Spam Emails Classification

Chau Hoang

STA 9891, Baruch College, Fall 2020

Overview

- In March 2016, the volume of spam emails is reported at 22,890,956 (Kaspersky Lab)
- The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time
- Users usually feel very irritating and may suffer from financial loss as victims of internet scams and other fraudulent practices (i.e: disclose sensitive information, credit card number, etc.)
- Source: ScienceDirect

Spambase Data Set & ML models used

- From UCI Machine Learning Repository
- n = 4601, p = 57, no missing values
- n+ = 1813, n- = 2788, n+/n- = 0.65
- 1 target variable: “class” (0 = non-spam, 1: spam)
- 57 Predictors:

+ 48 attributes of type “word_freq_WORD”: % of words in the email that match WORD)

+ 6 attributes of type “char_freq_CHAR”: % of characters in the email that match CHAR

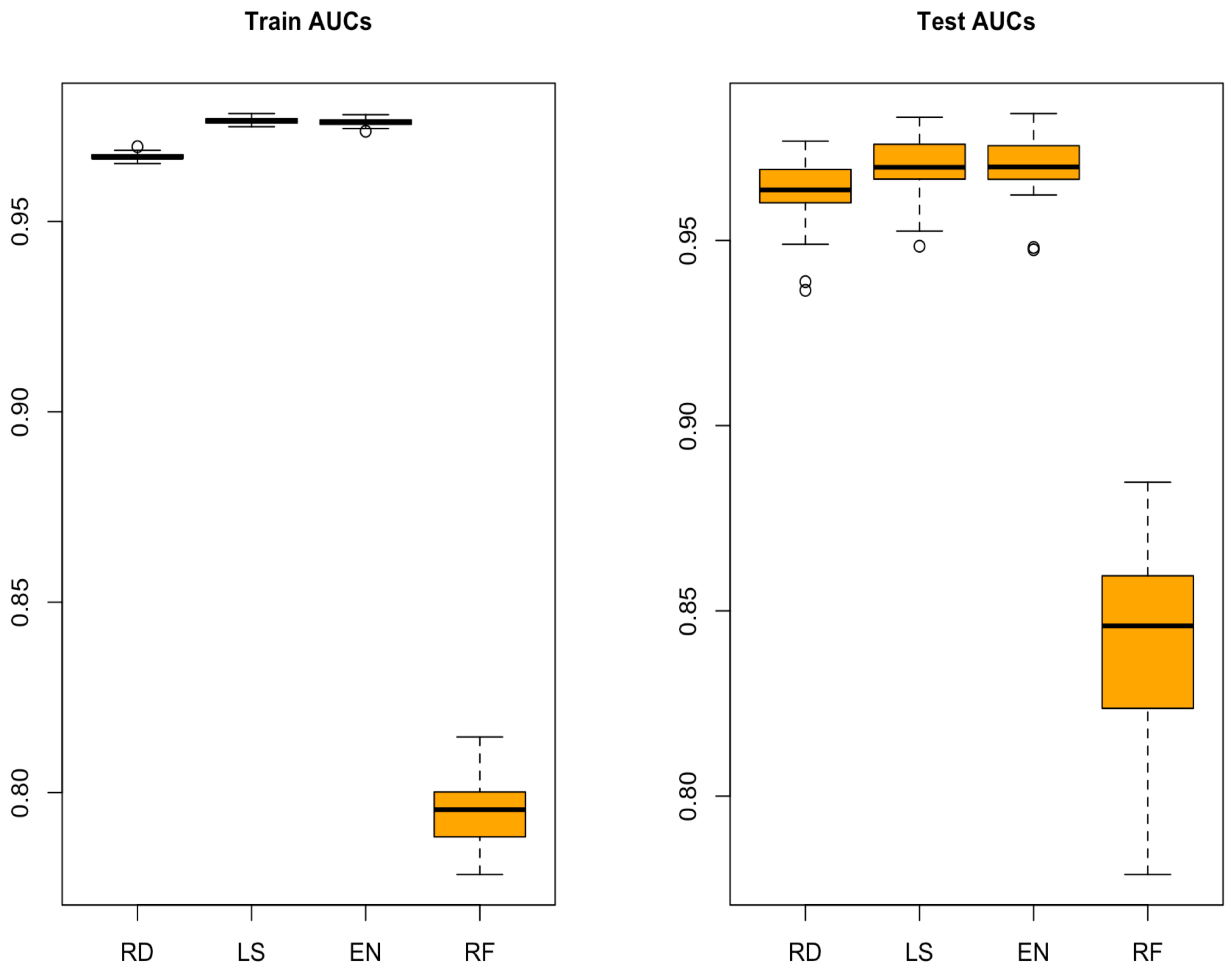
+ 1 attribute of type “capital_run_length_longest”:

length of longest uninterrupted sequences of capital letters

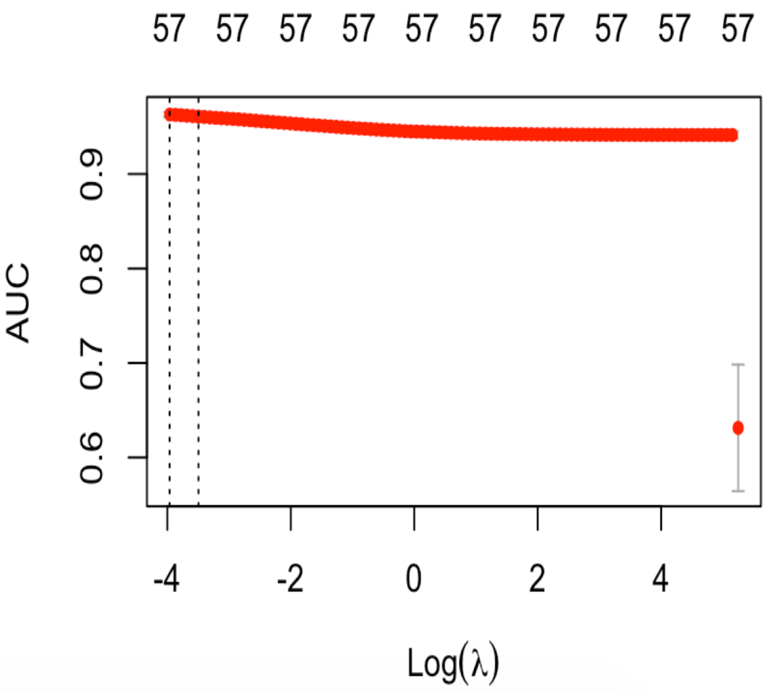
+1 attribute of type “capital_run_length_total”: total number of capital letters in the email

+ 1 attribute of type “capital_run_length_average”: average length of uninterrupted sequences of capital letters

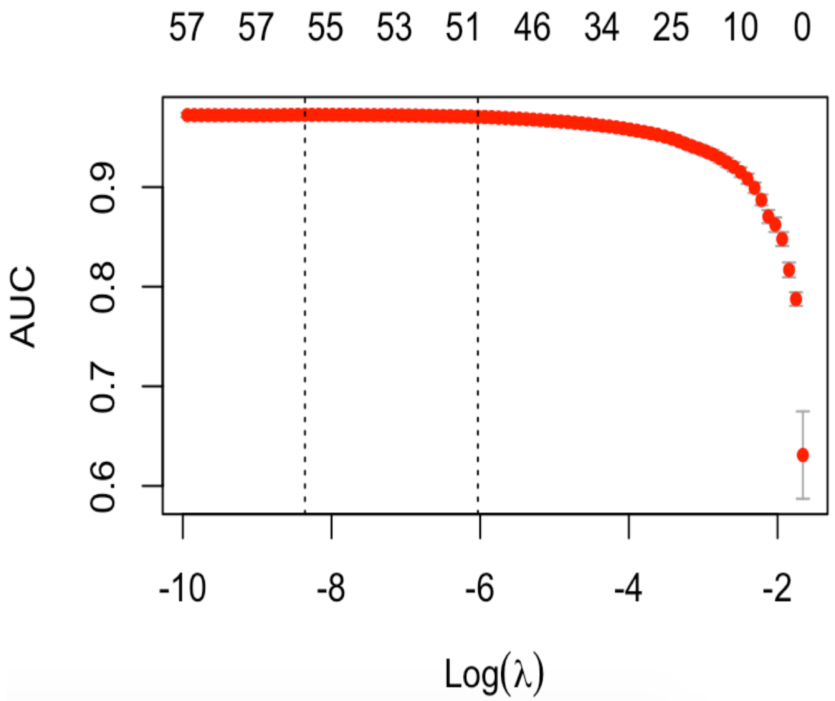
Boxplots of 50 AUCs for Train and Test set



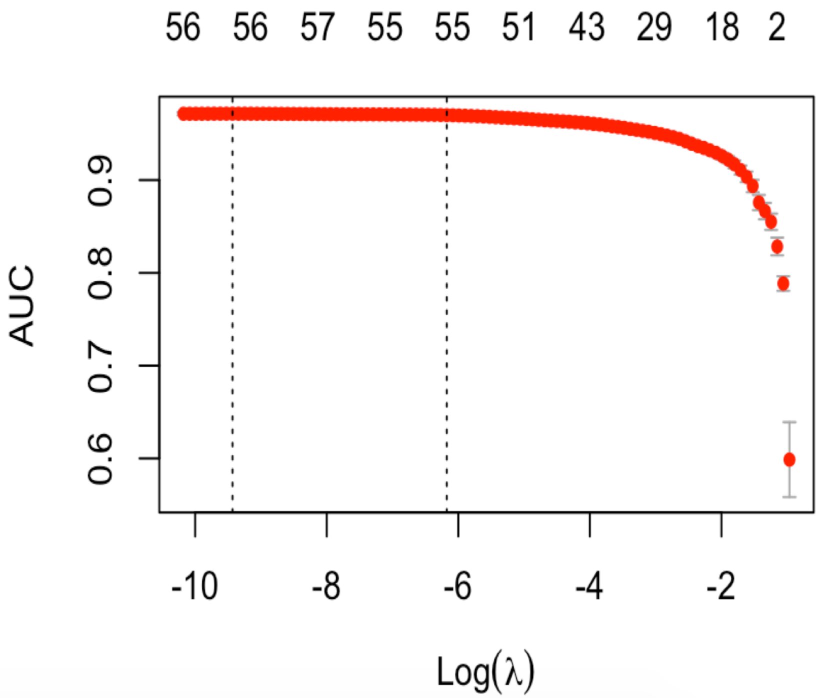
10-fold Cross Validation Curves for Regularized Methods



Ridge



Lasso



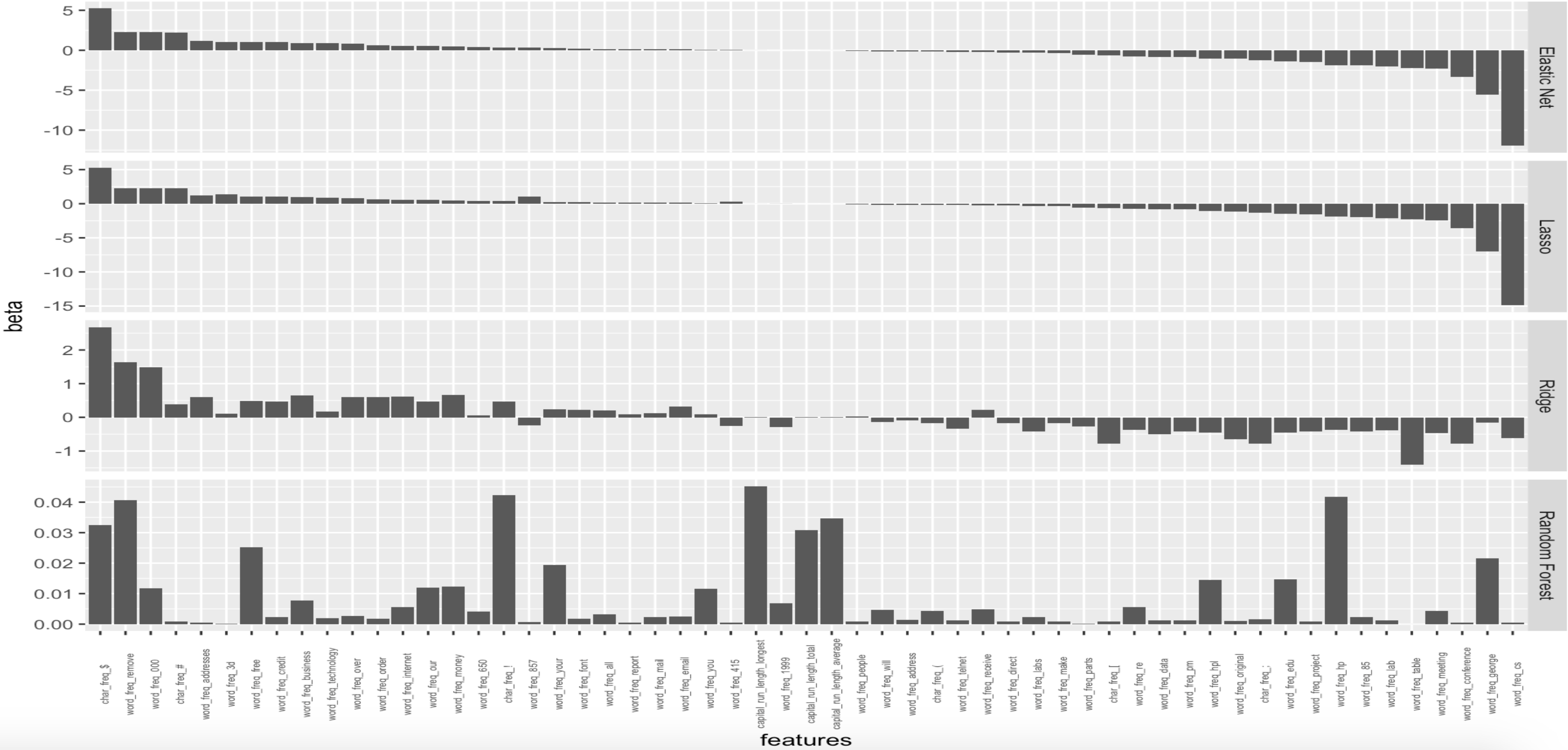
Elastic Net

Time taken to Cross Validate		
Ridge	Lasso	Elastic Net
10.99 seconds	36.14 seconds	31.18 seconds

90% AUCs and Time taken for each single fit

Methods	90% AUCs		Time taken for a single fit + CV (full data set)
Ridge	0.95	0.97	13.37 seconds
Lasso	0.96	0.98	1.13 mins
Elastic Net	0.96	0.98	1.12 mins
Random Forest	0.81	0.87	40.53 seconds

Bar plots of estimated coefficients and the importance of the parameters



Conclusion

- Random Forest doesn’t perform as well as the regularized methods in this dataset
- Ridge is the best method here based on its test AUCs and length of time for CV
- The importance of the features in Random Forest and the coefficients in the regularized methods are in agreement for a few features but also in disagreement in others. The important features are:

+ char_freq_$

+ word_freq_remove

+ word_freq_ooo