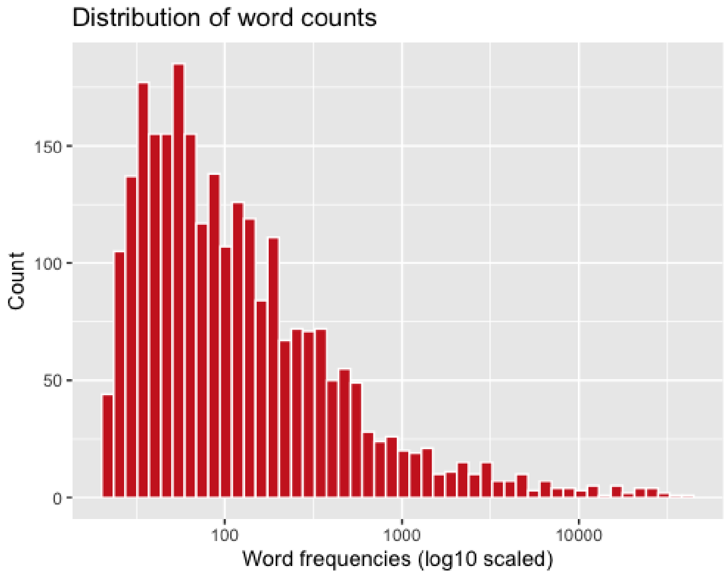
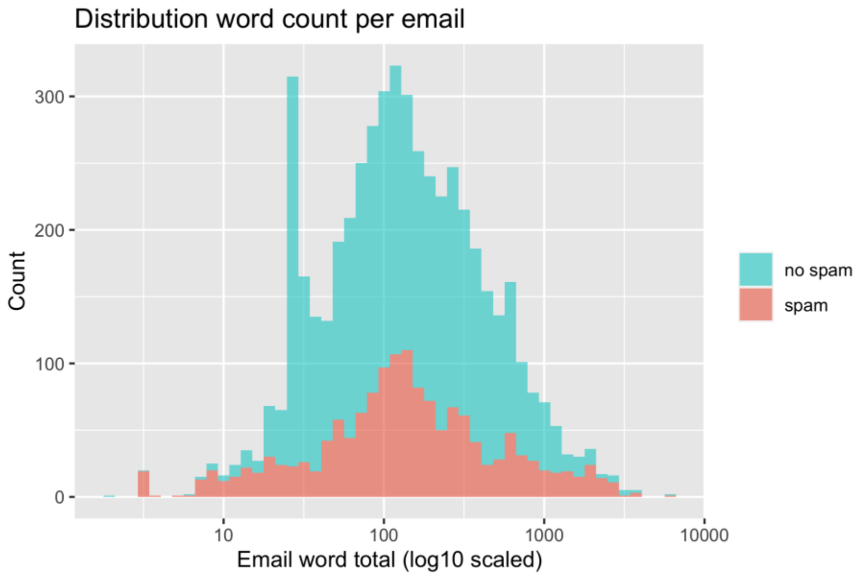
Email Spam Classification

Group 7: David Genfan, Brian Louie, Davin Yu

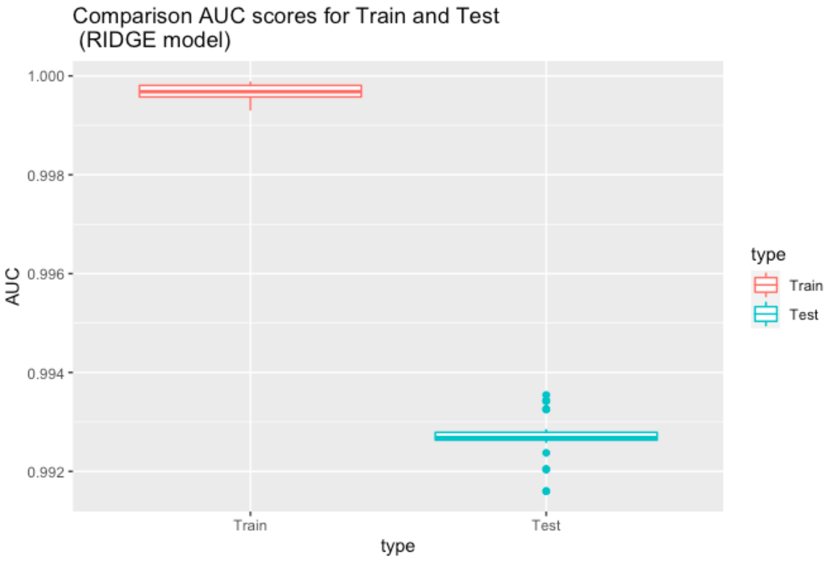
Overview

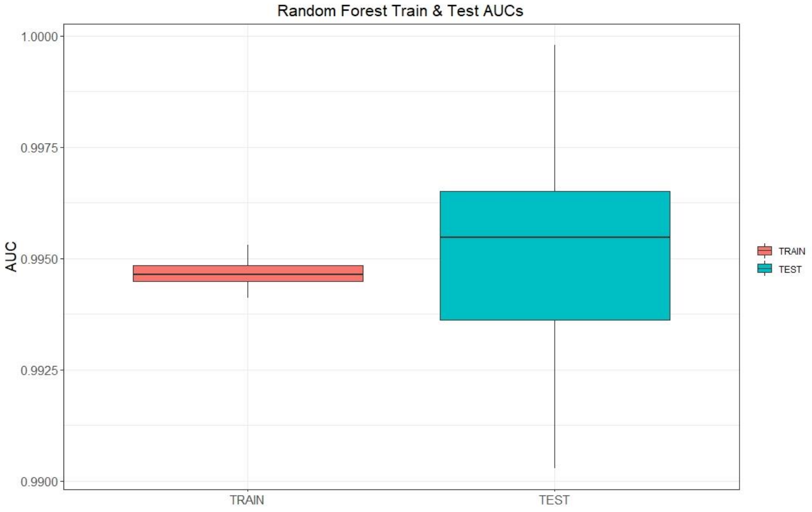
- Data contains 5172 emails with the 3000 most commonly used words from all emails and a label if the email is spam (1) or not spam (0)
  - Removed stop words from the features (“and”, “most”, “very”, etc.)
- Emails (n) = 5172, Words (P) = 2619
- Imbalance ratio: n+/n- = 1500 / 3672
- With this data set, we sought to determine if we could classify spam based on the composition of the email

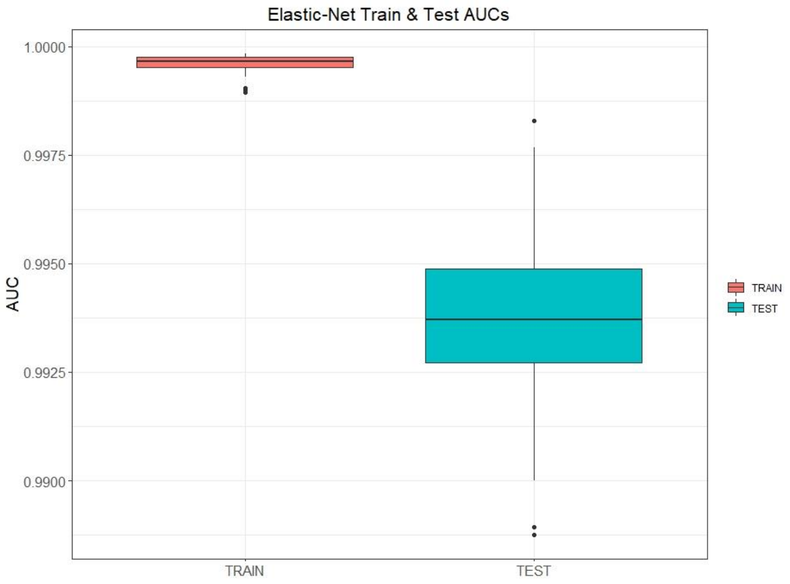


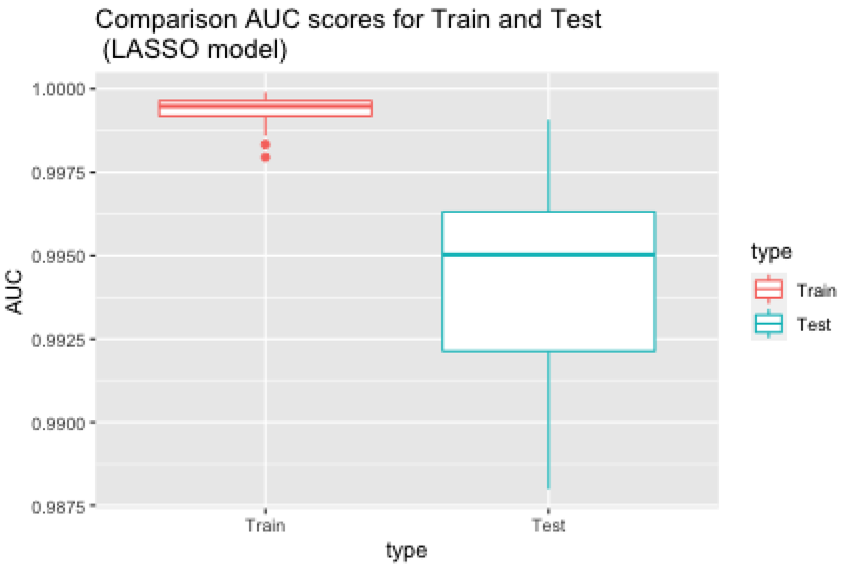


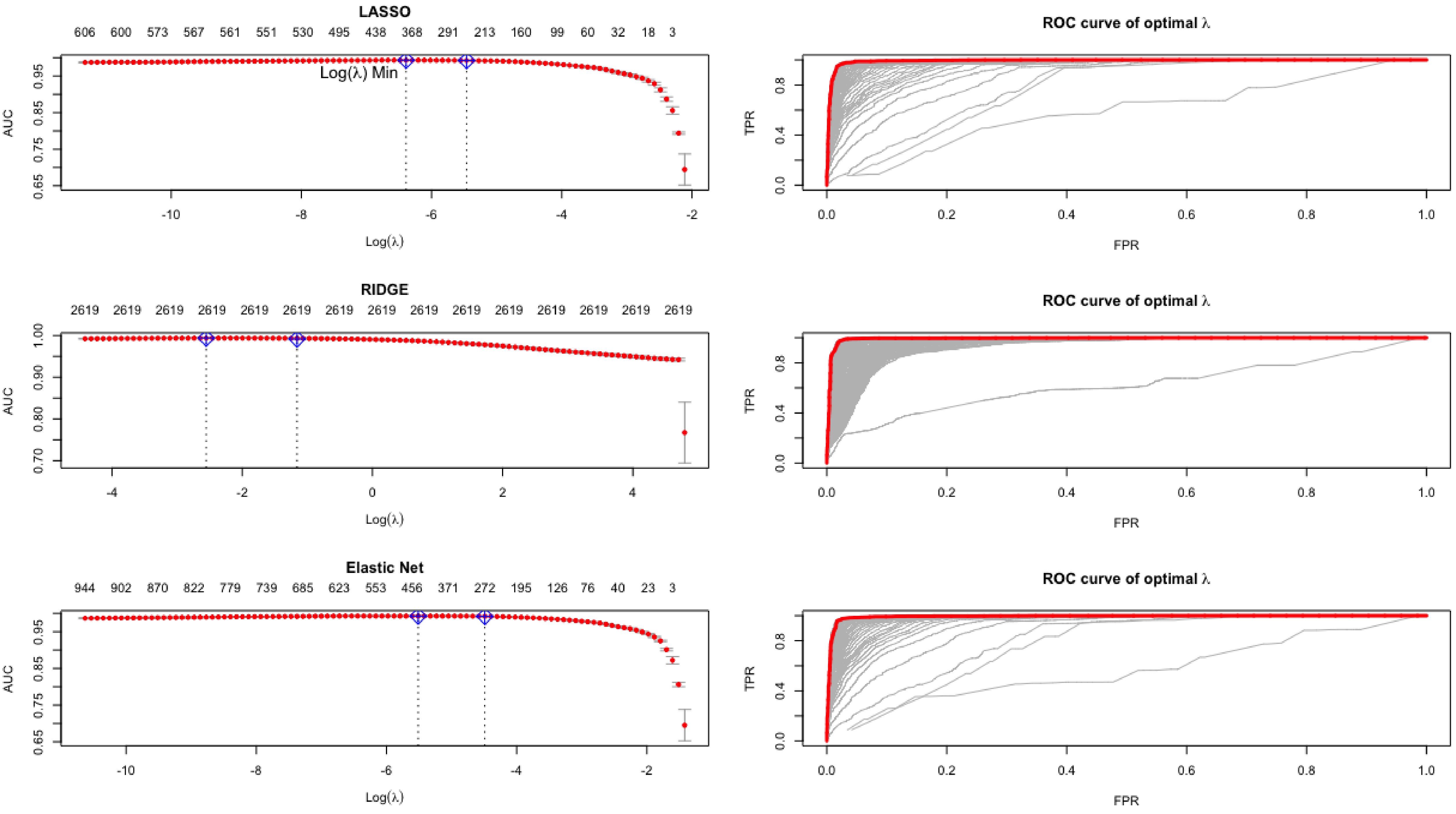
AUC's for D_{train} and $D_{validation}$









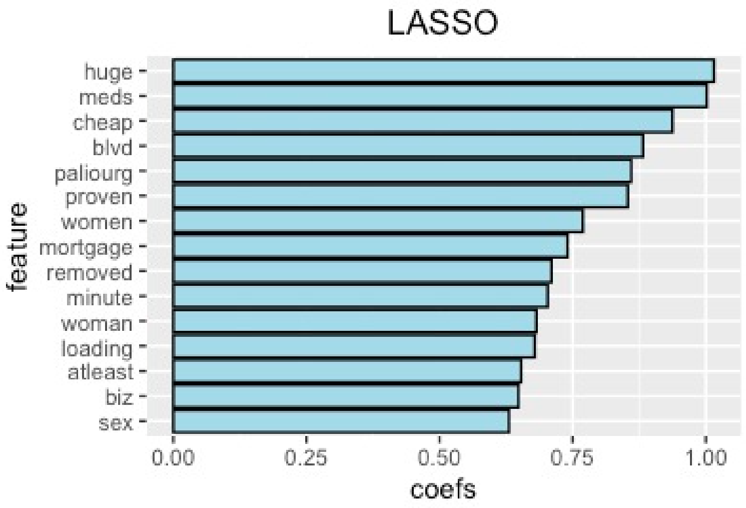


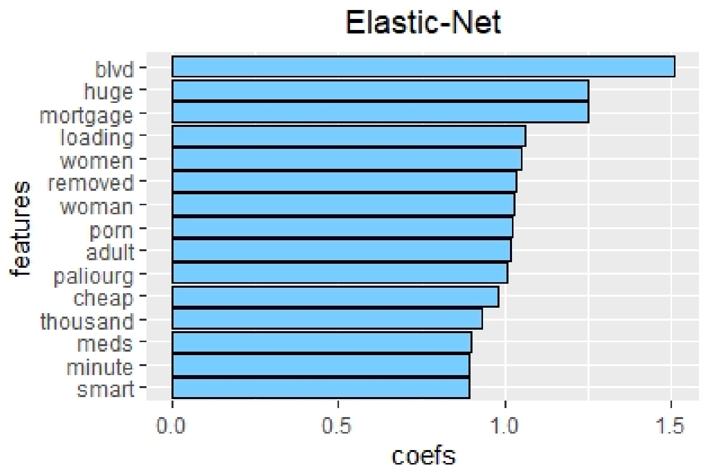
LASSO
606 600 573 567 561 551 530 495 438 368 291 213 160 99 60 32 18 3
Log(λ) Min
AUC
0.65 0.75 0.85 0.95
-10 -8 -6 -4 -2
Log(λ)
ROC curve of optimal λ
TPR
0.0 0.4 0.8
0.0 0.2 0.4 0.6 0.8 1.0
FPR
RIDGE
2619 2619 2619 2619 2619 2619 2619 2619 2619 2619 2619 2619 2619 2619 2619
AUC
0.70 0.80 0.90 1.00
-4 -2 0 2 4
Log(λ)
ROC curve of optimal λ
TPR
0.0 0.4 0.8
0.0 0.2 0.4 0.6 0.8 1.0
FPR
Elastic Net
944 902 870 822 779 739 685 623 553 456 371 272 195 126 76 40 23 3
AUC
0.65 0.75 0.85 0.95
-10 -8 -6 -4 -2
Log(λ)
ROC curve of optimal λ
TPR
0.0 0.4 0.8
0.0 0.2 0.4 0.6 0.8 1.0
FPR

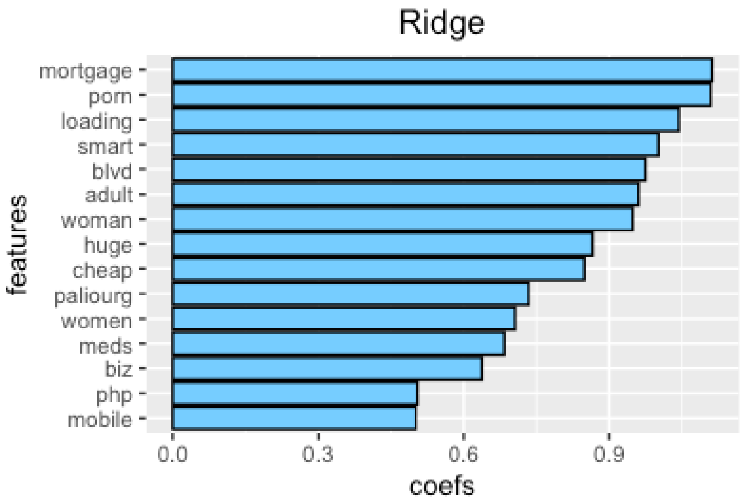
Method	Median test AUC	Time of single cross-validation	Time of full model fitting including cross-validating parameter tuning
Logistic Lasso	0.99	2 min 4 sec	4 min 45 sec
Logistic Ridge	0.99	6 min 9 sec	7 min 56 sec
Elastic Net	0.99	2 min 46 sec	3 min 4 sec
Random Forest	0.99	N/A	16 min 28 sec

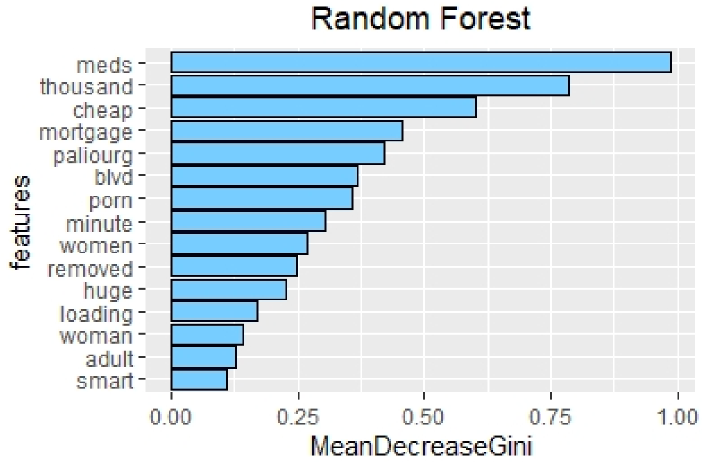
Top 15 Spam Features

Across all models, the most common features for spam emails were words such as "huge", "adult", "meds", typically associated with pornography



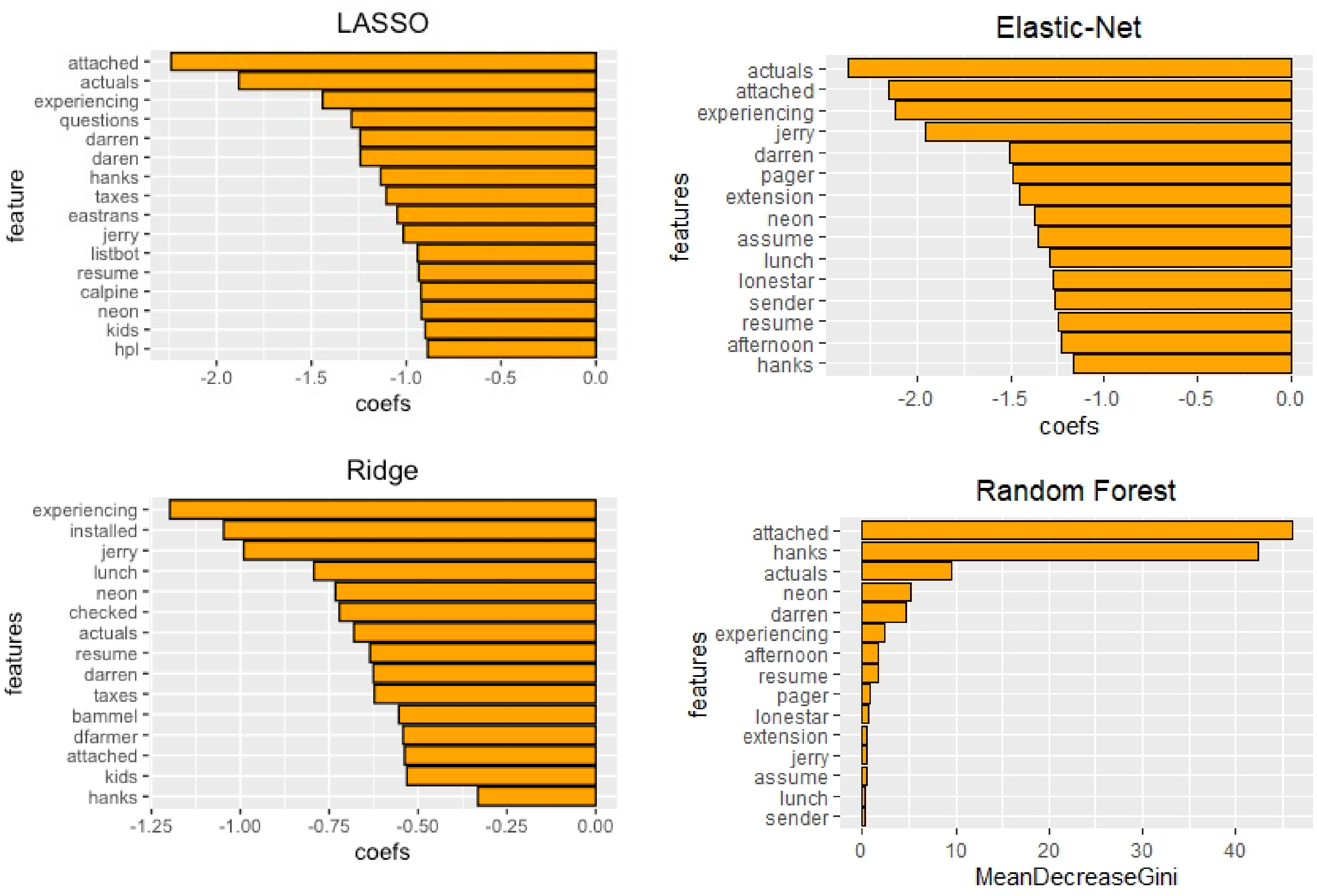




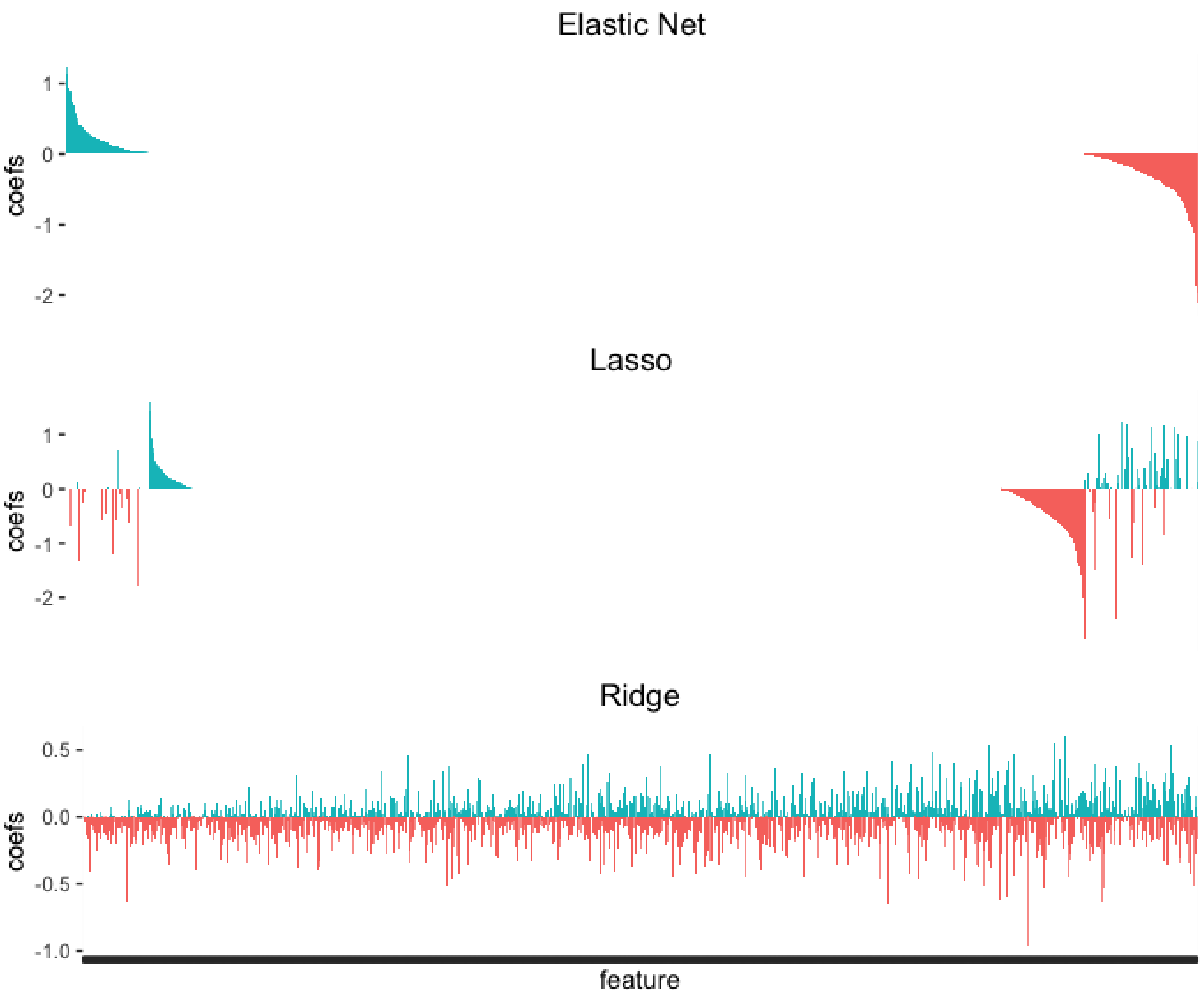


Top 15 Non-Spam Features

The top 15 non-spam features have more professional words such as "resume", "attached", "actuals".



Updated Feature Importance



Conclusion

- Performance is similar across all methods, with train AUCs being clustered closer to 1 and test AUCs hovering around 0.99
- Longer training time does not always mean more accuracy in results. Sometimes, overfitting may be occurring which would require more analysis.
- The top 15 Spam features were consistent across models, with words such as "adult", "mortgage", and "woman" included as spam triggers.