Lab 1

David Duffrin

Contents

Figure 1: spammy lappy

Lab 1

Intro

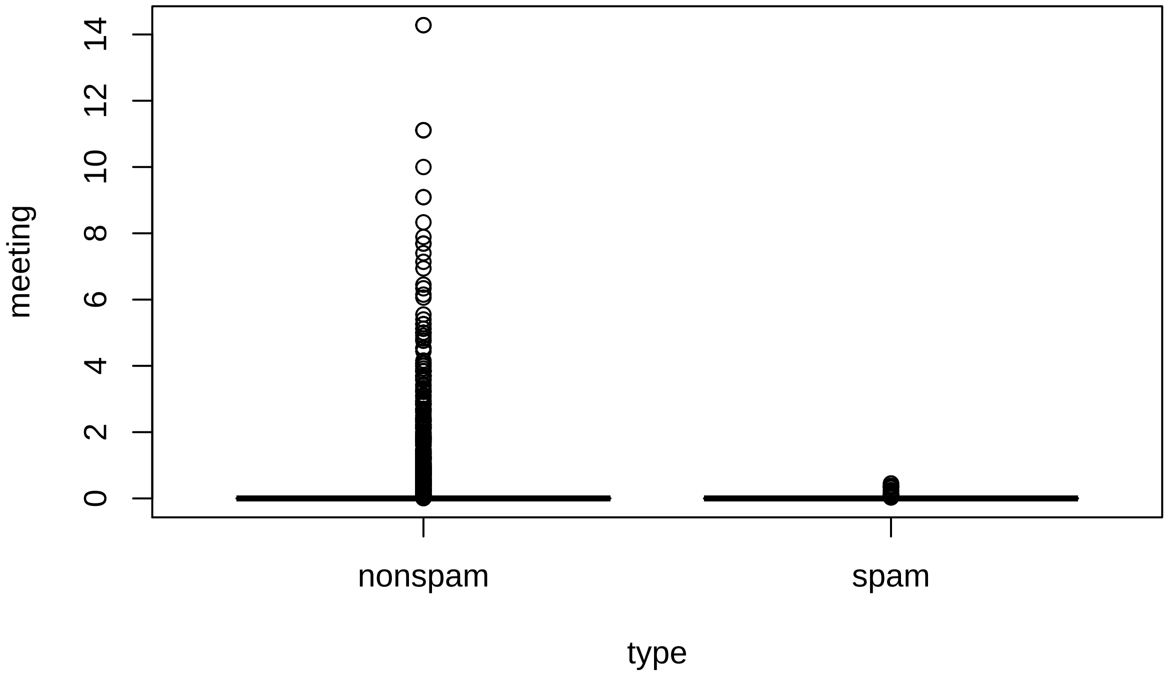
A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

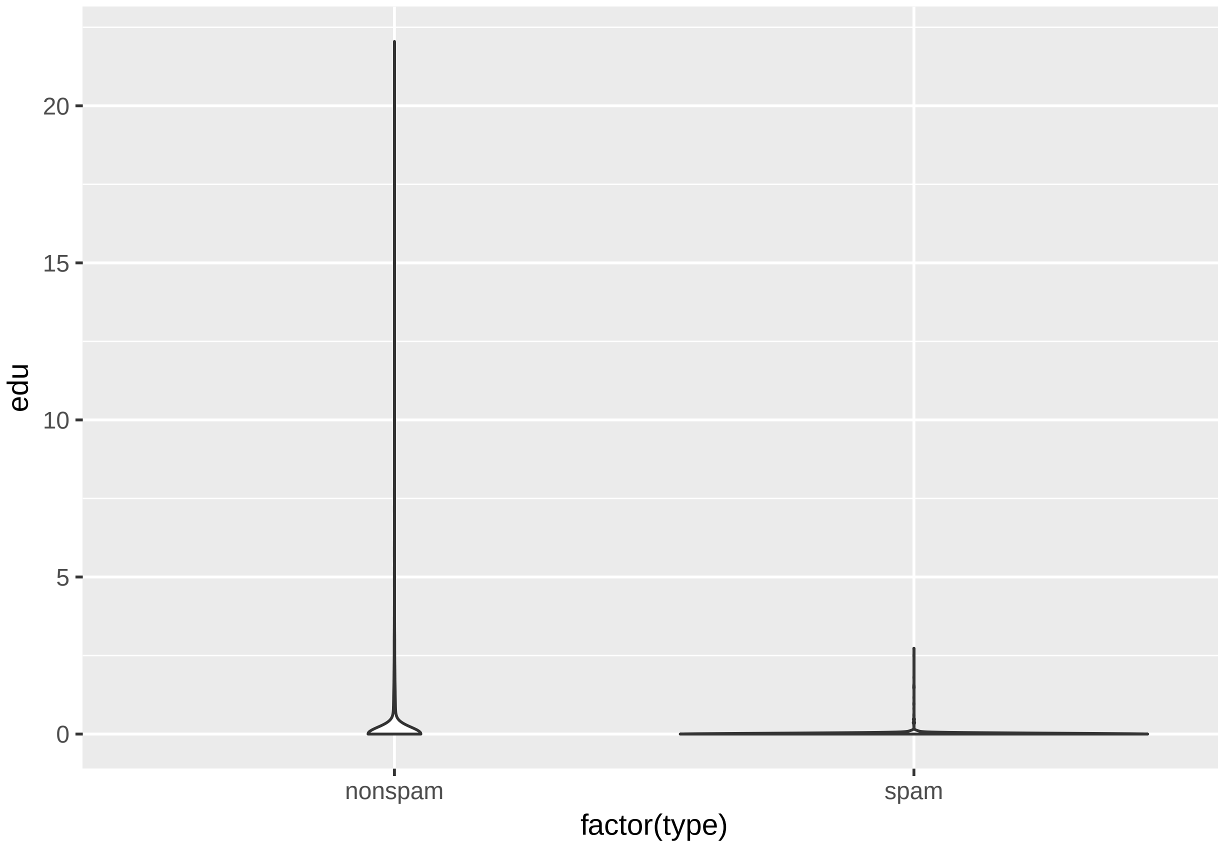
make	address	all	num3d	our	over	remove	internet	order	mail	receive	will	people	report	addres
0.00	0.64	0.64	0	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0
0.21	0.28	0.50	0	0.14	0.28	0.21	0.07	0.00	0.94	0.21	0.79	0.65	0.21	0
0.06	0.00	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0.00	1
0.00	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
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

make	address	all	num3d	our	over	remove	internet	order	mail	receive	will	people	report	addres
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

```
##      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.