Spam or not spam?

By Wei Bin Li, Nika Kondzhariya and Yulia Starovoytova

Project Description

Goal: classify emails for spam and non-spam

Techniques used: Lasso, Elastic-Net, Ridge and Random Forest

Data Structure

**Dataset:** The email Spam dataset collected from the UC Irvine Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/Spambase

Target Class Variable: denotes whether the e-mail was considered spam (1) or not (0)

Dimension:
- p = 57
- n = 4601:
  n1(spam) = 1813 (39.5%)
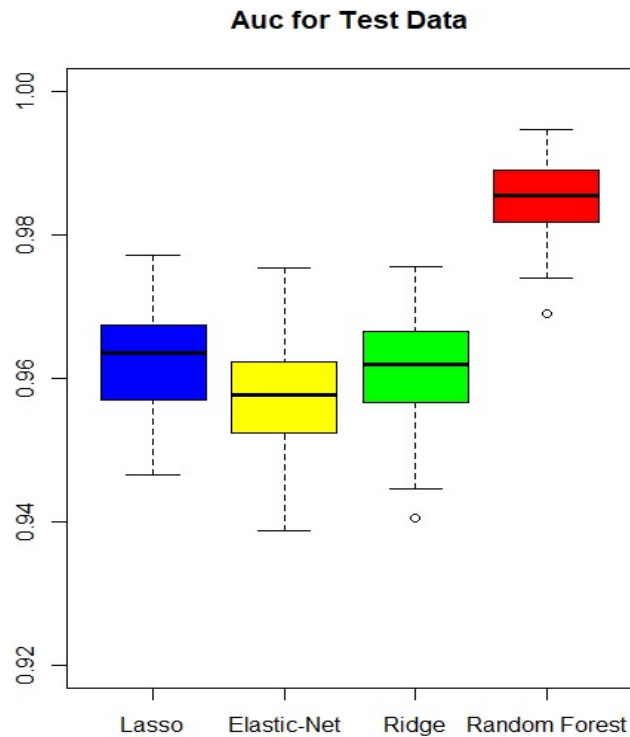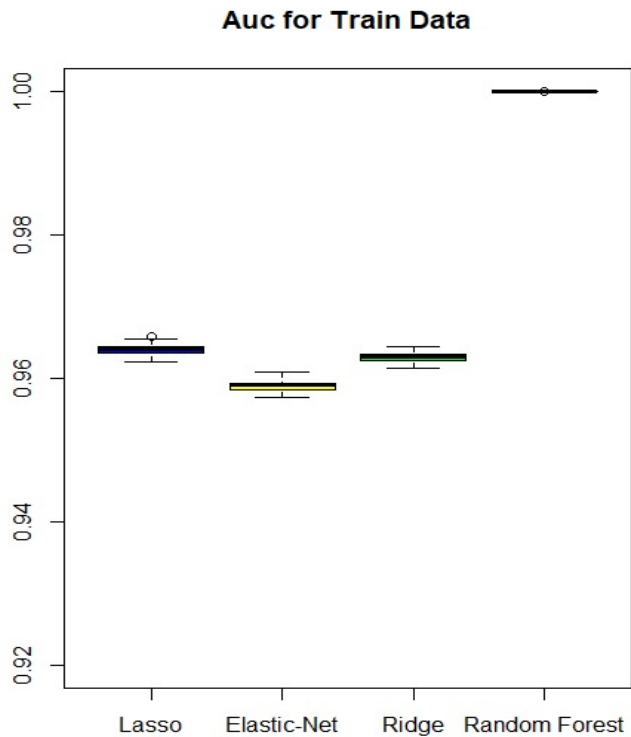 n0(non-spam) = 2788 (60.5%)

Features:
- % of a particular **word** occurring in the e-mail: "free", "credit", "money", "receive", "remove" and etc.

- % of a particular **character** occurring in the e-mail: ";", "(", "[", "!", "$", "#"

- length of sequences of consecutive **capital** letters
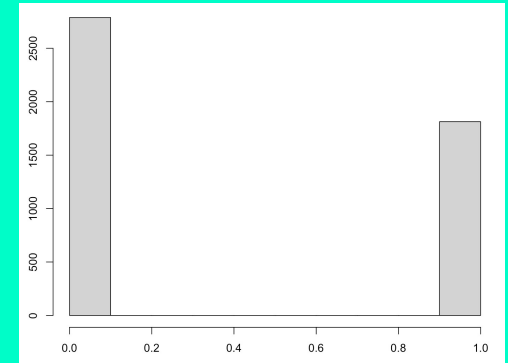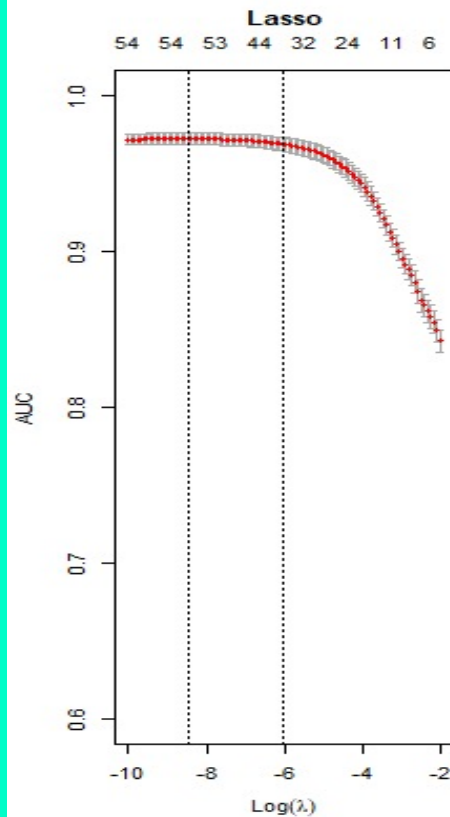
AUC result comparison for train and test set

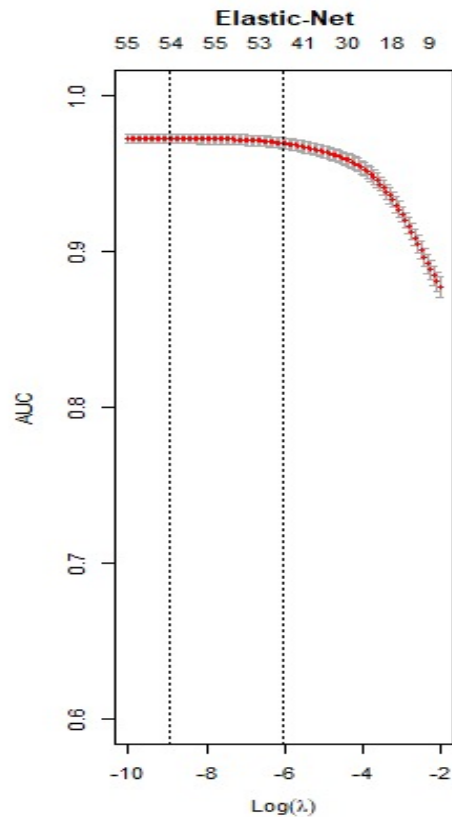**Train Data (n=4141)**

Test Data (n=460)

10 fold cross validation curves for one of the 50 samples

4.6 sec

Lasso

54 54 53 44 32 24 11 6

AUC

1.0
0.9
0.8
0.7
0.6

-10   -8   -6   -4   -2

Log(λ)

4.4 sec

Elastic-Net

55 54 55 53 41 30 18 9

AUC

1.0
0.9
0.8
0.7
0.6

-10   -8   -6   -4   -2

Log(λ)

6.8 sec

Ridge

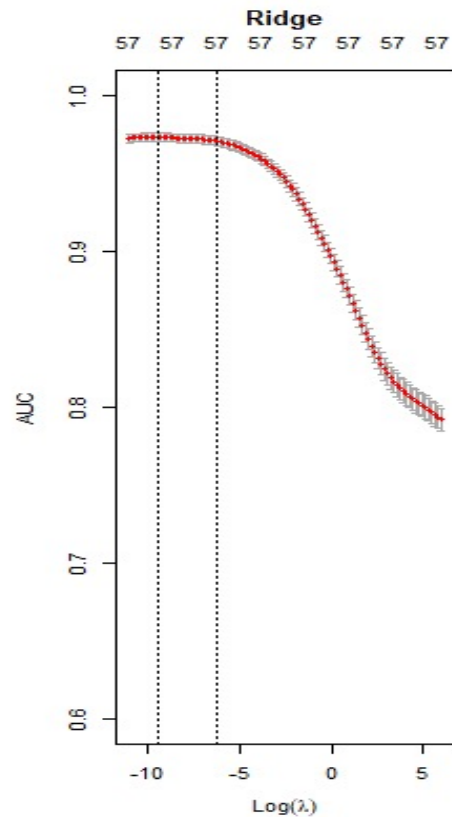57 57 57 57 57 57 57 57

AUC

1.0
0.9
0.8
0.7
0.6

-10   -5   0   5

Log(λ)
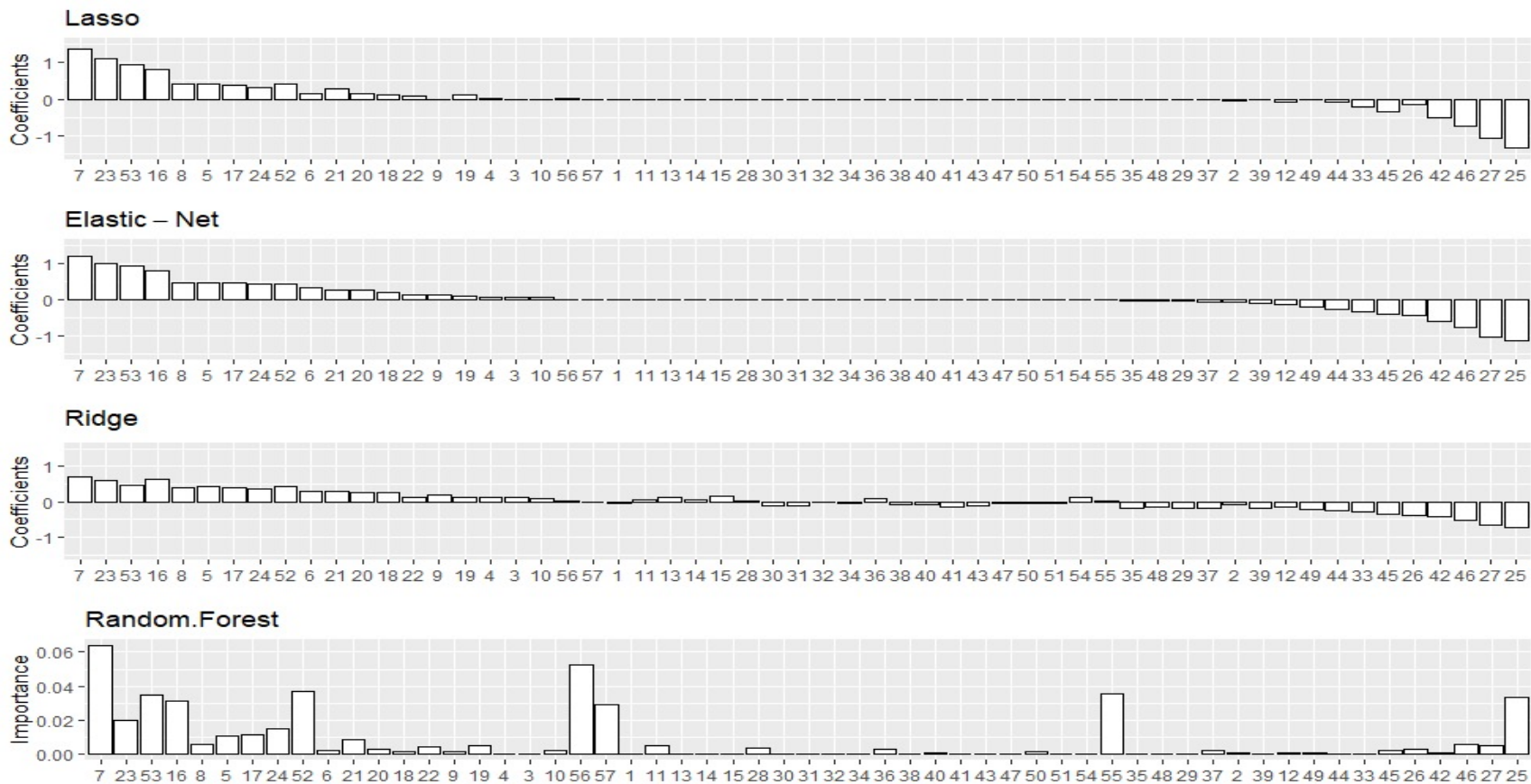
# Bar Plots of Estimated coefficients

More on coefficients

Lasso/Elastic-Net/Ridge:

- **Positive - spam:**
  - Word frequency "Remove"
  - Word frequency "000"
  - Word frequency "Free"

- **Negative - non spam:**
  - Word frequency "HP"
  - Word frequency "George"
  - Word frequency "Edu"

Random Forest:
  - Length of longest uninterrupted sequence of capital letters
  - Average length of uninterrupted sequences of capital letters
  - Sum of length of uninterrupted sequences of capital letters
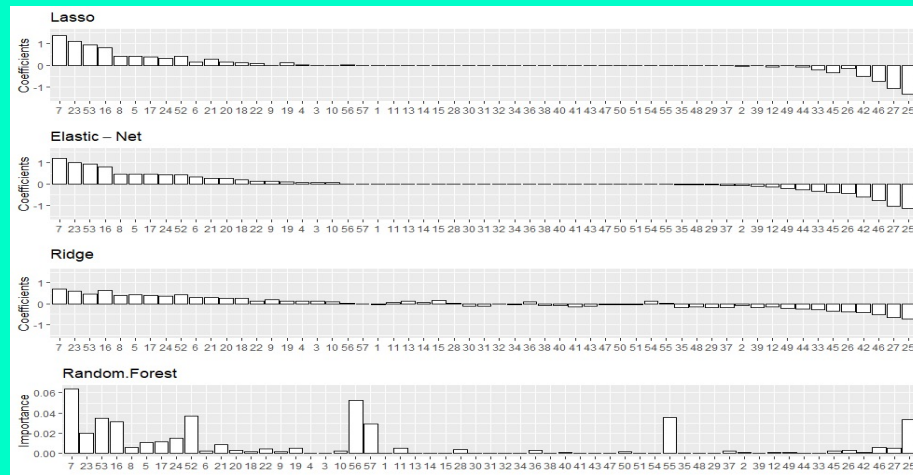
Model accuracy and run time

Model	90% CI on 50 Test AUCs*	Median of 50 Test AUCs	Time**
Random Forest	[0.977, 0.993]	0.992	23.3 sec
Lasso	[0.951, 0.972]	0.963	4.9 sec
Ridge	[0.949, 0.971]	0.961	7.1 sec
Elastic Net	[0.944, 0.968]	0.958	4.8 sec

Trade-off: The better performance – the more time required to build

Thank you for attention! Do you have any questions?

* – 90% test AUC based on the 50 samples with 90% confidence interval
** – The time it takes to fit the model on full dataset