CSCI323 - Modern AI

SPAM DETECTION

Data-Driven Feature Engineering and Multi-Model Optimization for Enhanced Spam Detection

Group No. 18

Jiwon Moon
Oh Hwee Xin
Yeonjae Lim
Chua Cheng Yi
Loh Chin Yee
Swam Pyae Aung

2024-S3

PART 1: DATA HANDLING

01 Intro

02 Data Cleaning

03 Data Preprocess

04 Data Exploration

05 Feature Engineering

06 Feature Selection

01 What are the public efforts on spam emails?

- CAN-SPAM act for opt-out mechanism
- Authentication headers
- X-Spam-Status header from SpamAssassin open source spam filtering program

01 Common Models in the Field

- Random Forests: Handle overfitting
- Logistic Regression: Simple yet powerful
- Naive Bayes: Simple and effective
- Dense Neural Networks: Understand complex pattern

01 Intro: Dataset

- *2007 TREC* Public Spam Corpus and *Enron*-Spam

```
wulvob get your medircations online qnb ikud viagra escapenumber escapenumber levitra escapenumber escapenumber cialis escapenumber
escapenumber imitrex escapenumber escapenumber flonax escapenumber escapenumber ultram escapenumber escapenumber vioxx escapenumber
escapenumber ambien escapenumber escapenumber valium escapenumber escapenumber xannax escapenumber escapenumber soma escapenumber
meridia escapenumber escapenumber escapenumber escapenumber cysfrt have you ever stopped to wonder how much an average man pays for
his mediecines ap painkillers drugs to improve the quality of life weight reducing tablets and many more escapenumber what's worse
the same mediucine costs a lot more if it is branded rfwur are you intrested so http dmvrwm remunerativ net dfuaeirxygiq visit our
website escapenumber
```

- Spam *Assassin* Dataset

```
From gort44@excite.com Mon Jun 24 17:54:21 2002 Return-Path: gort44@excite.com Delivery-Date: Tue Jun 4 05:31:16 2002 Received:
from mandark.labs.netnoteinc.com ([213.105.180.140]) by dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g544VF020182 for
<jm@jmason.org>; Tue, 4 Jun 2002 05:31:15 +0100 Received: from wi-poli.poli.cl ([200.54.149.34]) by mandark.labs.netnoteinc.com
(8.11.2/8.11.2) with SMTP id g544VC729935; Tue, 4 Jun 2002 05:31:13 +0100 Received: from 216.77.61.89 (unverified [218.5.180.148])
by wi-poli.poli.cl (EMWAC SMTPRS 0.83) with SMTP id <B0000918901@wi-poli.poli.cl>; Tue, 04 Jun 2002 00:14:29 -0400 Message-Id:
<B0000918901@wi-poli.poli.cl> To: <chrbader@telecom.at> From: "irese" <gort44@excite.com> Subject: Cash in on your home equity
Date: Tue, 04 Jun 2002 00:18:34 -1600 MIME-Version: 1.0 Content-Type: text/plain; charset="Windows-1252" X-Keywords: Content-
Transfer-Encoding: 7bit Mortgage Lenders & Brokers Are Ready to compete for your business. Whether a new home loan is what you
seek or to refinance your current home loan at a lower interest rate, we can help! Mortgage rates haven't been this low in years
take action now! Refinance your home with us and include all of those pesky credit card bills or use the extra cash for that pool
you've always wanted... Where others say NO, we say YES!!! Even if you have been turned down elsewhere, we can help! Easy terms!
Our mortgage referral service combines the highest quality loans with the most economical rates and the easiest qualifications!
Take just 2 minutes to complete the following form. There is no obligation, all information is kept strictly confidential, and you
must be at least 18 years of age. Service is available within the United States only. This service is fast and free. Free
information request form: PLEASE VISIT http://builtit4unow.com/pos
**************************************************************** Since you have received this message you have either responded to
one of our offers in the past or your address has been registered with us. If you wish to "OPT_OUT" please visit:
http://builtit4unow.com/pos ****************************************************************
```

01 Intro: Dataset

- 2007 TREC Public Spam Corpus

We chose this dataset.

```
Return-Path: <bounce-debian-mirrors=ktwarwic=speedy.uwaterloo.ca@lists.debian.org>
Received: from murphy.debian.org (murphy.debian.org [70.103.162.31])
        by speedy.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l38H9S0I003031
        for <ktwarwic@speedy.uwaterloo.ca>; Sun, 8 Apr 2007 13:09:28 -0400
Received: from localhost (localhost [127.0.0.1])
        by murphy.debian.org (Postfix) with QMQP
        id 90C152E68E; Sun,  8 Apr 2007 12:09:05 -0500 (CDT)
Old-Return-Path: <yan.morin@savoirfairelinux.com>
X-Spam-Checker-Version: SpamAssassin 3.1.4 (2006-07-26) on murphy.debian.org
X-Spam-Level:
X-Spam-Status: No, score=-1.1 required=4.0 tests=BAYES_05 autolearn=no
        version=3.1.4
X-Original-To: debian-mirrors@lists.debian.org
Received: from xenon.savoirfairelinux.net (savoirfairelinux.net [199.243.85.90])
        by murphy.debian.org (Postfix) with ESMTP id 827432E3E5
        for <debian-mirrors@lists.debian.org>; Sun,  8 Apr 2007 11:52:35 -0500 (CDT)
Received: from [192.168.0.101] (bas6-montreal28-1177925679.dsl.bell.ca [70.53.184.47])
        by xenon.savoirfairelinux.net (Postfix) with ESMTP id C1223F69B7
        for <debian-mirrors@lists.debian.org>; Sun,  8 Apr 2007 12:52:34 -0400 (EDT)
Message-ID: <46191DCE.3020508@savoirfairelinux.com>
Date: Sun, 08 Apr 2007 12:52:30 -0400
From: Yan Morin <yan.morin@savoirfairelinux.com>
User-Agent: Icedove 1.5.0.10 (X11/20070329)
MIME-Version: 1.0
To: debian-mirrors@lists.debian.org
Subject: Typo in /debian/README
X-Enigmail-Version: 0.94.2.0
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 7bit
X-Rc-Spam: 2007-01-18_01
X-Rc-Virus: 2006-10-25_01
X-Rc-Spam: 2007-01-18_01
Resent-Message-ID: <tHOiyB.A.jEC.xGSGGB@murphy>
Resent-From: debian-mirrors@lists.debian.org
X-Mailing-List: <debian-mirrors@lists.debian.org>
X-Loop: debian-mirrors@lists.debian.org
List-Id: <debian-mirrors.lists.debian.org>
List-Post: <debian-mirrors@lists.debian.org>
List-Help: <debian-mirrors-request@lists.debian.org?subject=help>
List-Subscribe: <debian-mirrors-request@lists.debian.org?subject=subscribe>
List-Unsubscribe: <debian-mirrors-request@lists.debian.org?subject=unsubscribe>
Precedence: list
Resent-Sender: debian-mirrors-request@lists.debian.org
Resent-Date: Sun,  8 Apr 2007 12:09:05 -0500 (CDT)
```

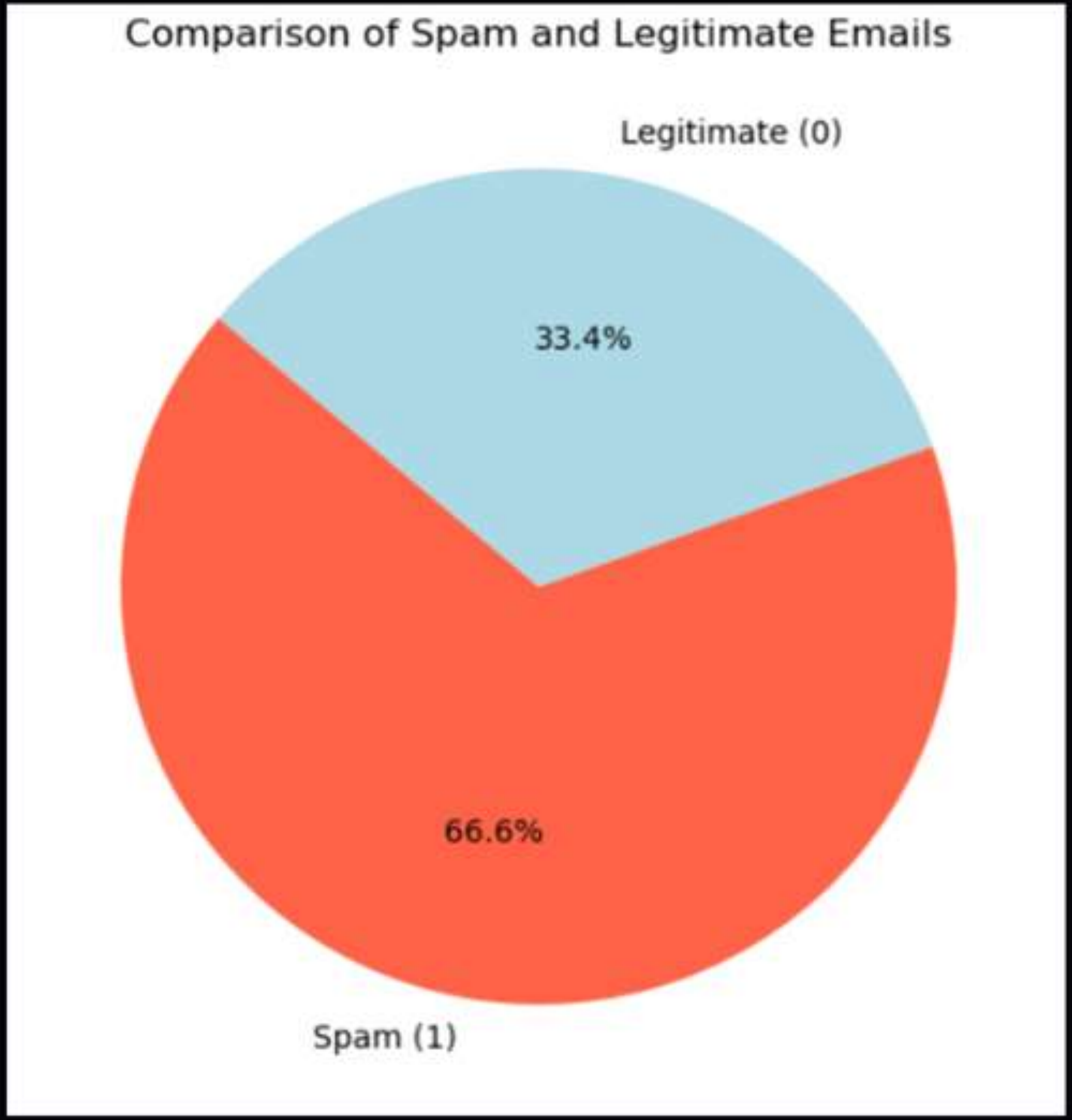
01 Intro: The credibility of the dataset

- Provided by University of Waterloo.
- Specifically for research focus use
- Dataset is part of TREC (Text Retrieval Conference) series, which is well-known in information retrieval community

https://plg.uwaterloo.ca/~gvcormac/treccorpus07/about.html

01 Intro: Initial Exploration

- 75419 email entries
- 2 columns : the label (spam(1) and ham(0)) and the origin
- 50199 labeled as spam, 25220 labeled as ham



01 Intro: Understand the data

- Want to understand the different characteristics of spam email and ham email

```python
import random

def print_rand_emails(dataset, n=3):
    for i in range(n):
        row = random.randint(1, dataset.shape[0])
        row_text = dataset.iloc[row, 1]
        row_label = dataset.iloc[row,0]
        print(f'Email {i}, identified as {"spam" if row_label == 1 else "ham"}')
        print('*'*50)
        print(row_text)


print_rand_emails(ham_emails, 2)
print_rand_emails(spam_emails, 2)
```

01 Why emails classified as Spam?

- Medication offers
  - with random words,
  - suspicious links
  - unsolicited advertising

- Promotion offers
  - could not opt out the marketing email
  - advertise lowest price
  - long list of items

01 Why emails classified as Ham?

- Technical discussion
  - specific and relevant context
  - personalised and contextual
  - no unsolicited offers or links
  - professional tone

- Subscription content
  - email from recognized organization
  - clear and relevant information
  - proper unsubscribe option

02 Data **Cleaning**

- Email often sent in multipart format to accomodate
  - Multiple content type
  - Handle attachment
  - Compatibility with email client
  - Improve deliverability

```
Return-Path: <abbkids@cox.net>
Received: from WISEGIGA ([218.158.93.211])
        by flax9.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l5544fhB003680
        for <smiles@speedy.uwaterloo.ca>; Tue, 5 Jun 2007 00:04:42 -0400
Message-ID: =?utf-8?b?PDAwMDAwMWM3YTcyNiRhMTRiMjIwMCRkMzVkOWVkYUDCu8O9wrvDqjHDhsOAwrHDqMOBw7jDiMKjPg==?=
From: "Philip" <abbkids@cox.net>
To: <smiles@flax9.uwaterloo.ca>
Subject: killer softwares for the price of nuts,vista new releases USA only
Date: Tue, 05 Jun 2007 13:04:36 +0100
MIME-Version: 1.0
Content-Type: multipart/alternative;
        boundary="------------ms050401020306070409060908"
X-Priority: 3
X-MSMail-Priority: Normal
X-Mailer: Microsoft Outlook Express 6.00.2900.3028
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.2900.3028

This is a multi-part message in MIME format.

--------------ms050401020306070409060908
```

02 Data Cleaning

- Multipart email handling
- Separation of bodies and headers
- Empty bodies or headers are replace with empty string

03 Data Preprocessing

- HTML and XML processing
  - extract the content and convert to plain text for textual analysis
  - extracts the mailto anchor tag for potential spam analysis

```
<hr size="1" noshade width="100%">
You have agreed to receive this email from CNN.com as a result of your CNN.com preference settings.<br>
To manage your settings click <a href="http://audience.cnn.com/services/cnn/memberservices/member_auth.jsp?url=http%3A%2F%2Faudience.c
To alter your alter your alert criteria or frequency or to unsubscribe from receiving custom email alerts, click <a href="http://audie
<a href="http://www.cnn.com/youralerts/refer/">Refer a friend or colleague</a> to CNN's FREE personalized alerting service!
<hr size="1" noshade width="100%">
<div class="cnnSectCopyright" style="padding-top:10px;">
Cable News Network LP, LLLP. One CNN Center, Atlanta, Georgia 30303<br>
<b>&#169; 2007 Cable News Network, LP, LLLP.</b><br>
A Time Warner Company. All Rights Reserved.<br>
<a href="http://www.cnn.com/interactive_legal.html">Terms</A>  under which this service is provided to you.<br>
Read our <a href="http://www.cnn.com/privacy.html">privacy guidelines</a>. <a href="http://www.cnn.com/feedback/">Contact us</a>.
</div>
```

03 Data Preprocessing

- Removal base64 encoded content like email and attachment

```
----5827277045207317362 9
```


```
----58272770452073173629
Content-Type: text/html;
Content-Transfer-Encoding: base64

PGh0bWwgeG1sbnM6dj0idXJuOnNjaGVtYXMtbWljcm9zb2Z0LWNvbTp2bWwiDQp4bWxuczpv
PSJ1cm46c2NoZW1hcy1taWNyb3NvZnQtY29tOm9mZmljZTpvZmZpY2UiDQp4bWxuczp3PSJ1
cm46c2NoZW1hcy1taWNyb3NvZnQtY29tOm9mZmljZTp3b3JkIg0KeG1sbnM9Imh0dHA6Ly93
d3cudzMub3JnL1RSL1JFQy1odG1sNDAiPg0KDQo8aGVhZD4NCjxtZXRhIGh0dHAtZXF1aXY9
Q29udGVudC1UeXBlIGNvbnRlbnQ9InRleHQvaHRtbDsgY2hhcnNldD13aW5kb3dzLTEyNTIi
Pg0KPG1ldGEgbmFtZT1Qcm9nSWQgY29udGVudD1Xb3JkLkRvY3VtZW50Pg0KPG1ldGEgbmFt
ZT1HZW5lcmF0b3IgY29udGVudD0iTWljcm9zb2Z0IFdvcmQgMTAiPg0KPG1ldGEgbmFtZT1P
cmlnaW5hdG9yIGNvbnRlbnQ9Ik1pY3Jvc29mdCBXb3JkIDEwIj4NCjxsaW5rIHJlbD1GaWxl
LUxpc3QgaHJlZj0ibWVuc3NhZ2UwMTBfYXJjaGl2b3MvZmlsZWxpc3QueG1sIj4NCjxsaW5r
IHJlbD1FZGl0LVRpbWUtRGF0YSBocmVmPSJtZW5zc2FnZTAxMF9hcmNoaXZvcy9lZGl0ZGF0
YS5tc28iPg0KPCEtLVtpZiAhbXNvXT4NCjxzdHlsZT4NCnZcOioge2JlaGF2aW9yOnVybCgj
ZGVmYXVsdCNWTUwpO30NCm9cOioge2JlaGF2aW9yOnVybCgjZGVmYXVsdCNWTUwpO30NCndc
Oioge2JlaGF2aW9yOnVybCgjZGVmYXVsdCNWTUwpO30NCi5zaGFwZSB7YmVoYXZpb3I6dXJs
KCNkZWZhdWx0I1ZNTCk7fQ0KPC9zdHlsZT4NCjwhW2VuZGlmXS0tPg0KPHRpdGxlPiAgPC90
aXRsZT4NCjwhLS1baWYgZ3RlIG1zbyA5XT48eG1sPg0KIDxvOkRvY3VtZW50UHJvcGVydGll
cz4NCiAgPG86QXV0aG9yPkJqPC9vOkF1dGhvcj4NCiAgPG86VGVtcGxhdGU+Tm9ybWFsPC9v
OlRlbXBsYXRlPg0KICA8bzpMYXN0QXV0aG9yPlJJQ0FSRE88L286TGFzdEF1dGhvcj4NCiAg
PG86UmV2aXNpb24+MjwvbzpSZXZpc2lvbj4NCiAgPG86VG90YWxUaW1lPjQ8L286VG90YWxU
aW1lPg0KICA8bzpDcmVhdGVkPjIwMDctMDItMDdUMTM6NDg6MDBaPC9vOkNyZWF0ZWQ+DQog
IDxvOkxhc3RTYXZlZD4yMDA3LTAyLTA3VDEzOjQ4OjAwWjwvbzpMYXN0U2F2ZWQ+DQogIDxv
OlBhZ2VzPjQ8L286UGFnZXM+DQogIDxvOldvcmRzPjYzMjwvbzpXb3Jkcz4NCiAgPG86Q2hh
cmFjdGVycz4zNDc5PC9vOkNoYXJhY3RlcnM+DQogIDxvOkNvbXBhbnk+VGhlIGhvdXplITwv
bzpDb21wYW55Pg0KICA8bzpMaW5lcz4yODwvbzpMaW5lcz4NCiAgPG86UGFyYWdyYXBocz44
PC9vOlBhcmFncmFwaHM+DQogIDxvOkNoYXJhY3RlcnNXaXRoU3BhY2VzPjQxMDM8L286Q2hh
```

03 Data Preprocessing

Text Processing on Email Bodies and Subject to prepare for textual analysis

- Lowercase text
- Remove punctuation
- Retain the number
  - Notice a lot number pop up in medication offer email, it might be helpful
- Tokenize words and remove stopwords with NLTK

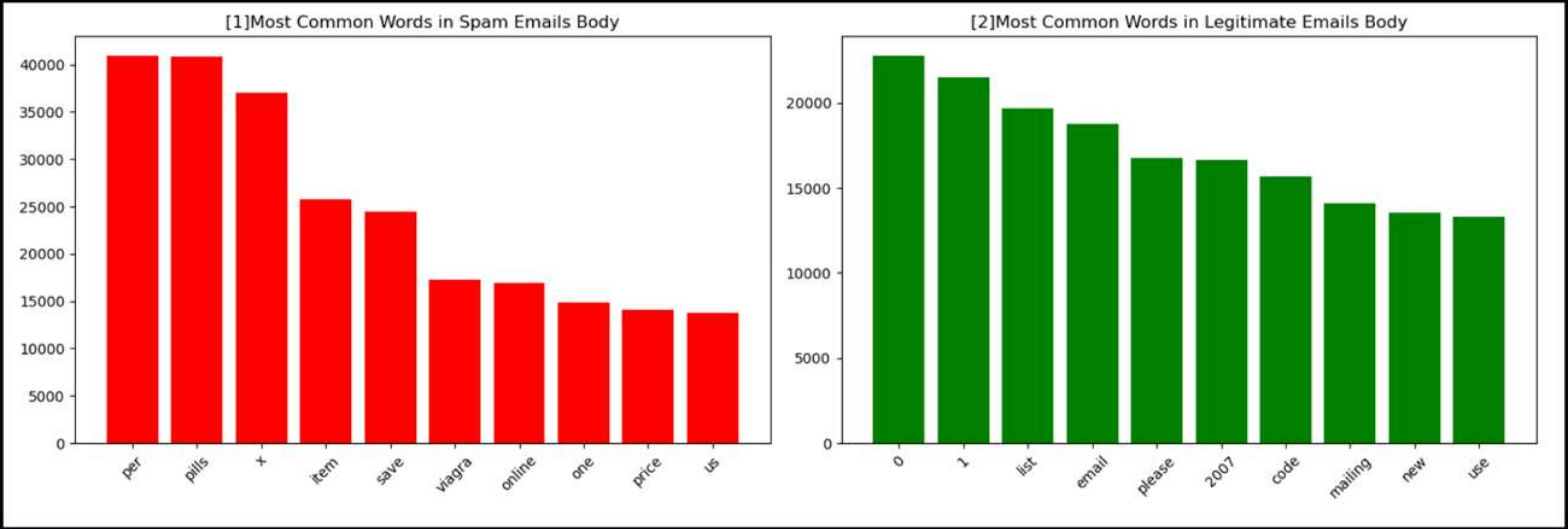
03 Data **Preprocessing**

Email headers processing

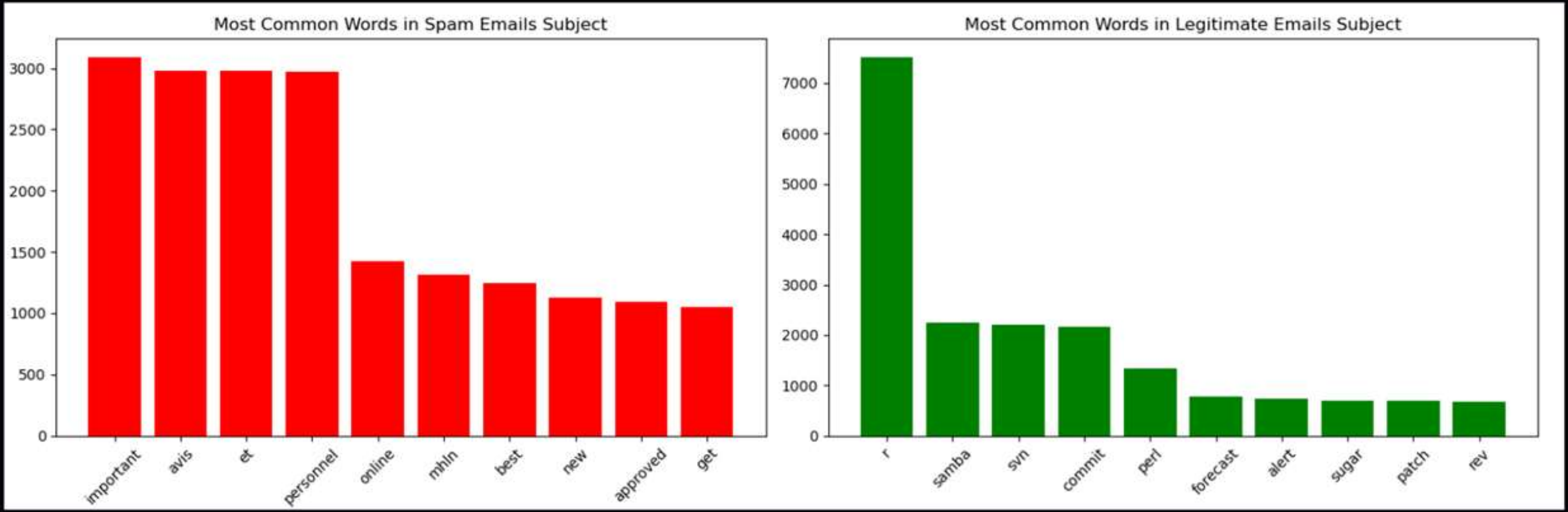
Following the initial exploration and research, we process headers that could be potentially useful (List-Unsubscribe, X-Spam-Status, etc) for spam email analysis. Empty row is replaced with ('') or ([]) or None depending on condition.

	label	return_path	from	received	list_unsubscribe	x_spam_status	authentication	list_subscribe	list_post	list_help
0	1	<RickyAmes@aol.com>	Tomas Jacobs <RickyAmes@aol.com>	[from 129.97.78.23 ([211.202.101.74])\tby spee...	None	no_info	{'SPF': False, 'DKIM': False, 'DMARC': False}	None	None	None
1	0	<bounce-debian-mirrors=ktwarwic=speedy.uwaterl...	Yan Morin <yan.morin@savoirfairelinux.com>	[from murphy.debian.org (murphy.debian.org [70...	<mailto:debian-mirrors-request@lists.debian.or...	no	{'SPF': False, 'DKIM': False, 'DMARC': False}	<mailto:debian-mirrors-request@lists.debian.or...	<mailto:debian-mirrors@lists.debian.org>	<mailto:debian-mirrors-request@lists.debian.or...

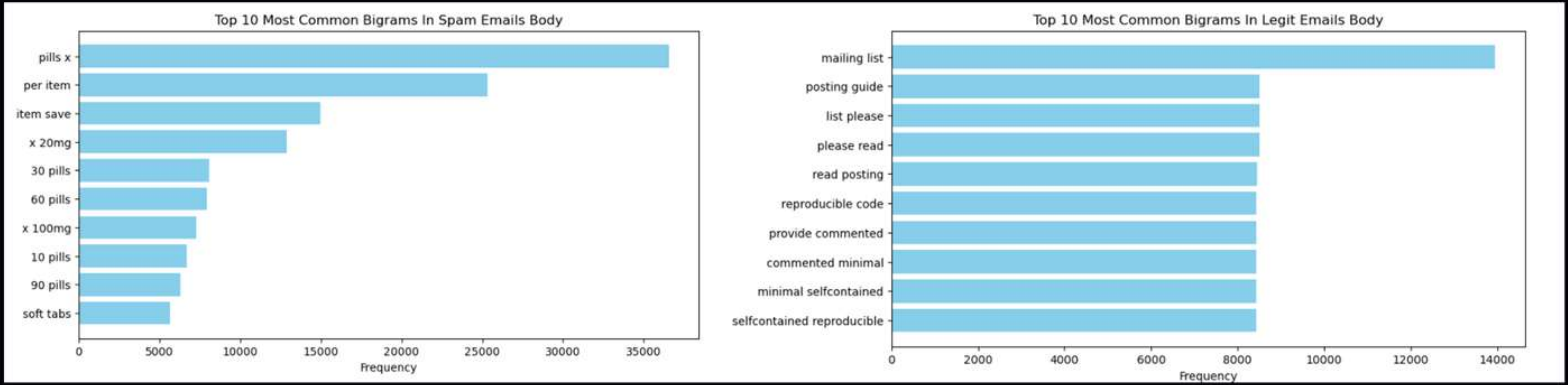
04 Data Exploration - Word Count Analysis



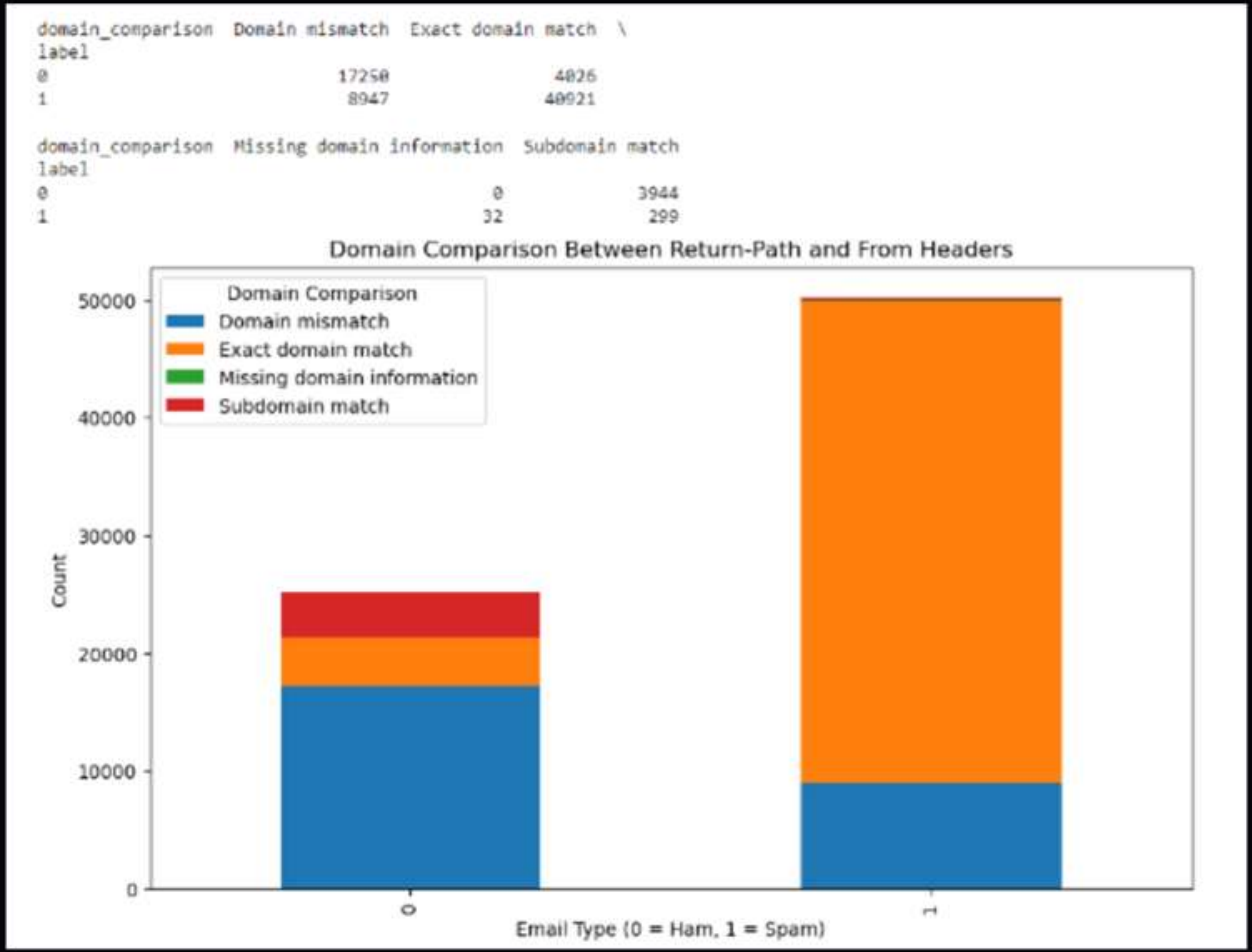
04 Data Exploration - Word Count Analysis



04 Data Exploration - Word Count Analysis



04 Data Exploration - Domain Comparison



<u>Spam</u>

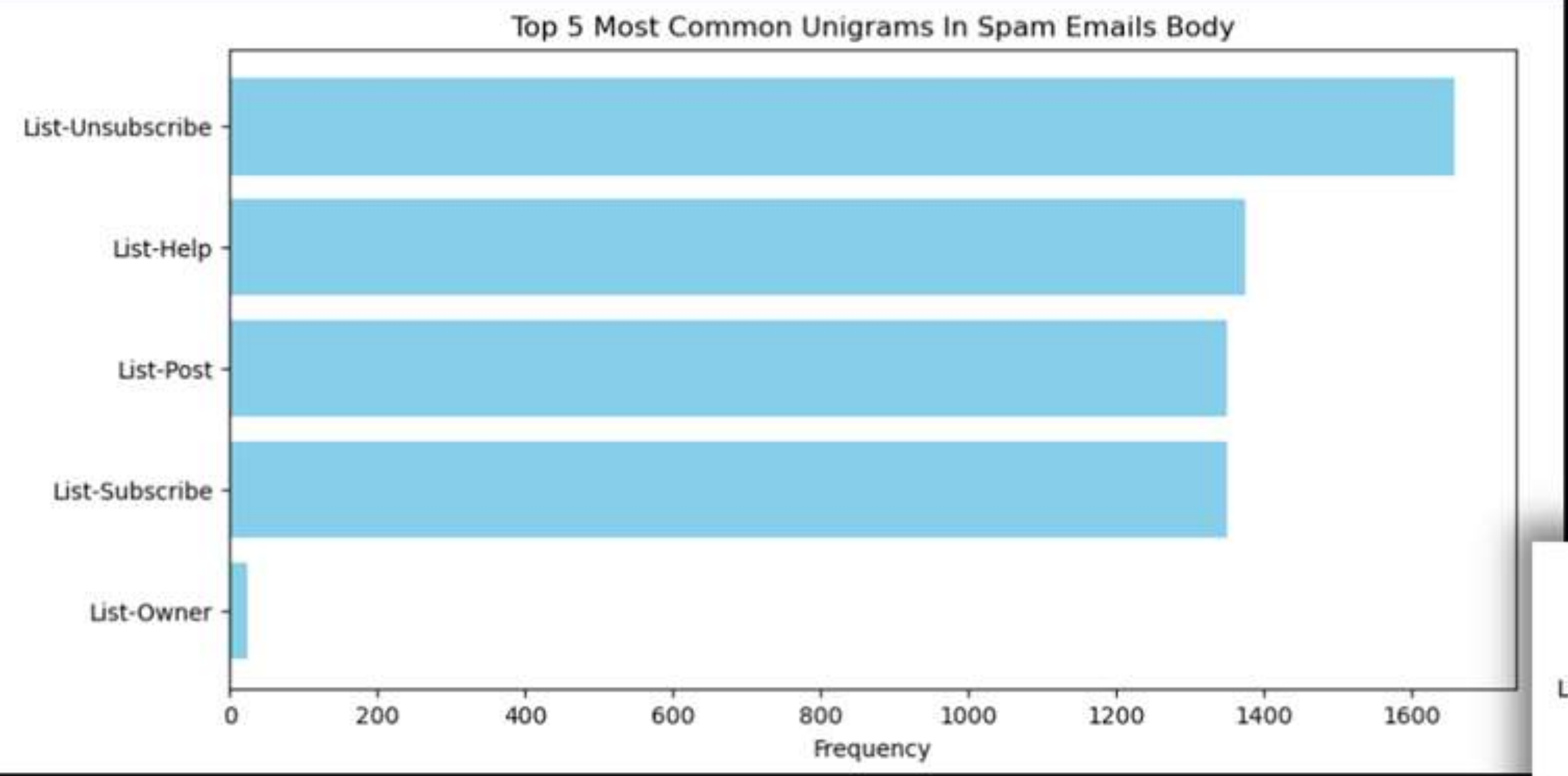
- Domain mismatch

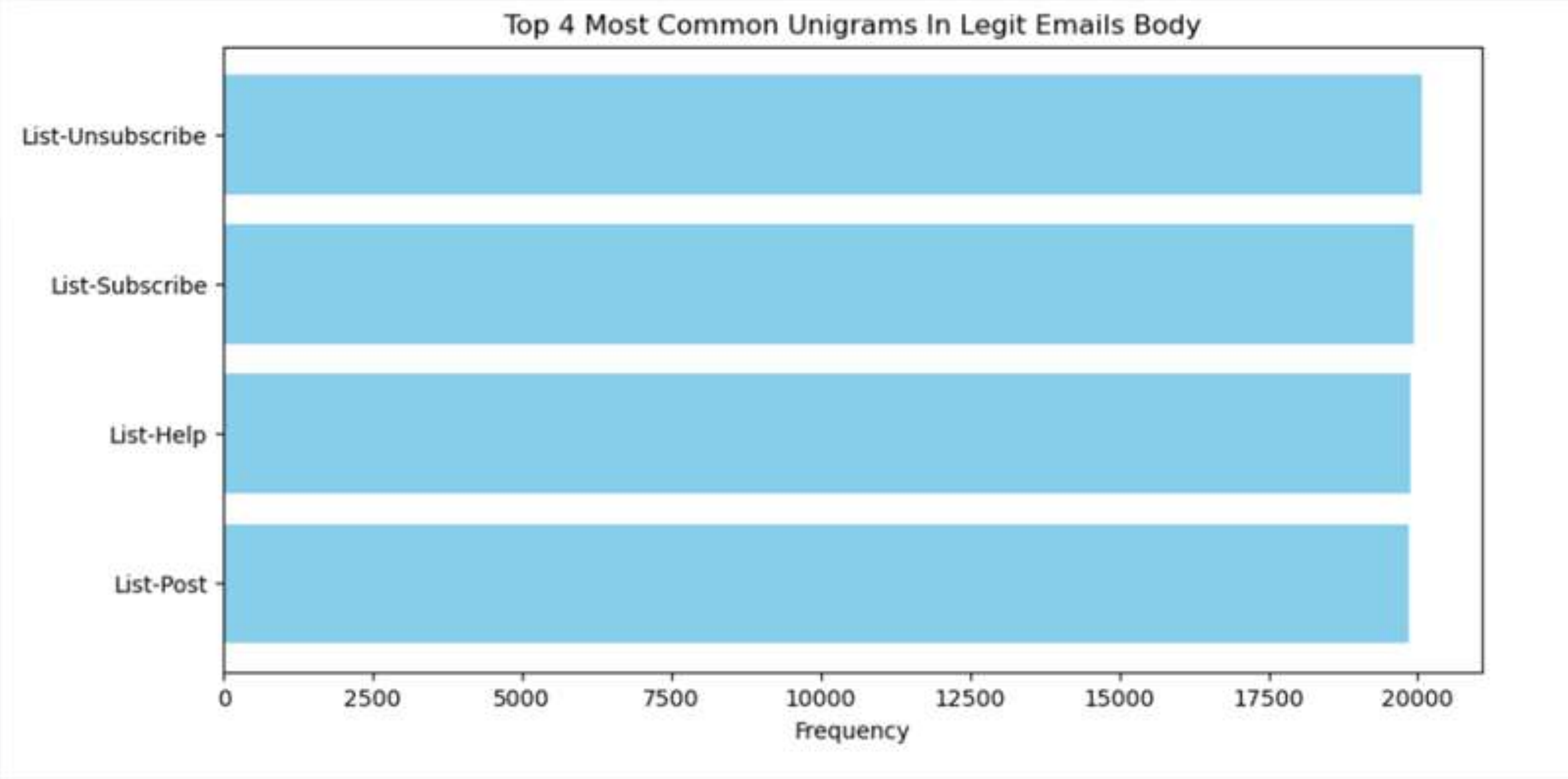
<u>Non-Spam</u>

- Exact Domain match

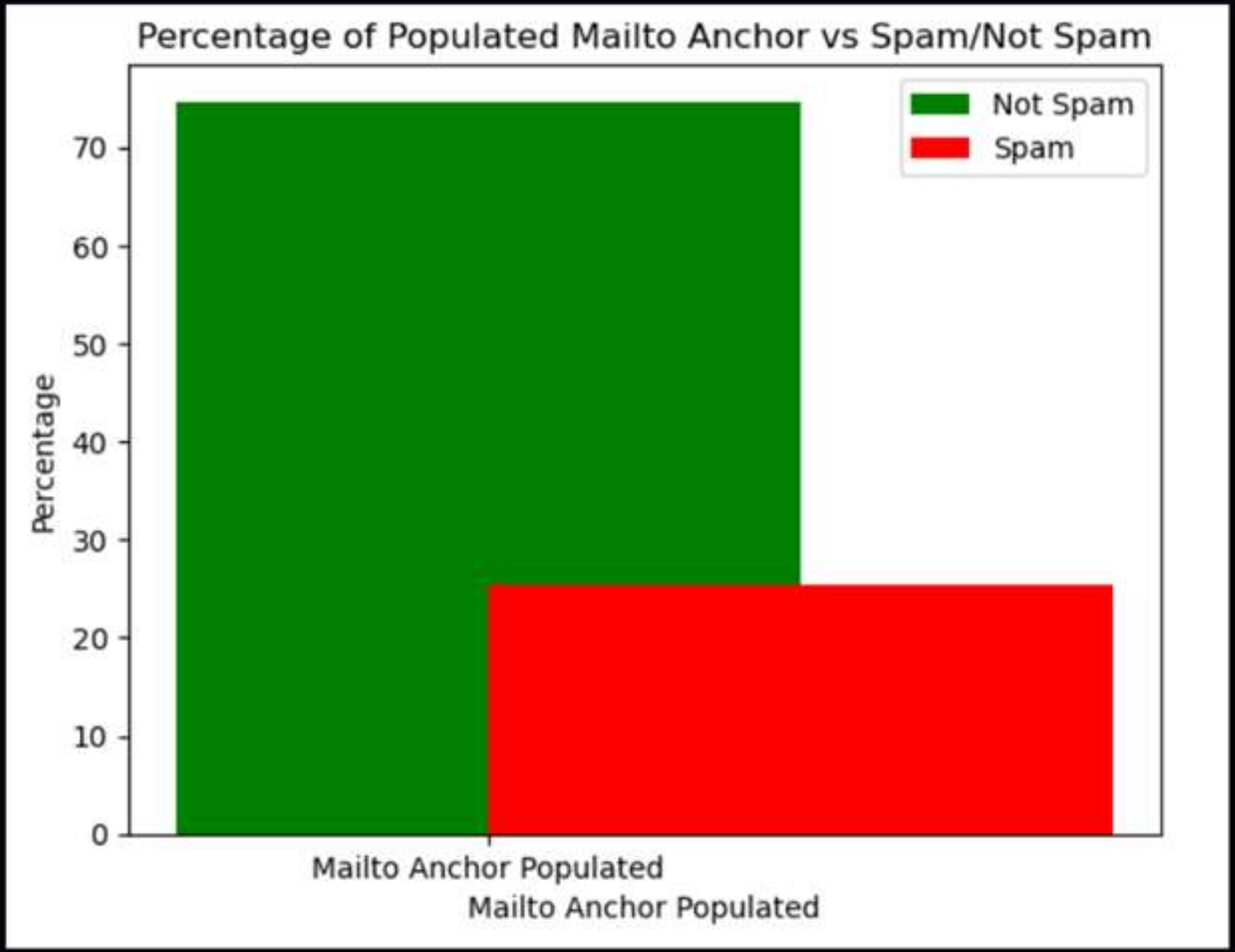
Spammers might align these domains to avoid detection.

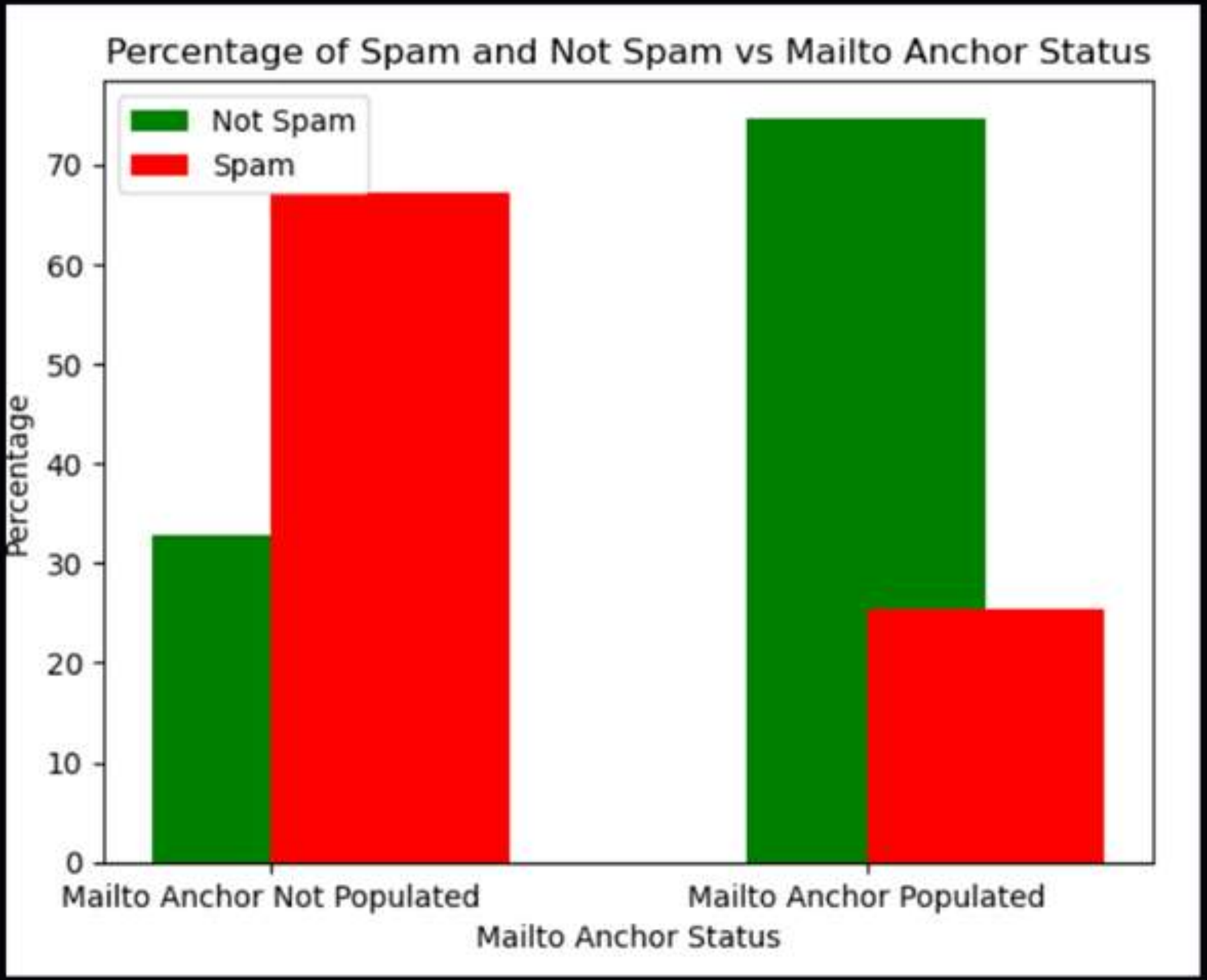
04 Data Exploration - 'mailto_header'



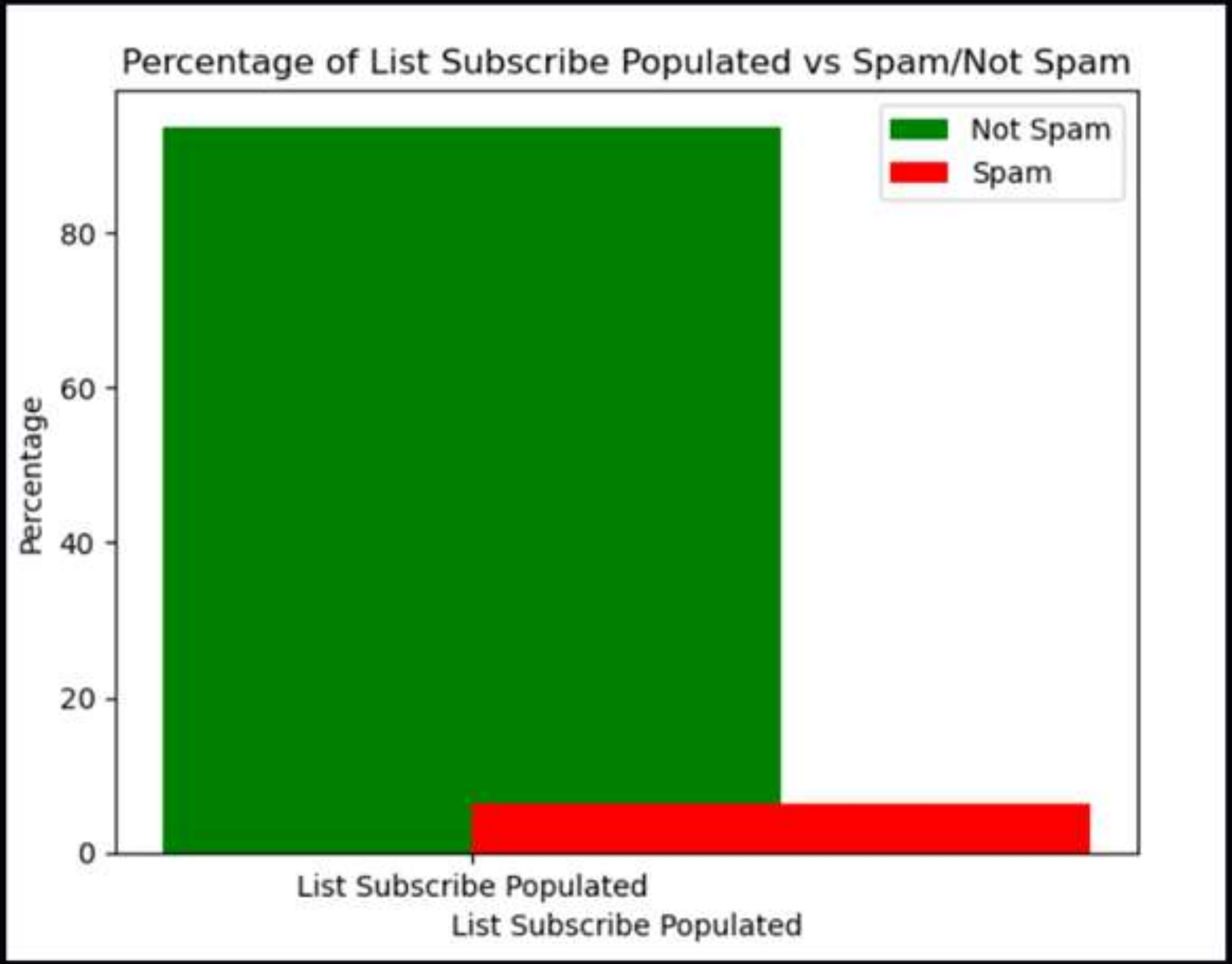


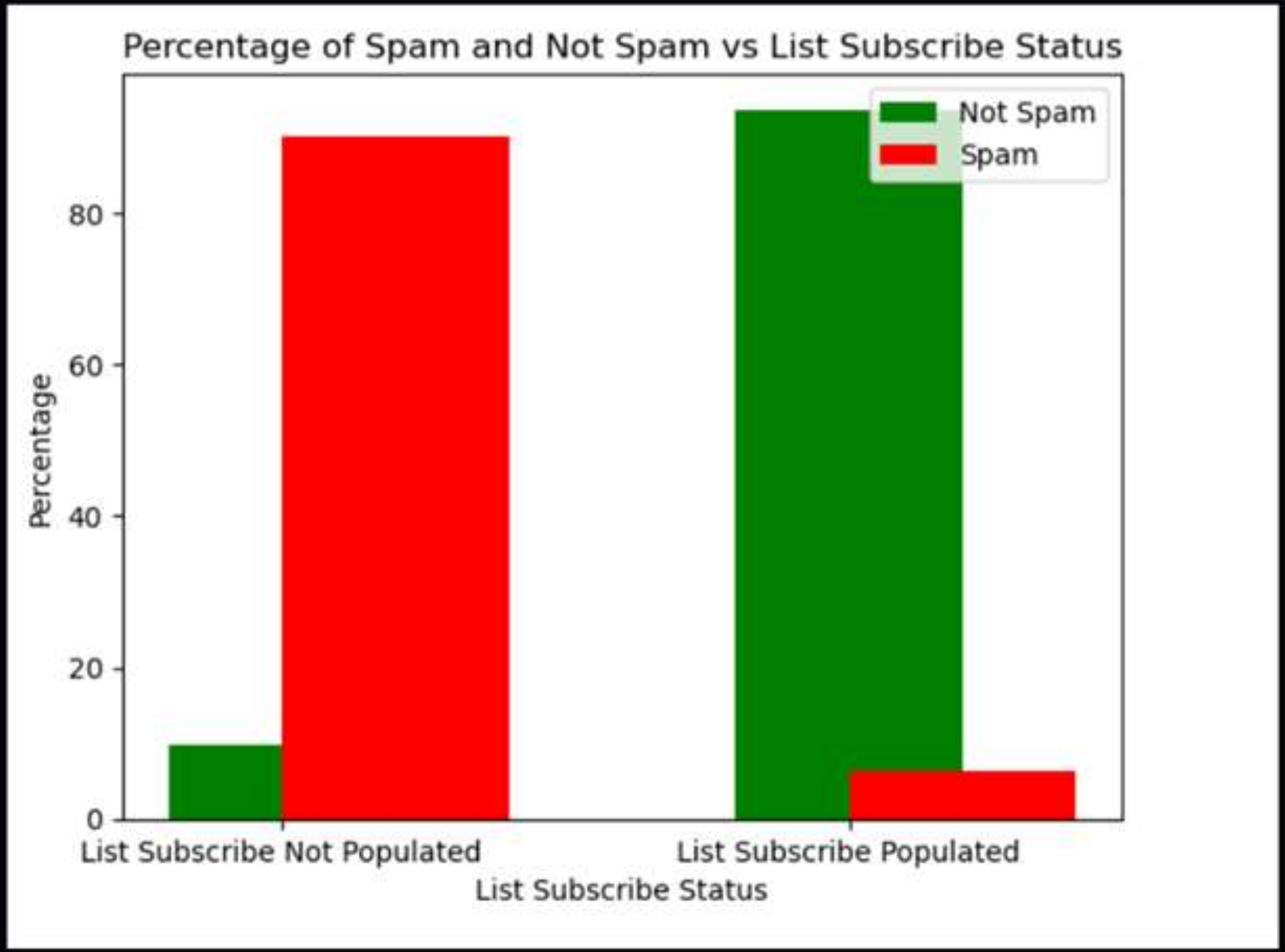
04 Data Exploration - 'mailto_anchor'



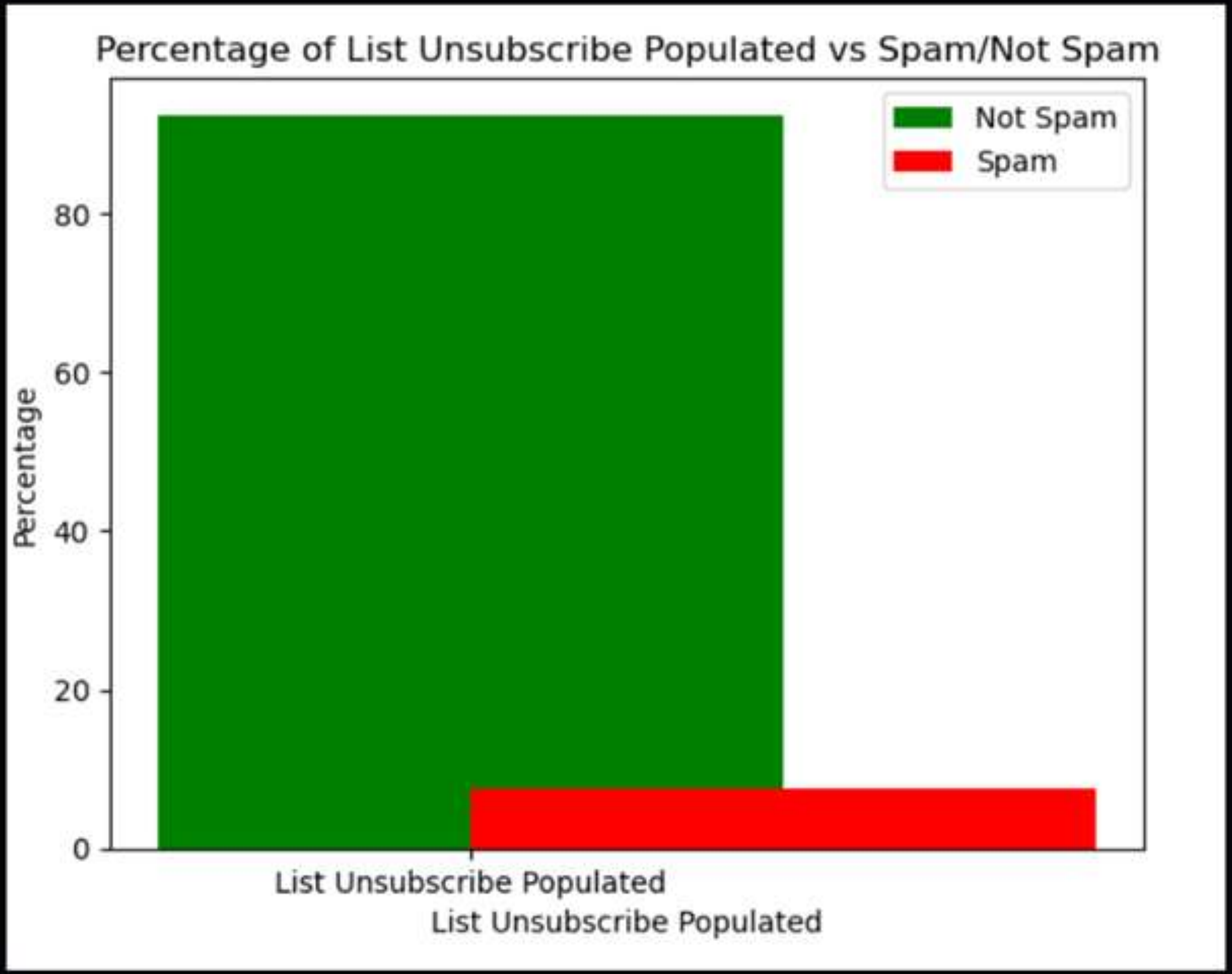


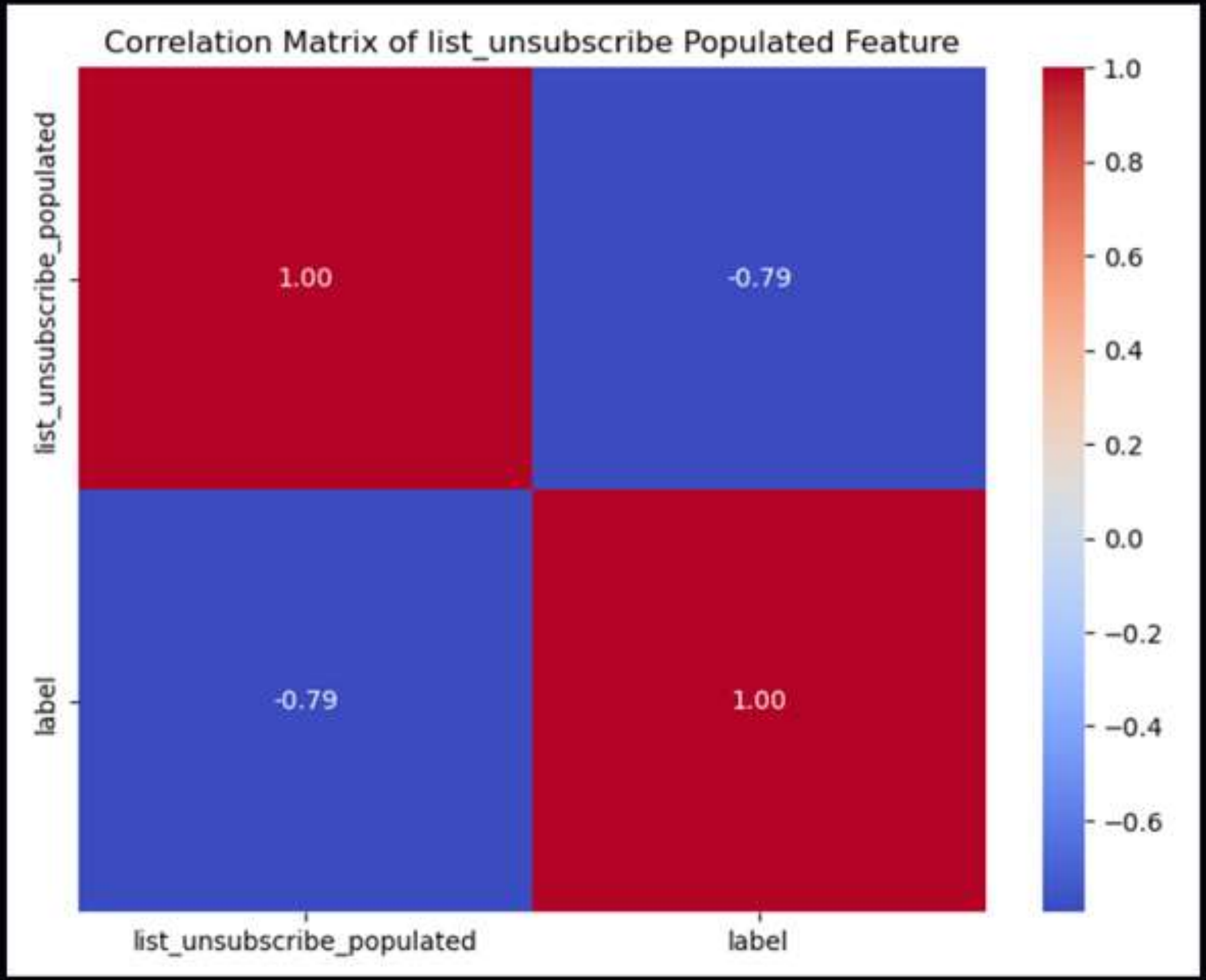
04 Data Exploration - 'list_subscribe'





04 Data Exploration - 'list_unsubscribe'





05

Feature Engineering (1)

New features that are more suitable for use in the model were extracted from existing features, while simultaneously conducting tests to verify their relationship with the labels.

	mailto_anchor, mailto_header	list_unsubscribe	authentication
New Feature	mailto_populated	list_unsubscribe_populated	is_authenticated
What to create	Binary indicator for the presence of a mailto link.	Binary indicator for the presence of a list-unsubscribe link.	Binary indicator for authentication by at least one method.
Phi Correlation Test	Phi coefficient: - 0.81 P-value ≈ 0.0	Phi coefficient: - 0.79 P-value ≈ 0.0	Phi coefficient: - 0.37 P-value ≈ 0.0

05

Feature Engineering (2)

New features that are more suitable for use in the model were extracted from existing features, while simultaneously conducting tests to verify their relationship with the labels.

	received	body
New Feature	relay_count	body_duplicates
What to create	Continuous variable indicating the number of received headers	Continuous variable that specifies the number of times content is duplicated
Point-biserial Correlation Test	Correlation: -0.59	Correlation: 0.14

05* Additional Correlation Test

Relationship verification tests were conducted between the unchanged features and the labels.

	x_spam_status	domain_comparison
Cramer's V Test	Cramer's V: 0.76	Cramer's V: 0.64

06

Feature Selection

Selected

- subject
- body
- list_unsubscribe_populated
- mailto_populated
- is_authenticated
- from_email
- return_email
- x_spam_status
- domain_comparison

Why?

- High correlation coefficient
- P-value converging to zero
- Contextually significant

Removed

- relay_count
- body_duplicates
- to

Why?

- Low correlation coefficient
- Interpretation differs from initial expectations despite not having low correlation
- Excluded to avoid contextual issues and logical conflicts

06* Data Transformation

Vectorization

Text Columns: subject, body

It is an essential step before inputting text columns into the model.
It can effectively control high cardinality.
By appropriately adjusting the max_features of the TfidfVectorizer, the vector dimension and overfitting can be regulated.

One- hot Encoding:
x_spam_status, domain_comparison

One-hot encode columns with three or more categories..

One-hot Encoding

07 Models

08 Fine Tuning

09 Evaluation

10 Conclusion

PART 2: IMPLEMENTATION

07 Models

Splitting Dataset

- ***Holdout*** Method for splitting training, validation, and test set (6:2:2)
- ***Stratified*** splitting is used to ensure that the label proportions of the original dataset are maintained within each subset.

Evaluation Metrics

- ***ROC-AUC score***: Indicates the overall performance of the model
- ***Classification Report***: Analyzes the detailed predictive capabilities of the model

07 Models

We chose 3 Models

Logistic Regression

- Simple and effective for binary classification tasks
- Less prone to overfitting, especially when regularized

Random Forest

- Capture complex, non-linear relationships between features
- Ensemble nature improves robustness and generalization

Dense Neural Network (DNN)

- Powerful feature extraction capabilities, especially for large and complex datasets
- Capture subtle patterns through multiple layers of neurons

07 Models

Logistic Regression

Basic Parameters:

- C = 0.01
- max_iter = 1000
- class_weight='balanced'

```
Logistic Regression Validation Results:
Validation AUC Score: 0.9808

Validation Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.91      0.92      5044
           1       0.95      0.97      0.96     10040

    accuracy                           0.95     15084
   macro avg       0.94      0.94      0.94     15084
weighted avg       0.95      0.95      0.95     15084

Logistic Regression Test Results:
Test AUC Score: 0.9826

Test Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.91      0.92      5044
           1       0.95      0.97      0.96     10040

    accuracy                           0.95     15084
   macro avg       0.94      0.94      0.94     15084
weighted avg       0.95      0.95      0.95     15084
```

07

Models

Random Forest

Basic Parameters:

- n_estimators = 100
- max_depth = 3
- min_samples_leaf=5
- class_weight='balanced'

```
Random Forest Validation Results:
Validation AUC Score: 0.9763

Validation Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.88      0.92      5044
           1       0.94      0.99      0.96     10040

    accuracy                           0.95     15084
   macro avg       0.96      0.93      0.94     15084
weighted avg       0.95      0.95      0.95     15084

Random Forest Test Results:
Test AUC Score: 0.9826

Test Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.91      0.92      5044
           1       0.95      0.97      0.96     10040

    accuracy                           0.95     15084
   macro avg       0.94      0.94      0.94     15084
weighted avg       0.95      0.95      0.95     15084
```

07 Models

Dense Neural Network

Basic Parameters:

- Dense layers: 64 units, ReLU
- Dropout rate: 0.5
- Output layer: 1 unit, Sigmoid
- Optimizer: Adam (lr=0.001)
- Loss: binary_crossentropy
- Epochs: 10
- Batch size: 32
- Class weight: 'balanced'

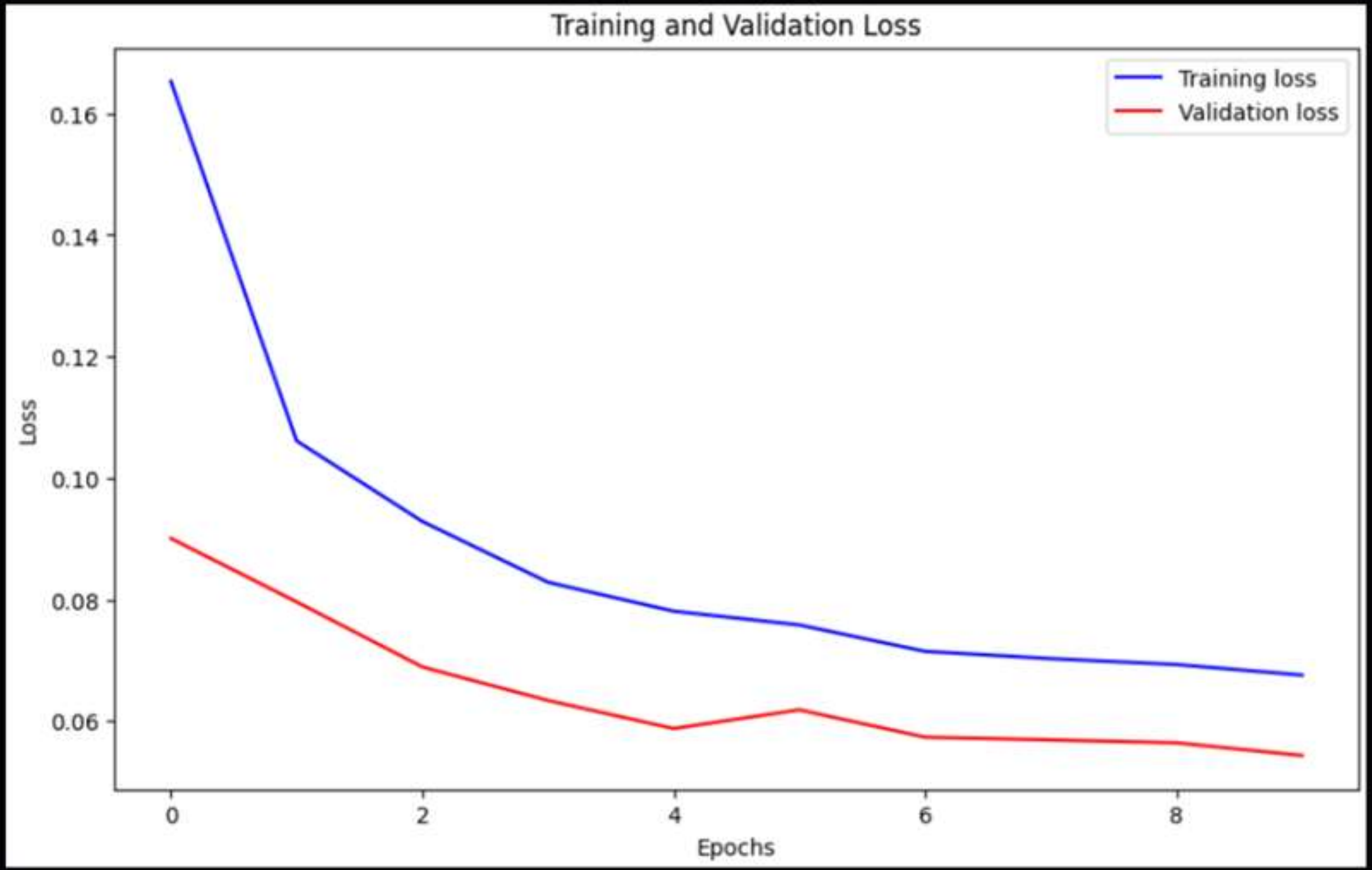
```
Dense Validation Results:
Validation AUC Score: 0.9963

Validation Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.97      5044
           1       0.99      0.99      0.99     10040

    accuracy                           0.98     15084
   macro avg       0.98      0.98      0.98     15084
weighted avg       0.98      0.98      0.98     15084

\Dense Test Results:
Test AUC Score: 0.9965

Test Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      5044
           1       0.99      0.99      0.99     10040

    accuracy                           0.98     15084
   macro avg       0.98      0.98      0.98     15084
weighted avg       0.98      0.98      0.98     15084
```

07 Models

Dense Neural Network

Basic Parameters:

- Dense layers: 64 units, ReLU
- Dropout rate: 0.5
- Output layer: 1 unit, Sigmoid
- Optimizer: Adam (lr=0.001)
- Loss: binary_crossentropy
- Epochs: 10
- Batch size: 32
- Class weight: 'balanced'



08

Fine Tuning

Key Components in ML Models

Hyperparameters are crucial settings that control the learning process in machine learning models. They are not learned from the data but are set prior to training. These parameters significantly influence model performance and play a vital role in optimizing spam detection algorithms.

08 Fine Tuning

Logistic Regression

Best Parameters:

- C = 1
- max_iter = 1000
- class_weight='balanced'

Validation AUC score: 0.9925

Validation classification report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	5044
1	0.97	0.98	0.98	10040
accuracy			0.97	15084
macro avg	0.96	0.96	0.96	15084
weighted avg	0.97	0.97	0.97	15084

Test AUC score: 0.9934

Test classification report:

	precision	recall	f1-score	support
0	0.96	0.95	0.96	5044
1	0.98	0.98	0.98	10040
accuracy			0.97	15084
macro avg	0.97	0.97	0.97	15084
weighted avg	0.97	0.97	0.97	15084

08

Fine Tuning

Random Forest

Best Parameters:

- n_estimators = 100
- max_depth = 5
- min_samples_leaf = 3
- class_weight = 'balanced'

```
Validation AUC Score: 0.9812

Validation Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.87      0.93      5044
           1       0.94      1.00      0.97     10040

    accuracy                           0.96     15084
   macro avg       0.97      0.94      0.95     15084
weighted avg       0.96      0.96      0.96     15084
Test AUC Score: 0.9832

Test Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.88      0.93      5044
           1       0.94      1.00      0.97     10040

    accuracy                           0.96     15084
   macro avg       0.97      0.94      0.95     15084
weighted avg       0.96      0.96      0.96     15084
```

08 Fine Tuning

Dense Neural Network

Basic Parameters:

- Dense layers: 64 units, ReLU
- Dropout rate: 0.5
- Output layer: 1 unit, Sigmoid
- Optimizer: Adam (lr=0.001)
- Loss: binary_crossentropy
- Epochs: 10
- Batch size: 64
- Class weight: 'balanced'

Validation AUC Score: 0.9963

Validation Classification Report:

	precision	recall	f1-score	support
0	1.00	0.96	0.97	5044
1	0.98	1.00	0.99	10040
accuracy			0.98	15084
macro avg	0.99	0.98	0.98	15084
weighted avg	0.98	0.98	0.98	15084

Test AUC Score: 0.9966

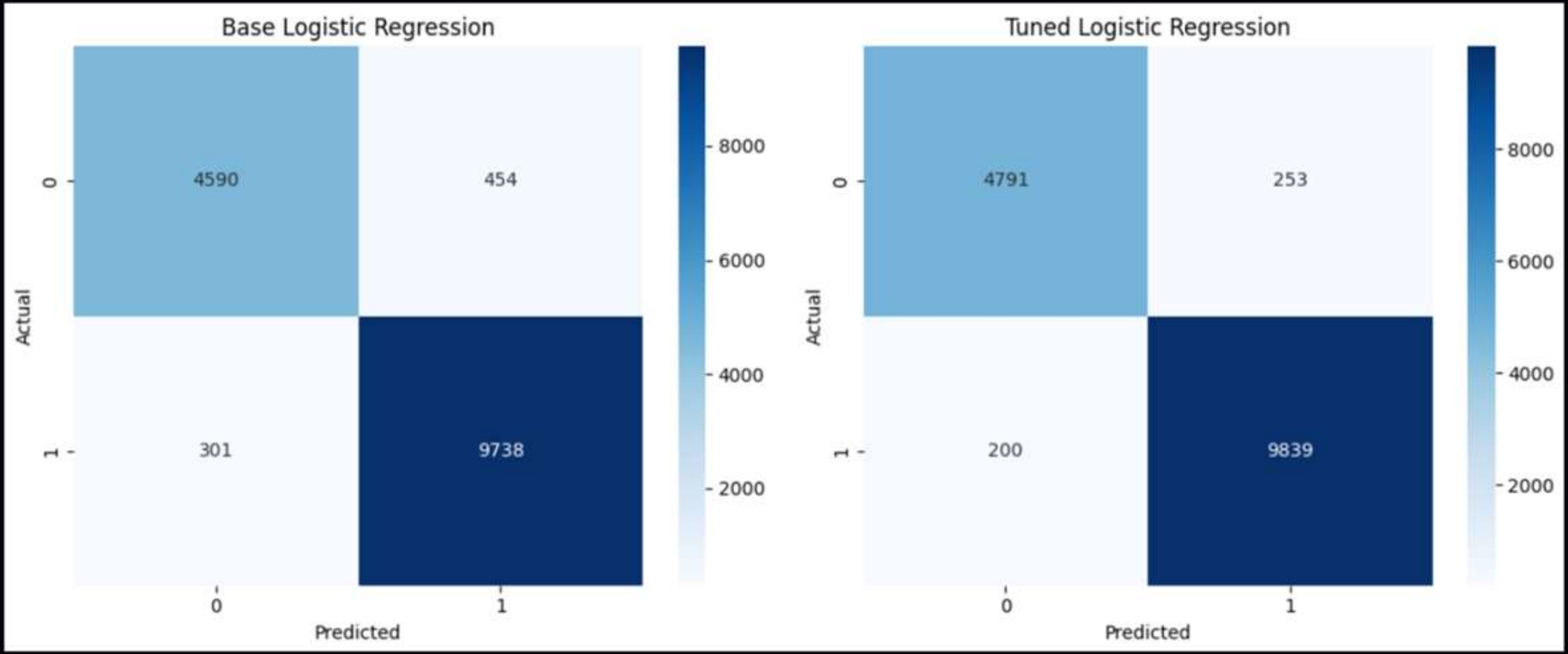
Test Classification Report:

	precision	recall	f1-score	support
0	0.99	0.96	0.97	5044
1	0.98	1.00	0.99	10040
accuracy			0.98	15084
macro avg	0.99	0.98	0.98	15084
weighted avg	0.98	0.98	0.98	15084

09 Evaluation

Logistic Regression

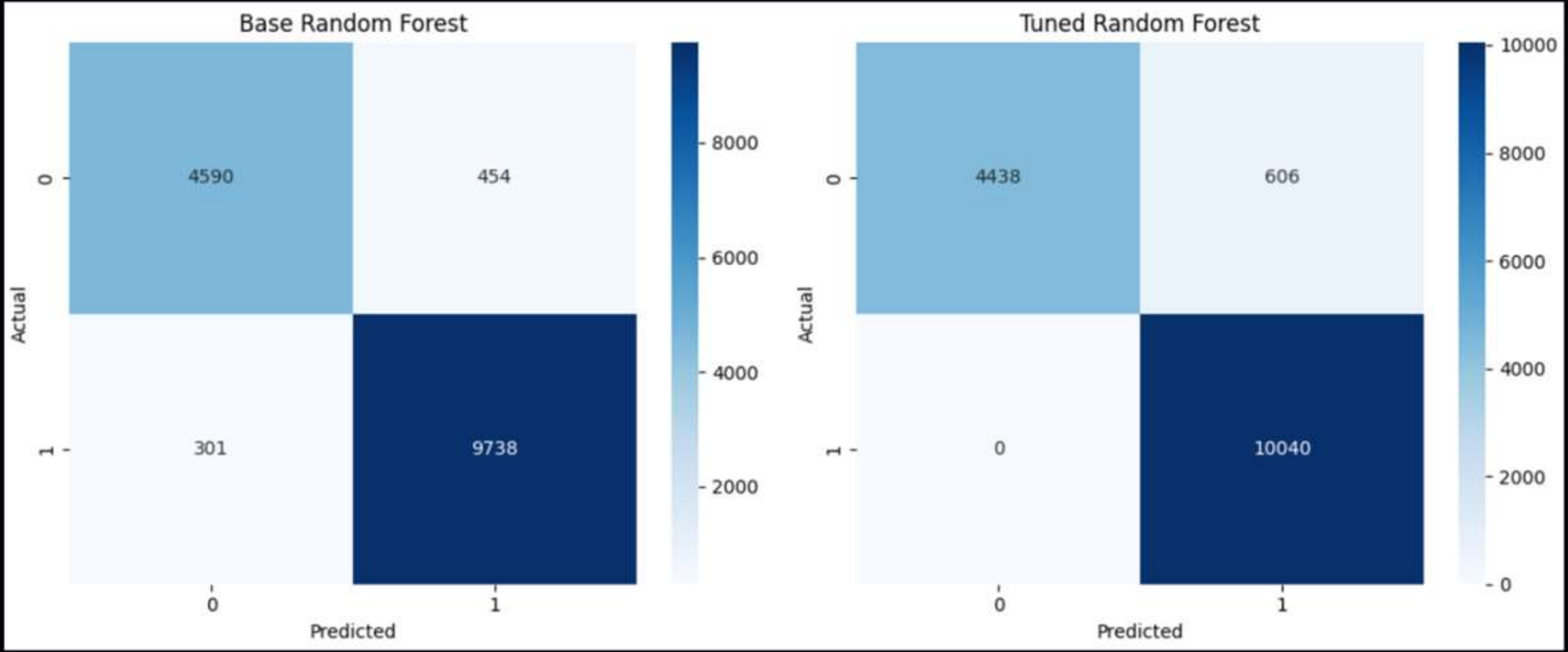
Base Model Test AUC: 0.9826
Tuned Model Test AUC: 0.9934



Evaluation

Random Forest

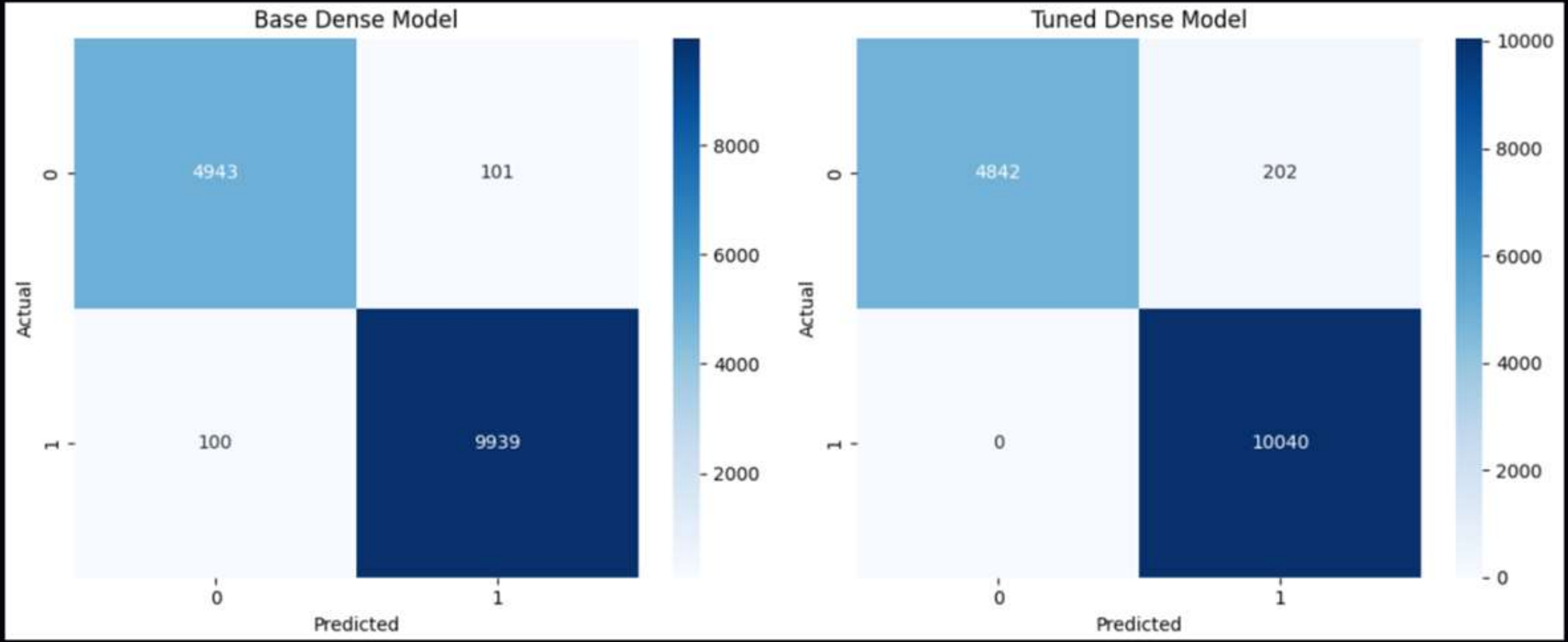
Base Model Test AUC: 0.9826
Tuned Model Test AUC: 0.9832



09 Evaluation

Dense Neural Network

Base Model Test AUC: 0.9965
Tuned Model Test AUC: 0.9966



10 Conclusion

Do we answer the 'Data Driven Feature Engineering and Multi Model Optimization for Enhanced Spam Email Detection'?

10 Conclusion : Feature Engineering

Comparing with and without feature engineering

- Before feature engineering with 2 columns: .90 accuracy
- After feature engineering: average around .98 to .99
- Utilize hold-out and stratified splitting to address overfitting concerns

What does feature engineering helps?

- Capture patterns and relationships within data better
- Improve the interpretability on the input variables
- Reduce the cardinalities of the complex data

10 Conclusion: Model Optimization

Comparing base and tuned models showed slight overall improvement, with the dense model performing best.
All base models had very high accuracy initially. This was due to subject and body being strong predictors of the label.

To dealing with this overfitting concern..

- Fine tuning for max_feature
- Balance the label distribution (class_weight = 'balanced')
- Adjusting parameters involved in controlling overfitting in each model
- Display learning curve graph

10 What can we improve further in future?

- Feature Importance Analysis
- Utilized interpretability tools like SHAP and LIME
- Apply model in different datasets for datashift problem and adversarial shift problem

Thank you for watching our Presentation!

Group No. 18

Jiwon Moon
Oh Hwee Xin
Yeonjae Lim
Chua Cheng Yi
Loh Chin Yee
Swam Pyae Aung