ANALYSIS ON EMAIL SPAM FILTERING

Data Analytics

PROJECT REPORT

Submitted by

R.P.L.LasyaPriya-17MIS7065

Suda Naga Sai Divya-17MIS7037

Under the Guidance of

Dr. Awnish Kumar

Associate Professor, CSE,

VIT-AP

TABLE OF CONTENTS

Chapter -1

Introduction

WHAT IS AN EMAIL?

Email, short for "electronic mail," is one of the most widely used features of the Internet, along with the web. It allows you to send and receive messages to and from anyone with an email address, anywhere in the world.

WHAT IS EMAIL SPAM?

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers. Often, spam email is sent for commercial purposes. While some people view it as unethical, many businesses still use spam. The cost per email is incredibly low, and businesses can send out mass quantities consistently. Spam email can also be a malicious attempt to gain access to your computer.

Most email providers offer a spam filter that automatically flags likely spam messages and separates them from the ham. There are many techniques for this but most commonly used technique is analysis on the content of the email. They use algorithms to predict whether an email is spam or ham with the help of previous data of emails. There are many algorithms that can be used to predict. Some algorithms work perfectly in filtering the spam but some may fail to filter the actual spam and some may do wrong filtering of spam too.

So, how to decide which algorithm will work perfectly and what if a wrong algorithm is considered? Our analysis helps to answer this question. It tells which algorithm gives perfect results and which doesn’t.

Chapter -2

Background Study

Email has become an indispensable part of daily business activities in nearly all aspects of commerce. The massive breadth of populations using email regularly can be attributed, in large part, to its accessibility and general usefulness. The first example of email was found in MIT. Users of MIT computers could leave messages with this program on computers at the university for other users, who would see the messages the next time they logged on to the computer.

As the internet started to gain popularity in the early 1990s, it was quickly recognized as an excellent advertising tool. At practically no cost, a person can use the internet to send an email message to thousands of people. These unsolicited junk electronic mails came to be called 'Spam'. The first email spam was sent to hundred users at a time. It was an advertisement for a presentation by Digital Equipment Corporation for their DECSYSTEM-20 products sent by Gary Thuerk, a marketer of theirs. The reaction to this first spam email was negative and then there are no others instances for long time after this.

Later a system called “Spam Filtering System” came into existence, which will filter the normal emails from the spam. So, every email service provider started using this system. But there are some disadvantages or effects of spam too. Some of them are follows:

- Communications overload
- Waste of time
- Irritation and discontent
- Criminalization of spam
- The loss of an important email that accidentally gets deleted along with the plethora of spam.

Our analysis mainly concentrates on the last effect of spam from the list mentioned above. Most of the people never check their spam mails. If some important mail accidentally get into spam, that will be definitely a loss. So, We decided to give a solution for this with our analysis.

Chapter -3

Problem Definition

We receive a number of emails in this internet world. Some of the emails we receive regularly are large nuisance creators. With spam filtering we can control it. But at the same time some of the useful emails may also get into spam. With this we may miss an important email.

This is actually due to the Machine Learning Algorithm that the email providers are using. An algorithm with less accuracy may send an important email into spam. So using an algorithm which is having a high accuracy is very important.

So, the main aim of the analysis is to know the spam filtering process and decide which algorithm helps in accurate spam filtering.

Chapter -4

Objective

Email Spam is an electronic spam where unsolicited messages are sent by email. There are many techniques for email spam filtering but most commonly used technique is analysis on the content of the email. This analysis is done using machine learning algorithms.

Our Analysis gives a clear idea about the filtering process of email spam and decides which Machine Learning Algorithm will be the best choice for email spam filtering.

Chapter -5

Methodology/Procedure

The procedure we followed in our analysis is as follows:

1. Load the data set
2. Pre-process the data
3. Extract the word frequency from the data
4. Split the dataset into train and test data
5. Apply Logistic Regression model
6. Apply CART model
7. Apply Random Forest model
8. Do the predictions on training data
9. Do the prediction on test data
10. Compare the predictions of three respective models in both train and test data.
11. Analyze the predictions and give a result from the analysis.

Chapter -6

Results and Discussion

Logistic Regression: Used when the dependent variable is categorical. In our analysis dependent variable is categorical(Spam or Not Spam).

CART Model: Classification And Regression Tree can be used for classification or regression predictive modeling problems. It is used in our analysis, as we need to classify the mail as Spam or Not Spam.

Random Forest: Used in all purposes like classification, regression and other tasks as well. In most of the cases it overcomes the drawback of CART i.e, over fitting the data.

In our analysis, we used logistic regression, CART and random forest models for training the data. And evaluated the predictions of these three algorithms on test data. The predictions of three algorithms on both training and testing data of our analysis are as follows:

- Logistic Regression – 99.99%(training); 96.27 %(testing)
- CART Model – 96.96%(training); 96.31%(testing)
- Random Forest – 99.78%(training); 99.75%(testing).

From the above mentioned predictions, we analyzed the following:

- Logistic Regression got most accurate percentage in case of training data. But it was far-a-part in case of testing data. This is called over fitting because of large data.
- CART Model is at an average accuracy both in case of training and testing data.
- Random Forest got its impressive accuracy percentage both in case of training and testing data.

Chapter -7

Conclusion and Future Scope

From our analysis we like to conclude that using an imperfect algorithm may send an important mail also into spam. Most of the people never check their spam box, so they will never see that mail. To avoid this we need to use a high accuracy based algorithm.

Random Forest is the most suitable algorithm for Spam filtering. As it worked well with a large dataset on both training and test data, we can apply random forest in spam filtering process to have a perfect Spam Filtering System.

And the future scope of our analysis is to apply our analysis result to a practically working model so it would be useful to filter the spam perfectly in future.

References

1. https://kharshit.github.io/blog/2017/08/25/email-spam-filtering-text-analysis-in-r

2. https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,average%20prediction%20(regression)%20of%20the

3. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

4. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Appendix – A

Team Work and Work Management

WHAT IS TEAM WORK?

The importance of teamwork in the workplace is demonstrated by the feelings of unity, collaboration, and motivation that it brings. An effective team works as a collective unit by utilizing the skills and talents of each team member to support and achieve a common goal.

The reasons why the team work is important and why it matters are as follows:

- Motivates unity in work place
- Offers differing perspectives and feedback
- Provides improved efficiency and productivity
- Provides great learning opportunities
- Promotes work place synergy

WHAT IS WORK MANAGEMENT?

In simple terms work management is a combination of Resource Management, Time Management, Project Management, Client Relationship Management, Process Management and Business Intelligence.

It helps the organization in crucial business processes in terms of the following:

- Schedule work more effectively
- Meet client needs
- Utilize assets and resources
- Evaluate performance

Appendix – B

Coding and Snap Shot

```
# Give the path our the dataset
setwd("C:/Users/user/Desktop/SEM7/Data Analytics/Project")

# Load the dataset
emails = read.csv('emails.csv', stringsAsFactors = FALSE)
str(emails)

# Create confusion matrix of the dependent variable
table(emails$spam)
```

```
> setwd("C:/Users/user/Desktop/SEM7/Data Analytics/Project")
> emails = read.csv('emails.csv', stringsAsFactors = FALSE)
> str(emails)
'data.frame':   5728 obs. of  2 variables:
 $ text: chr  "Subject: naturally irresistible your corporate identity  lt is really hard
 to recollect a company : the  market"| __truncated__ "Subject: the stock trading gunslin
ger  fanny is merrill but muzo not colza attainder and penultimate like esmar"| __truncat
ed__ "Subject: unbelievable new homes made easy  im wanting to show you this  homeowner
 you have been pre - approved"| __truncated__ "Subject: 4 color printing special  request
 additional information now ! click here  click here for a printable "| __truncated__ ...
 $ spam: int  1 1 1 1 1 1 1 1 1 1 ...
> table(emails$spam)

   0    1
4360 1368
>
```

```
# Load the text mining package and start preprocessing
library(tm)
```

Extract the word frequency from the dataset

corpus = VCorpus(VectorSource(emails$text))

#Convert the words into lower case

corpus = tm_map(corpus, content_transformer(tolower))

#Convert the entire text into a plain text

corpus = tm_map(corpus, PlainTextDocument)

#Remove all the punctuation marks

corpus = tm_map(corpus, removePunctuation)

corpus = tm_map(corpus, removeWords, stopwords("en"))

corpus = tm_map(corpus, stemDocument)

```
> library(tm)
Loading required package: NLP
> corpus = VCorpus(VectorSource(emails$text))
> corpus = tm_map(corpus, content_transformer(tolower))
> corpus = tm_map(corpus, PlainTextDocument)
> corpus = tm_map(corpus, removePunctuation)
> corpus = tm_map(corpus, removeWords, stopwords("en"))
> corpus = tm_map(corpus, stemDocument)
>
```

Convert the text into document term matrix

dtm = DocumentTermMatrix(corpus)

dtm

spdtm = removeSparseTerms(dtm, 0.95)

spdtm

```
> dtm = DocumentTermMatrix(corpus)
> dtm
<<DocumentTermMatrix (documents: 5728, terms: 28687)>>
Non-/sparse entries: 481719/163837417
Sparsity           : 100%
Maximal term length: 24
Weighting          : term frequency (tf)
> spdtm = removeSparseTerms(dtm, 0.95)
> spdtm
<<DocumentTermMatrix (documents: 5728, terms: 330)>>
Non-/sparse entries: 213551/1676689
Sparsity           : 89%
Maximal term length: 10
Weighting          : term frequency (tf)
>
```

Convert the text document into a dataframe

emailsSparse = as.data.frame(as.matrix(spdtm))

colnames(emailsSparse) = make.names(colnames(emailsSparse))

#Sort the words in a sequence

sort(colSums(emailsSparse))

```
> emailsSparse = as.data.frame(as.matrix(spdtm))
> colnames(emailsSparse) = make.names(colnames(emailsSparse))
> sort(colSums(emailsSparse))
    vkamin      begin     either       done      sorri        lot    mention
       301        317        318        337        343        348        355
   thought      bring       idea     better     immedi    without       mean
       367        374        378        383        385        389        390
     write      happi      repli       life     experi     involv     specif
       390        396        397        400        405        405        407
    arrang      creat       read       wish       open     realli       link
       410        413        413        414        416        417        421
       say    respond      sever       keep        etc      anoth        run
       423        430        430        431        434        435        437
      info     togeth      short     sincer        buy        due    alreadi
       438        438        439        441        442        445        446
      line      allow     recent    special      given     believ     design
       448        450        451        451        453        456        457
       put      remov       X853  wednesday       type     public       full
       458        460        462        464        466        468        469
      hear       join     effect     effort    tuesday     robert      locat
       469        469        471        473        474        482        485
     check       area      final    increas       soon    analysi       sure
       488        489        490        491        492        495        495
      deal     return      place      onlin    success       sinc understand
       498        509        516        518        519        521        521
     still     import    comment    confirm      hello       long      thing
       523        530        531        532        534        534        535
     point    appreci       feel      howev     member       hour        net
       536        541        543        545        545        548        548
```

```
      talk       last      phone       X000      chang        fax       john
       981        998       1001       1007       1035       1038       1042
   current    stinson       give    univers      offic        gas    schedul
      1044       1051       1055       1059       1068       1070       1071
    financ      state       name       X713       good      posit   crenshaw
      1073       1086       1089       1097       1097       1104       1115
    system       well       sent      visit       free      next.      avail
      1118       1125       1126       1126       1141       1145       1152
  question    address      offer     attach     number       date    product
      1152       1154       1171       1176       1182       1187       1197
     order      think     includ     report       best     confer        now
      1210       1216       1238       1279       1291       1297       1300
       www    discuss  interview     servic   communic    request       just
      1323       1326       1333       1337       1343       1344       1354
      take      trade       send     provid       list       help    program
      1361       1366       1379       1405       1410       1430       1438
    option       want    project    contact    present     follow     receiv
      1488       1488       1522       1543       1543       1552       1557
       see    houston       http        edu       call    shirley       corp
      1567       1582       1609       1627       1687       1689       1692
      week   interest        day       also    develop       make       year
      1758       1814       1860       1864       1882       1884       1890
       let     messag       look     regard      email        one      power
      1963       1983       2003       2045       2066       2108       2117
    energi      model       risk       mail        new    compani       busi
      2179       2199       2267       2269       2281       2290       2313
      need        use       like        get        may      manag      group
      2328       2330       2352       2462       2465       2600       2604
      know       meet      price     inform       work     market   research
      2614       2623       2694       2701       2708       2750       2820
     X2001       time    forward      thank        can   kaminski      X2000
      3089       3145       3161       3730       4257       4801       4967
     pleas        com        hou       will       vinc    subject        ect
      5113       5443       5577       8252       8532      10202      11427
     enron
     13388
> |
```

emailsSparse$spam = emails$spam

Sort the words that are frequent in ham

sort(colSums(subset(emailsSparse, spam == 0)))

```
> emailsSparse$spam = emails$spam
> sort(colSums(subset(emailsSparse, spam == 0)))
      spam       life      remov      money      onlin    without     websit
         0         80        103        114        173        191        194
     click    special       wish      repli        buy        net       link
       217        226        229        239        243        243        247
    immedi       done       mean     design        lot     effect       info
       249        254        259        261        268        270        273
    either       read      write       line      begin      sorri    success
       279        279        286        289        291        293        293
    involv      creat    softwar     better     vkamin        say       keep
       294        299        299        301        301        305        306
     bring     believ       full    increas     realli    mention    thought
       311        313        317        320        324        325        325
      idea     invest       secur     specif      sever     experi      thing
       327        327        337        338        340        346        347
     allow      check        due       type      happi     return     expect
       348        351        351        352        354        355        356
     short     effort       open   internet     sincer     public     recent
       357        358        360        361        361        364        368
     anoth    alreadi       home       made    respond      given        etc
       369        372        375        380        382        383        385
       put     within      place      right    version      hello       sure
       385        386        388        390        390        395        396
      area        run     arrang    account       join       hour      locat
       397        398        399        401        403        404        406
    togeth      engin     import        per     corpor       high     result
       406        411        411        412        414        416        418
      hear      final       deal     applic       even        web     custom
       420        422        423        428        429        430        433
      soon       long       sinc      futur     member       x000      event
       435        436        439        440        446        447        447
       don       part       feel    tuesday  wednesday      still       unit
       450        450        453        454        456        457        457
      site       x853    continu understand    resourc     robert    analysi
       458        461        464        464        466        466        468
      form      point     assist    confirm     differ     intern     might
       468        474        475        485        489        489        490
     state     system    process       hope     london       just     receiv
       806        816        826        828        828        830        830
     chang     review    current      shall     friday       team      phone
       831        834        841        844        847        850        858
      issu       data      avail       last       good       give        www
       865        868        872        874        876        883        897
       gas       list      posit      visit     includ      resum       best
       905        907        917        920        924        928        933
     offic     servic       talk     number       well        fax     provid
       935        942        943        951        961        963        970
      sent      next.       send       http       john    univers     financ
       971        975        986       1009       1022       1025       1038
   stinson    schedul       take       date       want   question    program
      1051       1054       1057       1060       1068       1069       1080
     think       x713   crenshaw     attach      trade       help      email
      1084       1097       1115       1155       1167       1168       1201
   compani    request        see   communic     confer    discuss       make
      1225       1227       1238       1251       1264       1270       1281
   contact     follow  interview    project       mail    present       busi
      1301       1308       1320       1328       1352       1397       1416
  interest     option        day       call        one       year       week
      1429       1432       1440       1497       1516       1523       1527
     messag    houston       also       look        edu       corp    shirley
      1538       1577       1604       1607       1620       1643       1687
   develop        get        new        use        let    regard     inform
      1691       1768       1777       1784       1856       1859       1883
      need      power        may       like       risk    energi     market
      1890       1972       1976       1980       2097       2124       2150
     model      price       work      manag       know      group       meet
      2170       2191       2293       2334       2345       2474       2544
      time   research    forward      x2001        can      thank        com
      2552       2752       2952       3060       3426       3558       4444
     pleas   kaminski       x2000        hou       will       vinc    subject
      4494       4801       4935       5569       6802       8531       8625
       ect      enron
     11417      13388
>
```

#Sort the words that are frequent in spam

sort(colSums(subset(emailsSparse, spam == 1)))

```
> sort(colSums(subset(emailsSparse, spam == 1)))
      x713   crenshaw      enron     gibner   kaminski     stinson      vkamin
         0          0          0          0          0          0          0
      x853       vinc        doc      kevin    shirley      deriv      april
         1          1          2          2          2          3          5
   houston      resum        edu     friday       hou  wednesday        ect
         5          5          7          7          8          8         10
    arrang  interview     attend     london     robert    student    schedul
        11         13         15         15         16         16         17
  thursday     monday       john    tuesday     attach    suggest    appreci
        17         19         20         20         21         21         23
      mark      begin    comment    analysi      x2001      model       hope
        25         26         26         27         29         29         30
   mention      x2000     togeth     confer      invit    univers     financ
        30         32         32         33         33         34         35
      talk     either        run       morn      shall      happi    thought
        38         39         39         40         40         42         42
    depart    confirm    respond     school       corp        etc       hear
        46         47         48         48         49         49         49
     howev      sorri       idea     energi    discuss       open     option
        49         50         51         55         56         56         56
     soon understand      cours     experi     associ      point      bring
        57         57         59         59         62         62         63
  director   particip     anoth       join      still      final   research
        65         65         66         66         66         68         68
     case        set     specif      given       juli    problem        put
        69         69         69         70         71         73         73
   alreadi        ask        abl       deal        fax       book       team
        74         74         75         75         75         76         76
     issu      locat       meet      updat        lot     sincer     better
        79         79         79         79         80         80         82
    short       sinc       done   question     recent    possibl   contract
        82         82         83         83         83         84         85
      end       move       data      might    continu       note       feel
        85         86         87         87         88         88         90
   resourc      sever       area   communic     realli        due     direct
        90         90         92         92         93         94         96

     posit      thing       call    develop    complet       much       even
       187        188        190        191        192        192        193
   project     design       form     expect     person    without        buy
       194        196        196        198        198        198        199
     trade     effect       rate       base       find    current      first
       199        201        201        202        202        203        203
     chang      visit    financi       high       mani    forward       good
       204        206        207        208        208        209        221
   special        don    success        per     number       week     result
       225        226        226        230        231        231        237
       web   industri    contact       made    follow      month      right
       238        239        242        242        244        249        249
     today       also       help   internet      manag       know        way
       251        260        262        262        266        269        278
     avail      state      futur       home      start     system       take
       280        280        282        285        300        302        304
       net     includ       life        see       name      onlin     within
       305        314        320        329        344        345        346
     remov       best    program      peopl     custom       year       like
       357        358        358        359        363        367        372
  interest       send     servic       look       work        day       want
       385        393        395        396        415        420        420
   product        www    account     provid       need    softwar     messag
       421        426        428        435        438        440        445
      site    address        may       list      price        new     websit
       455        461        489        503        503        504        506
    report      secur       just      offer     invest      order        use
       507        520        524        528        540        541        546
     click       x000        now        one       time       http     market
       552        560        575        592        593        600        600
      make       free      pleas      money        get     receiv     inform
       603        606        619        662        694        727        818
       can      email       busi       mail        com    compani       spam
       831        865        897        917        999       1065       1368
      will    subject
      1450       1577
>
```

Split the data frame into train and test data frames

```
emailsSparse$spam = as.factor(emailsSparse$spam)
library(caTools)
set.seed(123)
spl = sample.split(emailsSparse$spam, 0.7)
train = subset(emailsSparse, spl == TRUE)
test = subset(emailsSparse, spl == FALSE)
```

```
> emailsSparse$spam = as.factor(emailsSparse$spam)
> library(caTools)
> set.seed(123)
> spl = sample.split(emailsSparse$spam, 0.7)
> train = subset(emailsSparse, spl == TRUE)
> test = subset(emailsSparse, spl == FALSE)
>
```

Build a logistic regression model

```
spamLog = glm(spam~., data=train, family="binomial")
summary(spamLog)
```

```
> spamLog = glm(spam~., data=train, family="binomial")
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(spamLog)

Call:
glm(formula = spam ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.011   0.000    0.000    0.000    1.354

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.082e+01  1.055e+04  -0.003    0.998
X000         1.474e+01  1.058e+04   0.001    0.999
X2000       -3.631e+01  1.556e+04  -0.002    0.998
X2001       -3.215e+01  1.318e+04  -0.002    0.998
X713        -2.427e+01  2.914e+04  -0.001    0.999
X853        -1.212e+00  5.942e+04   0.000    1.000
abl         -2.049e+00  2.088e+04   0.000    1.000
access      -1.480e+01  1.335e+04  -0.001    0.999
account      2.488e+01  8.165e+03   0.003    0.998
addit        1.463e+00  2.703e+04   0.000    1.000
address     -4.613e+00  1.113e+04   0.000    1.000
allow        1.899e+01  6.436e+03   0.003    0.998
alreadi     -2.407e+01  3.319e+04  -0.001    0.999
also         2.990e+01  1.378e+04   0.002    0.998
analysi     -2.405e+01  3.860e+04  -0.001    1.000
anoth       -8.744e+00  2.032e+04   0.000    1.000
applic      -2.649e+00  1.674e+04   0.000    1.000
appreci     -2.145e+01  2.762e+04  -0.001    0.999
approv      -1.302e+00  1.589e+04   0.000    1.000
april       -2.620e+01  2.208e+04  -0.001    0.999
area         2.041e+01  2.266e+04   0.001    0.999
```

```
process      -2.957e-01  1.191e+04    0.000    1.000
product       1.016e+01  1.345e+04    0.001    0.999
program       1.444e+00  1.183e+04    0.000    1.000
project       2.173e+00  1.497e+04    0.000    1.000
provid        2.422e-01  1.859e+04    0.000    1.000
public       -5.250e+01  2.341e+04   -0.002    0.998
put          -1.052e+01  2.681e+04    0.000    1.000
question     -3.467e+01  1.859e+04   -0.002    0.999
rate         -3.112e+00  1.319e+04    0.000    1.000
read         -1.527e+01  2.145e+04   -0.001    0.999
real          2.046e+01  2.358e+04    0.001    0.999
realli       -2.667e+01  4.640e+04   -0.001    1.000
receiv        5.765e-01  1.585e+04    0.000    1.000
recent       -2.067e+00  1.780e+04    0.000    1.000
regard       -3.668e+00  1.511e+04    0.000    1.000
relat        -5.114e+01  1.793e+04   -0.003    0.998
remov         2.325e+01  2.484e+04    0.001    0.999
repli         1.538e+01  2.916e+04    0.001    1.000
report       -1.482e+01  1.477e+04   -0.001    0.999
request      -1.232e+01  1.167e+04   -0.001    0.999
requir        5.004e-01  2.937e+04    0.000    1.000
research     -2.826e+01  1.553e+04   -0.002    0.999
resourc      -2.735e+01  3.522e+04   -0.001    0.999
respond       2.974e+01  3.888e+04    0.001    0.999
respons      -1.960e+01  3.667e+04   -0.001    1.000
result       -5.002e-01  3.140e+04    0.000    1.000
 [ reached getOption("max.print") -- omitted 81 rows ]

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4409.49  on 4009  degrees of freedom
Residual deviance:   13.46  on 3679  degrees of freedom
AIC: 675.46

Number of Fisher Scoring iterations: 25

> |
```

#Build a CART model

library(rpart)

library(rpart.plot)

spamCART = rpart(spam~., data=train, method="class")

```
> library(rpart)
> library(rpart.plot)
> spamCART = rpart(spam~., data=train, method="class")
```

Build a random forest model

library(randomForest)

set.seed(123)

spamRF = randomForest(spam~., data=train)

```
> library(randomForest)
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
> set.seed(123)
> spamRF = randomForest(spam~., data=train)
>
```

#Prediction on training data

predTrainLog = predict(spamLog, type="response")

predTrainCART = predict(spamCART)[,2]

predTrainRF = predict(spamRF, type="prob")[,2]

```
> predTrainLog = predict(spamLog, type="response")
> predTrainCART = predict(spamCART)[,2]
> predTrainRF = predict(spamRF, type="prob")[,2]
>
```

Evaluate the performance of the logistic regression model on training set

table(train$spam, predTrainLog > 0.5)

training set accuracy of logistic regression

(3052+954)/nrow(train)

training set AUC of logistic regression

library(ROCR)

predictionTrainLog = prediction(predTrainLog, train$spam)

as.numeric(performance(predictionTrainLog, "auc")@y.values)

```
> # Evaluate the performance of the logistic regression model on training set
> table(train$spam, predTrainLog > 0.5)

    FALSE TRUE
  0  3052    0
  1     4  954
> # training set accuracy of logistic regression
> (3052+954)/nrow(train)
[1] 0.9990025
> # training set AUC of logistic regression
> library(ROCR)
> predictionTrainLog = prediction(predTrainLog, train$spam)
> as.numeric(performance(predictionTrainLog, "auc")@y.values)
[1] 0.9999959
>
```

Evaluate the performance of the CART model on training set

table(train$spam, predTrainCART > 0.5)

training set accuracy of CART

(2885+894)/nrow(train)

training set AUC of CART

predictionTrainCART = prediction(predTrainCART, train$spam)

as.numeric(performance(predictionTrainCART, "auc")@y.values)

```
> # Evaluate the performance of the CART model on training set
> table(train$spam, predTrainCART > 0.5)

    FALSE TRUE
  0  2885  167
  1    64  894
> # training set accuracy of CART
> (2885+894)/nrow(train)
[1] 0.942394
> # training set AUC of CART
> predictionTrainCART = prediction(predTrainCART, train$spam)
> as.numeric(performance(predictionTrainCART, "auc")@y.values)
[1] 0.9696044
>
```

```
# Evaluate the performance of the random forest model on training set

table(train$spam, predTrainRF > 0.5)

# training set accuracy of random forest

(3013+914)/nrow(train)

# training set AUC of random forest

predictionTrainRF = prediction(predTrainRF, train$spam)

as.numeric(performance(predictionTrainRF, "auc")@y.values)
```

```
> # Evaluate the performance of the random forest model on training set
> table(train$spam, predTrainRF > 0.5)

    FALSE TRUE
  0  3015   37
  1    42  916
> # training set accuracy of random forest
> (3013+914)/nrow(train)
[1] 0.9793017
> # training set AUC of random forest
> predictionTrainRF = prediction(predTrainRF, train$spam)
> as.numeric(performance(predictionTrainRF, "auc")@y.values)
[1] 0.9978155
>
```

“In terms of both accuracy and AUC, logistic regression is nearly perfect and outperforms the other two models.”

#Prediction on testing data

predTestLog = predict(spamLog, newdata=test, type="response")

predTestCART = predict(spamCART, newdata=test)[,2]

predTestRF = predict(spamRF, newdata=test, type="prob")[,2]

```
> predTestLog = predict(spamLog, newdata=test, type="response")
> predTestCART = predict(spamCART, newdata=test)[,2]
> predTestRF = predict(spamRF, newdata=test, type="prob")[,2]
>
```

Evaluate the performance of the logistic regression model on testing set

table(test$spam, predTestLog > 0.5)

(1257+376)/nrow(test)

predictionTestLog = prediction(predTestLog, test$spam)

as.numeric(performance(predictionTestLog, "auc")@y.values)

```
> # Evaluate the performance of the logistic regression model on testing set
> table(test$spam, predTestLog > 0.5)

    FALSE TRUE
  0  1257   51
  1    34  376
> (1257+376)/nrow(test)
[1] 0.9505239
> predictionTestLog = prediction(predTestLog, test$spam)
> as.numeric(performance(predictionTestLog, "auc")@y.values)
[1] 0.9627517
>
```

Evaluate the performance of the CART model on testing set

table(test$spam, predTestCART > 0.5)

(1228+386)/nrow(test)

predictionTestCART = prediction(predTestCART, test$spam)

as.numeric(performance(predictionTestCART, "auc")@y.values)

```
> # Evaluate the performance of the CART model on testing set
> table(test$spam, predTestCART > 0.5)

    FALSE TRUE
  0  1228   80
  1    24  386
> (1228+386)/nrow(test)
[1] 0.9394645
> predictionTestCART = prediction(predTestCART, test$spam)
> as.numeric(performance(predictionTestCART, "auc")@y.values)
[1] 0.963176
> |
```

Evaluate the performance of the random forest model on testing set

table(test$spam, predTestRF > 0.5)

(1290+385)/nrow(test)

predictionTestRF = prediction(predTestRF, test$spam)

as.numeric(performance(predictionTestRF, "auc")@y.values)

```
> # Evaluate the performance of the random forest model on testing set
> table(test$spam, predTestRF > 0.5)

    FALSE TRUE
  0  1291   17
  1    23  387
> (1290+385)/nrow(test)
[1] 0.9749709
> predictionTestRF = prediction(predTestRF, test$spam)
> as.numeric(performance(predictionTestRF, "auc")@y.values)
[1] 0.9975899
> |
```

“The random forest outperformed logistic regression and CART in both measures, obtaining an impressive AUC of 0.997 on the test set.”