Lab 1

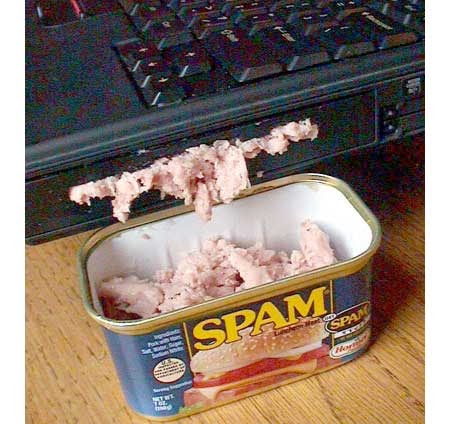
David Duffrin

Table of Contents

# Lab 1

## Intro

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

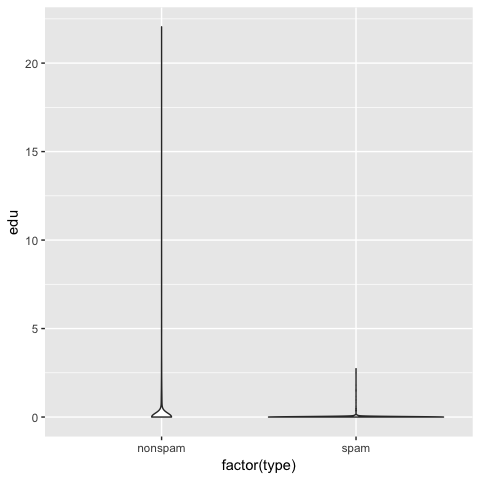
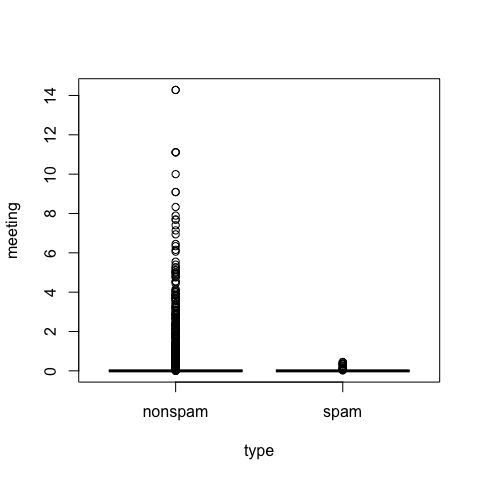


spammy lappy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| make | address | all | num3d | our | over | remove | internet | order | mail | receive | will | people | report | addresses | free | business | email | you | credit | your | font | num000 | money | hp | hpl | george | num650 | lab | labs | telnet | num857 | data | num415 | num85 | technology | num1999 | parts | pm | direct | cs | meeting | original | project | re | edu | table | conference | charSemicolon | charRoundbracket | charSquarebracket | charExclamation | charDollar | charHash | capitalAve | capitalLong | capitalTotal | type |
| 0.00 | 0.64 | 0.64 | 0 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.64 | 0.00 | 0.00 | 0.00 | 0.32 | 0.00 | 1.29 | 1.93 | 0.00 | 0.96 | 0 | 0.00 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.000 | 0 | 0.778 | 0.000 | 0.000 | 3.756 | 61 | 278 | spam |
| 0.21 | 0.28 | 0.50 | 0 | 0.14 | 0.28 | 0.21 | 0.07 | 0.00 | 0.94 | 0.21 | 0.79 | 0.65 | 0.21 | 0.14 | 0.14 | 0.07 | 0.28 | 3.47 | 0.00 | 1.59 | 0 | 0.43 | 0.43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.07 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.132 | 0 | 0.372 | 0.180 | 0.048 | 5.114 | 101 | 1028 | spam |
| 0.06 | 0.00 | 0.71 | 0 | 1.23 | 0.19 | 0.19 | 0.12 | 0.64 | 0.25 | 0.38 | 0.45 | 0.12 | 0.00 | 1.75 | 0.06 | 0.06 | 1.03 | 1.36 | 0.32 | 0.51 | 0 | 1.16 | 0.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0.06 | 0 | 0 | 0.12 | 0 | 0.06 | 0.06 | 0 | 0 | 0.01 | 0.143 | 0 | 0.276 | 0.184 | 0.010 | 9.821 | 485 | 2259 | spam |
| 0.00 | 0.00 | 0.00 | 0 | 0.63 | 0.00 | 0.31 | 0.63 | 0.31 | 0.63 | 0.31 | 0.31 | 0.31 | 0.00 | 0.00 | 0.31 | 0.00 | 0.00 | 3.18 | 0.00 | 0.31 | 0 | 0.00 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.137 | 0 | 0.137 | 0.000 | 0.000 | 3.537 | 40 | 191 | spam |
| 0.00 | 0.00 | 0.00 | 0 | 0.63 | 0.00 | 0.31 | 0.63 | 0.31 | 0.63 | 0.31 | 0.31 | 0.31 | 0.00 | 0.00 | 0.31 | 0.00 | 0.00 | 3.18 | 0.00 | 0.31 | 0 | 0.00 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.135 | 0 | 0.135 | 0.000 | 0.000 | 3.537 | 40 | 191 | spam |
| 0.00 | 0.00 | 0.00 | 0 | 1.85 | 0.00 | 0.00 | 1.85 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0 | 0.00 | 0 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.223 | 0 | 0.000 | 0.000 | 0.000 | 3.000 | 15 | 54 | spam |

## money edu type   
## Min. : 0.00000 Min. : 0.0000 nonspam:2788   
## 1st Qu.: 0.00000 1st Qu.: 0.0000 spam :1813   
## Median : 0.00000 Median : 0.0000   
## Mean : 0.09427 Mean : 0.1798   
## 3rd Qu.: 0.00000 3rd Qu.: 0.0000   
## Max. :12.50000 Max. :22.0500

The percentage of emails that are spam in the dataset is 0.3940448



## Data Analysis

## Generalized Linear Model   
##   
## 2761 samples  
## 57 predictor  
## 2 classes: 'nonspam', 'spam'   
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 2761, 2761, 2761, 2761, 2761, 2761, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9221363 0.8368392

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction nonspam spam  
## nonspam 1044 78  
## spam 71 647  
##   
## Accuracy : 0.919   
## 95% CI : (0.9056, 0.9311)  
## No Information Rate : 0.606   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8301   
## Mcnemar's Test P-Value : 0.623   
##   
## Sensitivity : 0.9363   
## Specificity : 0.8924   
## Pos Pred Value : 0.9305   
## Neg Pred Value : 0.9011   
## Prevalence : 0.6060   
## Detection Rate : 0.5674   
## Detection Prevalence : 0.6098   
## Balanced Accuracy : 0.9144   
##   
## 'Positive' Class : nonspam   
##

##   
## Call:  
## roc.default(response = testing$typeNum, predictor = resultsNum)  
##   
## Data: resultsNum in 1115 controls (testing$typeNum 0) < 725 cases (testing$typeNum 1).  
## Area under the curve: 0.9144

My test dataset contains 60.5939877 percent nonspam emails, so I will use this as a baseline for accuracy. After training a binomial model on 60.0086938 percent of the data, I predicted the type of email in the test dataset (which consisted of the remaining rows). I got an accuracy of 0.9221363 and AUC of 0.9143683.