HEDGE FUND X: FINANCIAL MODELING CHALLENGE

FINAL ML PROJECT

STA 9891 UMA
DECEMBER 13, 2021

MACHINE LEARNING FOR DATA MINING
PROF. KAMIAR RAHNAMA RAD

GROUP 1:

MICHAEL S. BONETTI
https://github.com/mbonetti-nyc/Hedge-Fund-X

ZICKLIN BUSINESS 100 YEARS

BRIEF DESCRIPTION

HEDGE FUND X | SIGNATE COMPETITION ON KAGGLE

About this dataset

Summarizes heterogenous set of features

- Signate, a Japanese AI / Data Analytics platform
- Sample of a training dataset used in a DeepAnalytics competition in 2018
- Predict the future movement of financial products (unknown)
  - Datapoints: 10,000 *(stated & observed)* (n)
  - Response ($target$): predictive binary classification (0 or 1)

Attributes

- 91 attributes (p)
  - 88 predictive attributes used (all numerical): 2 non-predictive, 1 target
    - 2 attributes removed $(ID, train9)$
- No missing values were found upon preliminary analysis

DeepAnalytics

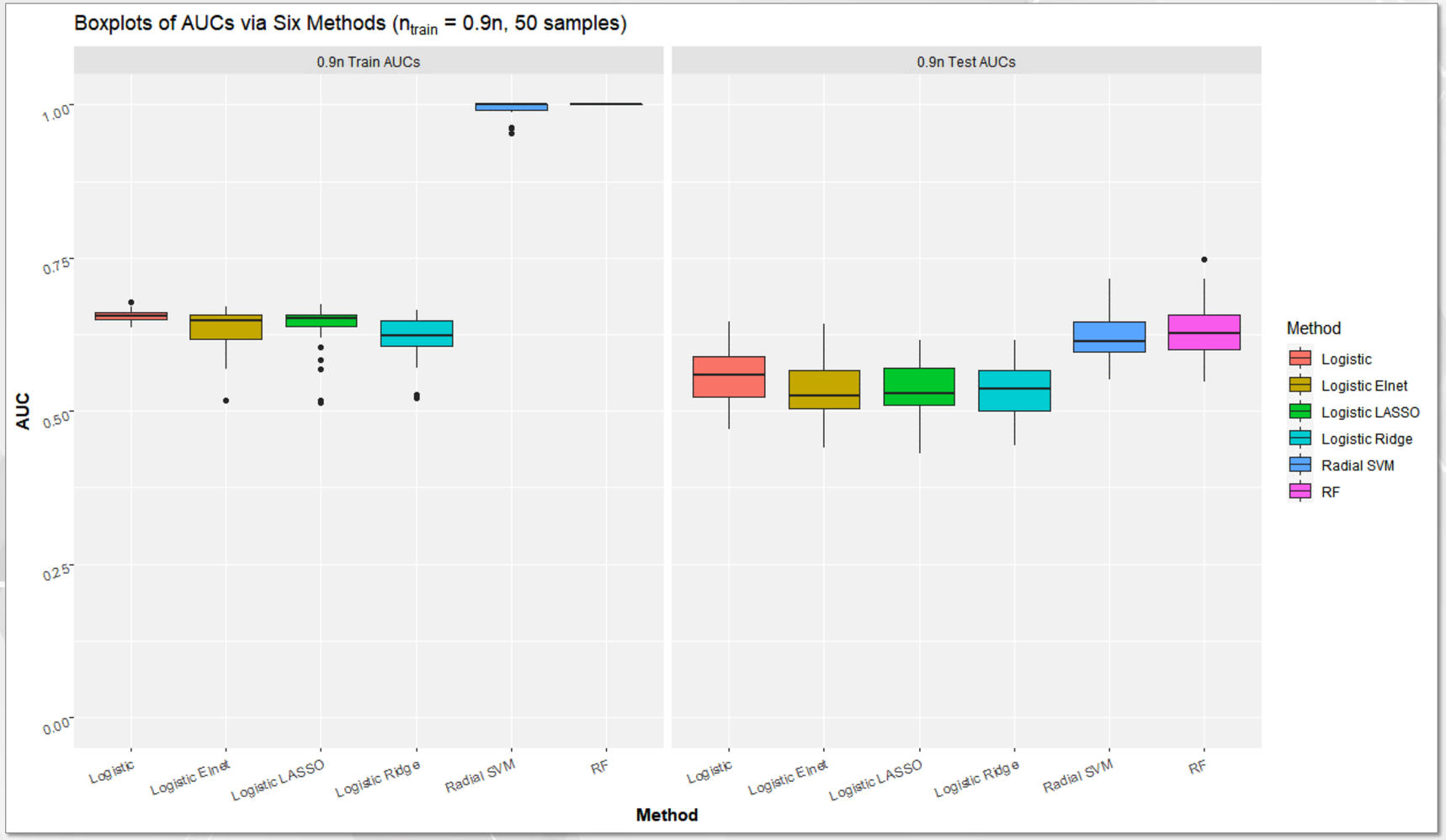
For this project

- 1,000 observation (n_{subset}) subset taken
  - Speeds up processing, troubleshooting, and improvements
  - 2 runs
    - R1: 1K static obs. for reproducible results
    - R2: 1K randomly-chosen obs. for varying results
- Imbalance ratio (IR) *(* denotes majority class)*
  - Original dataset is relatively balanced
  - $n_0 = 4{,}994, n_1 = 5{,}006 \ \therefore \ \frac{n_0}{n_1} = 99.76\%$
  - Run 1: $n_0^* = 512, n_1 = 488 \ \therefore \ \frac{n_1}{n_0} = 95.31\%$
  - Run 2: $n_0 = 473, n_1^* = 527 \ \therefore \ \frac{n_0}{n_1} = 89.75\%$

ZICKLIN BUSINESS 100 YEARS

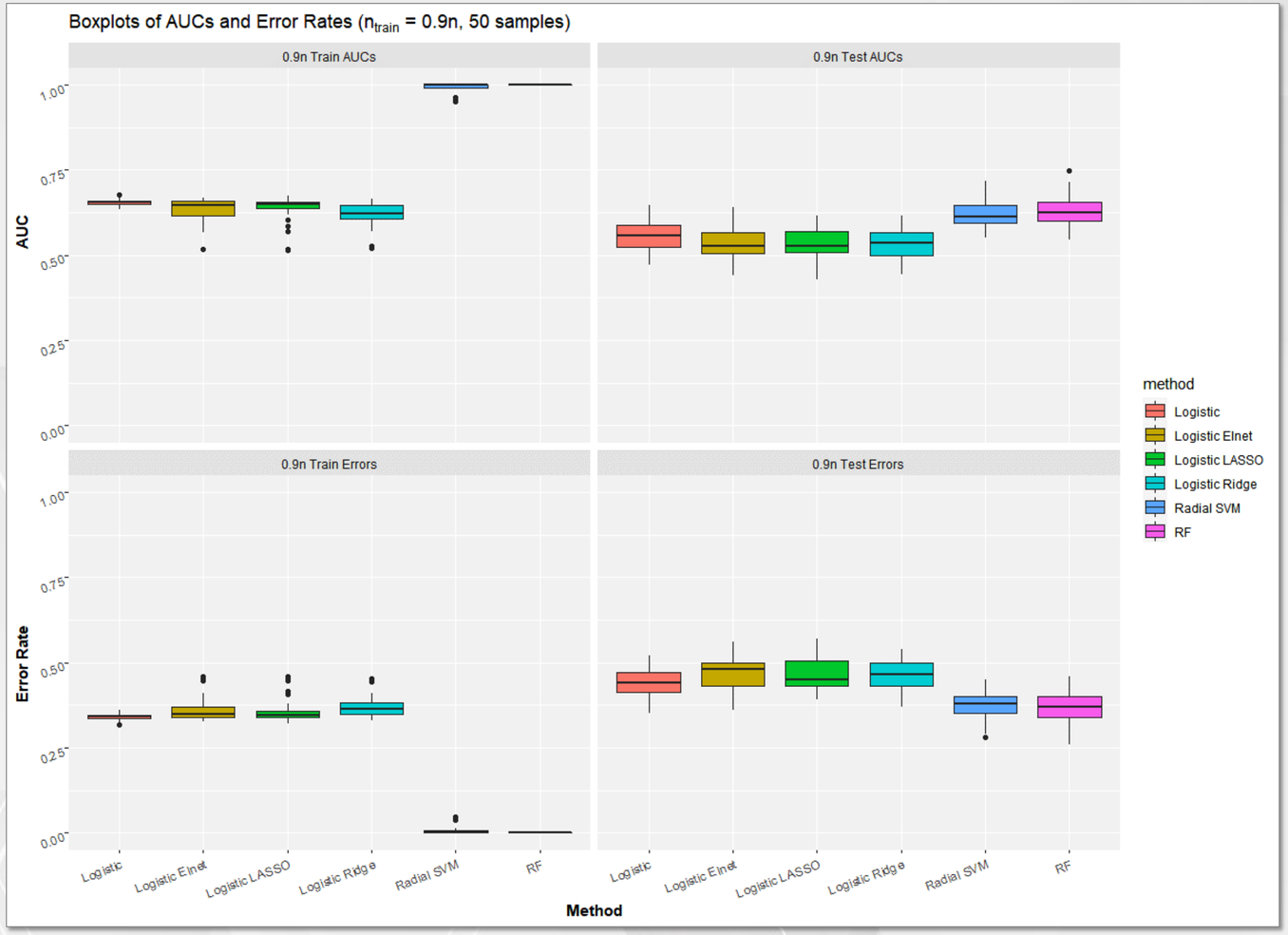
BOXPLOTS

$(0.9n)$ | AUCs via 6 ML METHODS



BOXPLOTS

$(0.9n)$ | AUCs & ERROR RATES via 6 ML METHODS



Observations:

- **0.9n AUC training**
  - SVM and RF medians at, or near, 1
- **0.9n AUC testing**
  - Larger variances overall compared to AUC training boxplots
  - Logistic, Elnet, LASSO, and Ridge medians are close; SVM and RF higher

---

- **0.9n training errors**
  - Near mirror-image of AUC training boxplots, with Elnet variance slightly smaller
  - SVM and RF medians at, or near, 0
- **0.9n testing errors**
  - Larger variances overall; medians for SVM and RF are still lower

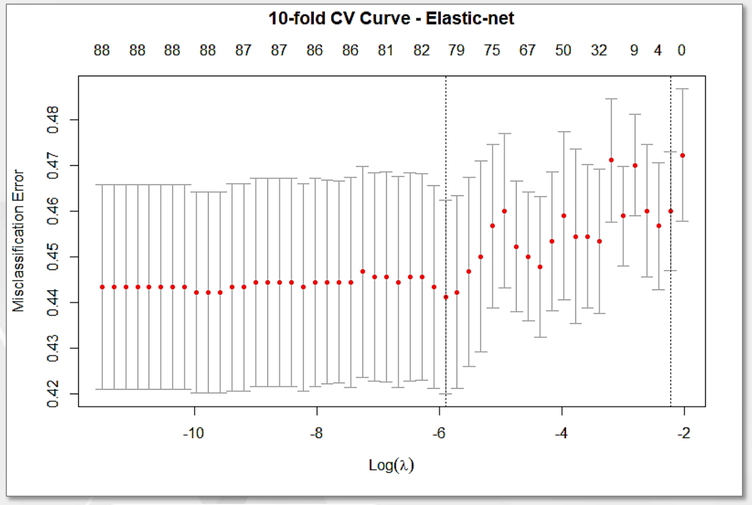
CV CURVES

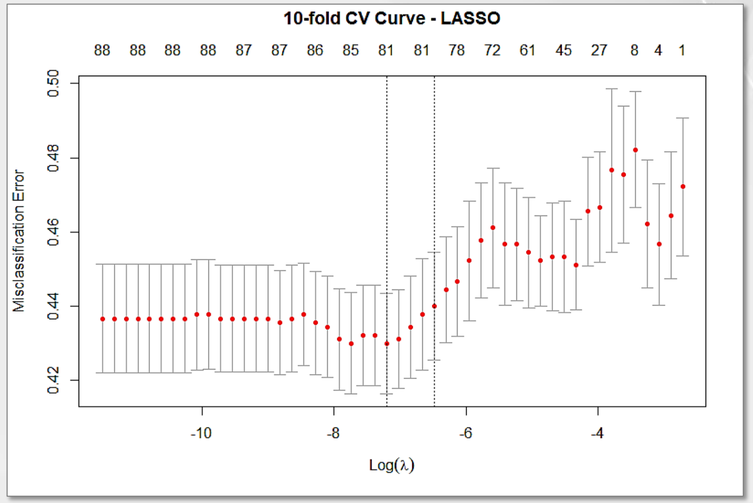
10-fold (0.9n) | ELASTIC-NET, LASSO, & RIDGE

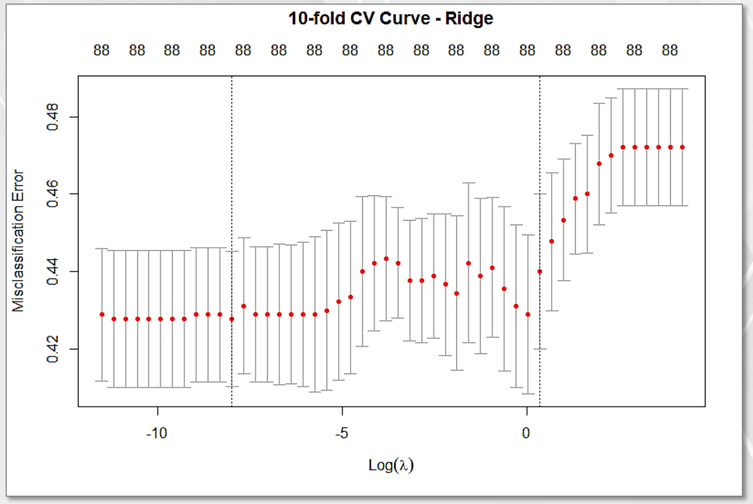
EN Run 1: 0.910 secs.
EN Run 2: 0.780 secs.

LASSO Run 1: 0.910 secs.
LASSO Run 2: 0.830 secs.

Ridge Run 1: 0.860 secs.
Ridge Run 2: 0.780 secs.



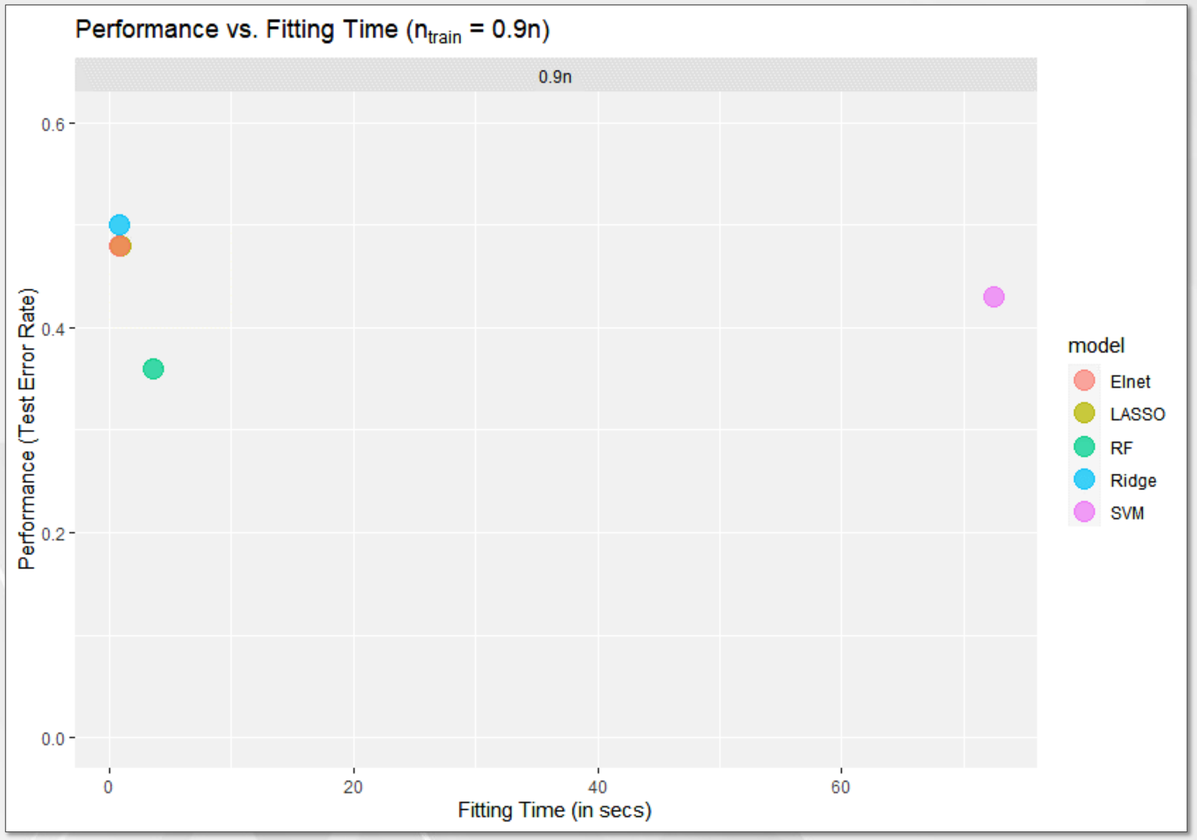




Run 1: Using static 1,000 observation subset
Run 2: Using 1,000 randomly-selected observation subset

All CV plots were generated from Run 2

PERFORMANCE *vs.* RUNTIME

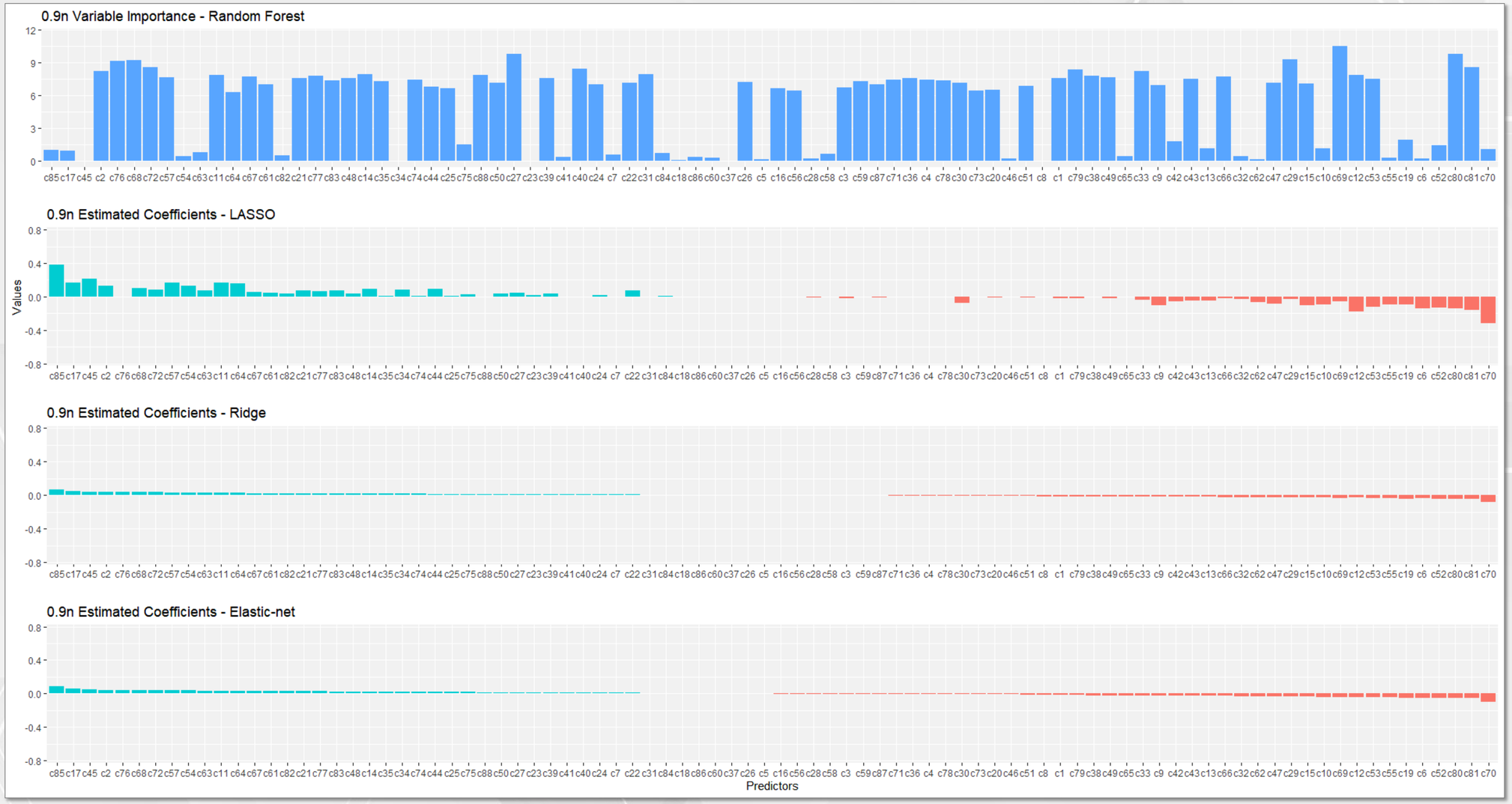


Method	Performance *(error rates)*		AUC_{test}		Runtime *(in secs)*	
	Run 1	Run 2	R1	R2	R1	R2
Elastic-net	0.390	**0.480**	0.523	**0.535**	1.000	**0.840**
LASSO	0.430	**0.480**	0.518	**0.537**	0.980	**0.910**
Random Forest	0.380	**0.360**	0.578	**0.629**	3.640	**3.610**
Ridge	0.390	**0.500**	0.528	**0.535**	0.940	**0.860**
SVM	0.380	**0.430**	0.605	**0.619**	71.08	**72.40**

Observations:

- **Avg. Performance: 0.422**
- **Fast runtimes**
  - However, the AUCs are no better than 50 / 50
- **Trade-off? *Slightly*.**
  - RF consistently performed the best, albeit with a runtime of +3.50 secs.
  - SVM takes the longest, with minimal performance improvement

VARIABLE IMPORTANCE



VARIABLE IMPORTANCE

CLOSING REMARKS

HEDGE FUND X | SIGNATE COMPETITION ON KAGGLE

Variable Importance (Top 3)

- **Positive influence** ↗
  c85, c17, c45
- **Negative influence** ↙
  c70, c81, c80

However, …

- **Too many unknowns**
  - Cannot tell which specific funds / stocks affected performance
- **No better than a coin toss**
  - Except for RF and SVM, AUCs are no better than 50 / 50

Improvements can be made…

- **RF & SVM (best performers), AdaBoost, SGD, kNN, NB**
- **… but the financial market is unpredictable!**
  - Therefore, this makes financial modeling decidedly more difficult to perform

THANK YOU!

Q&A

https://github.com/mbonetti-nyc/Hedge-Fund-X