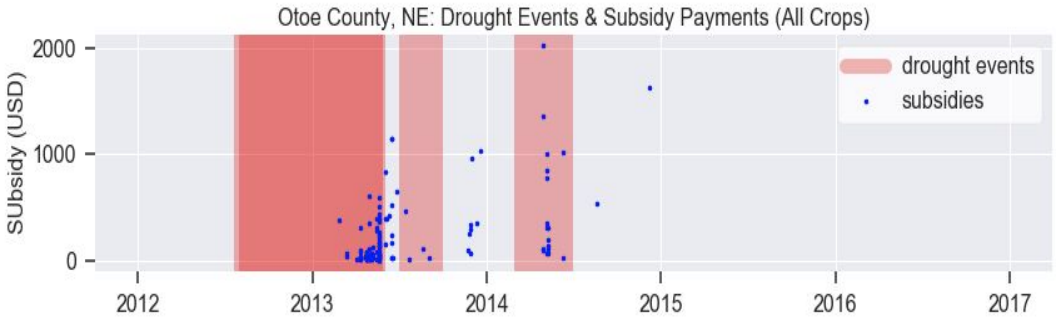


t_{loss} ← ? → $t_{payment}$

→ Aggregate for each crop:

- Monthly bins
- Avg subsidy (all counties)
- Focus on drought events ($$)

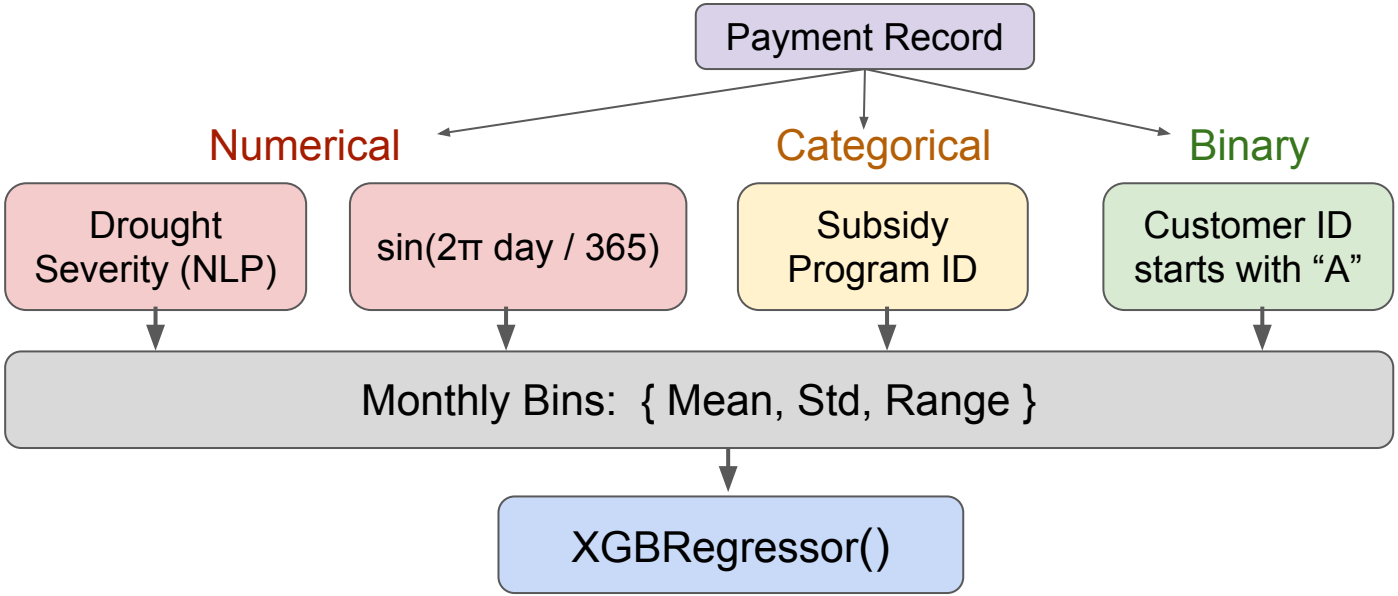
→ <u>Few data points</u> for forecasting



Model-Driven Feature Engineering:

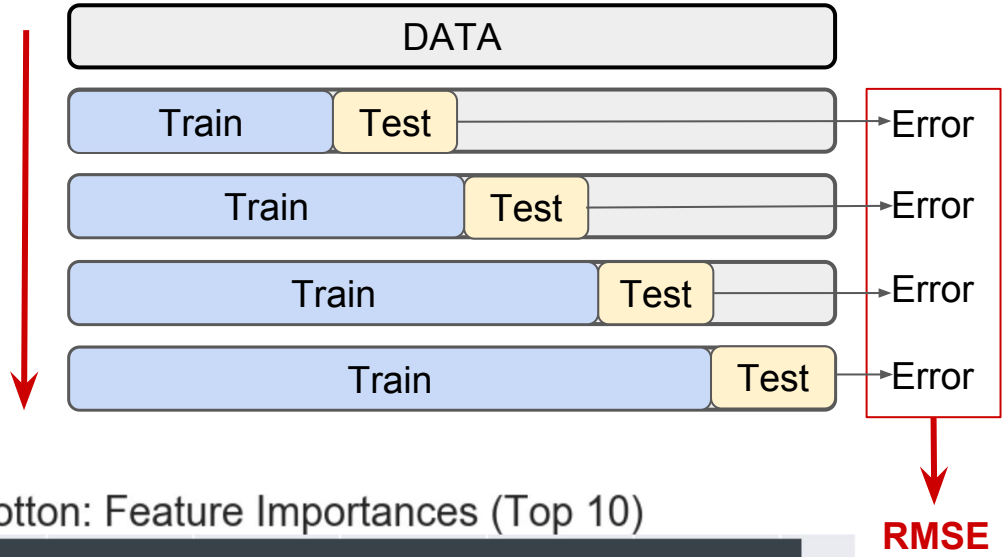
Two models:

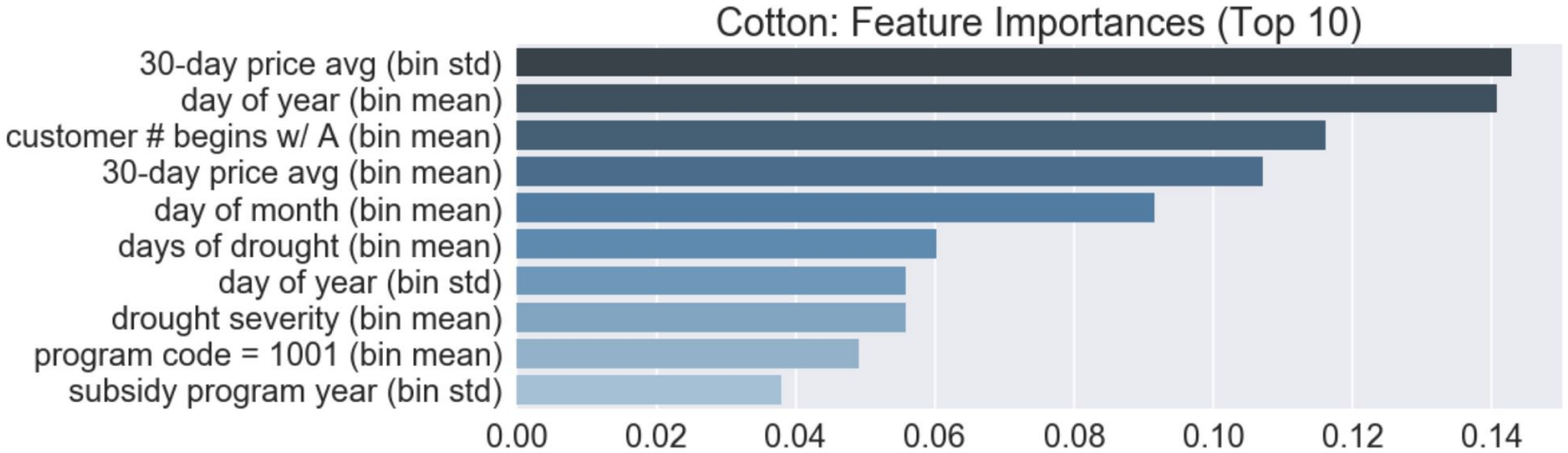
- ARIMA: Auto-Regressive Integrated Moving Average (no engineering)
- XGBoost Regression Tree: > 600 engineered features



Methodology:

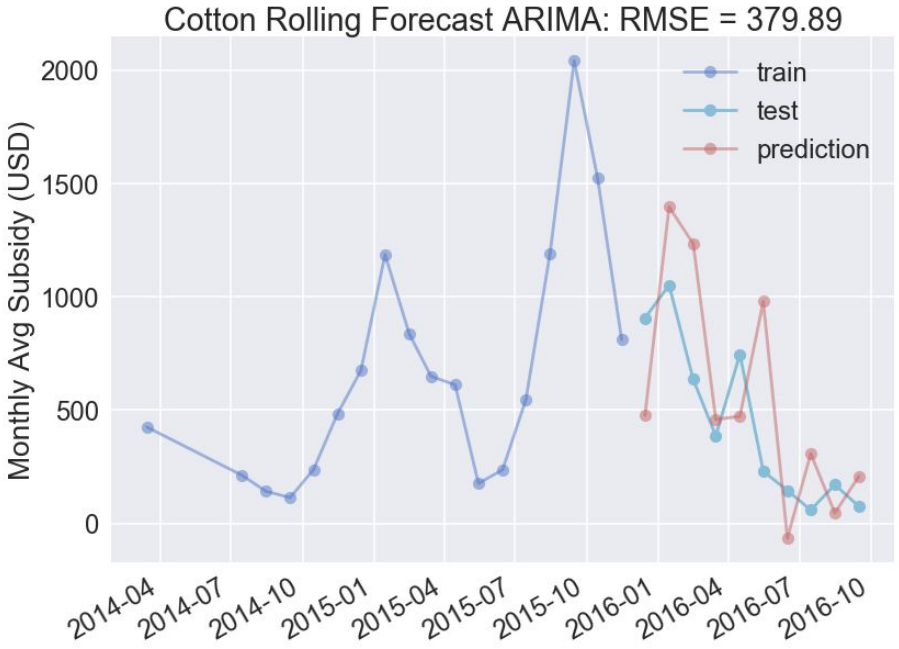
- Target: log(Avg Monthly Subsidy)
- Hyperparams: grid search
- Validation: walk-forward forecast, expanding window
- XGB: examine feature importance



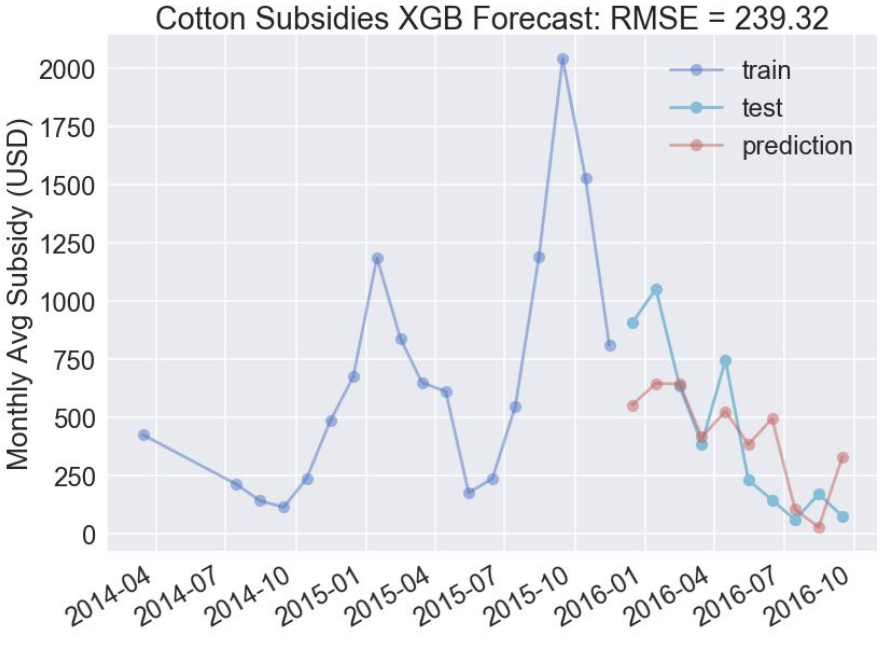


Results: Cotton

ARIMA:

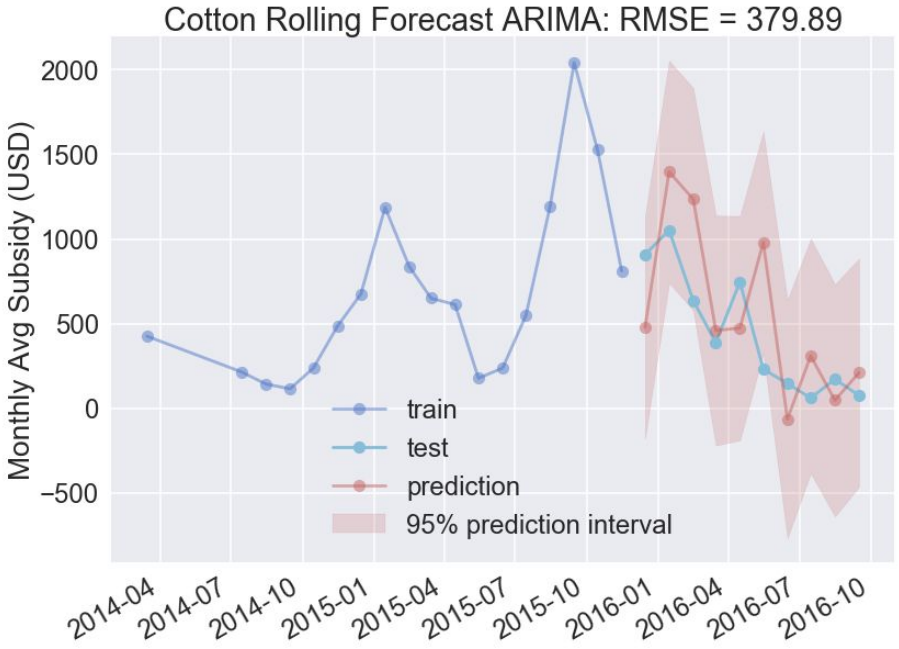


XGB Regressor:

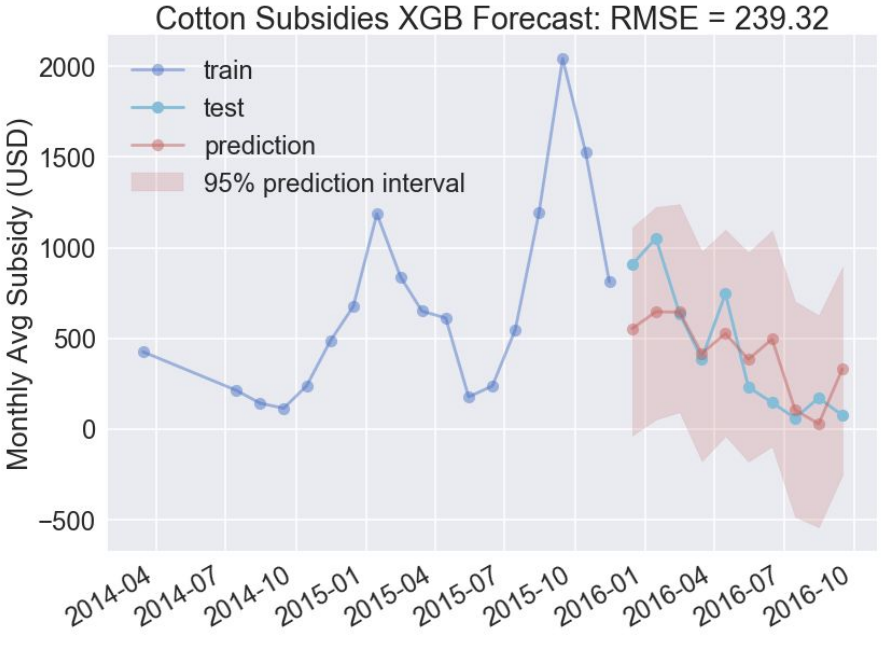


Results: Cotton

ARIMA:



XGB Regressor:



Utility Assessment / Improvements:

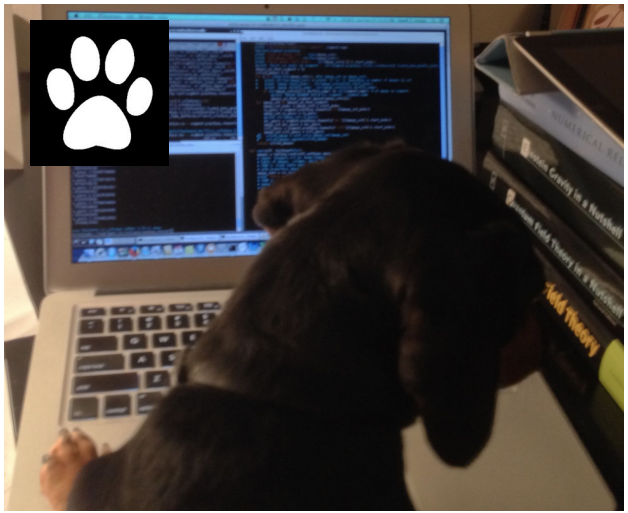
- Wide error bars; but indicative of trend
- Find missing data: (volume, date of sale)
- Other paths forward:
  - ARIMA → ARIMAX (add exogenous variables)
  - Feature selection / PCA

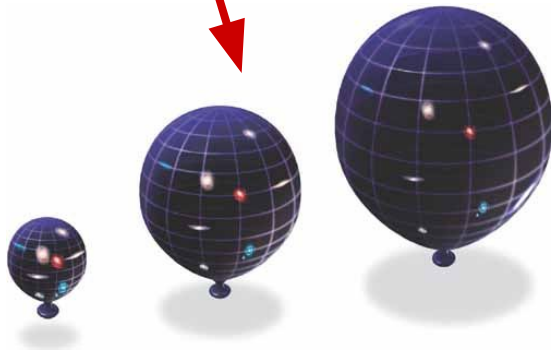
About Derek:

Harvard College
A.B., Physics

Harvard Law School
J.D.

Columbia University
Ph.D., Physics





Inflationary Big Bang Cosmology