PROOFPOINT: LEARNING RULES TO TRIAGE CYBERSECURITY ALERTS

Abraham Kwok, Jade Mon, Sofiia Surzhak

Motivation and Context

Business Context

Our mentors are two Data Scientists from Proofpoint. Proofpoint develops software to protect people, data and brands against cyber attacks. Over 1,000,000 emails go through Proofpoint's system every hour. Each needs to be classified as malicious or safe.
ML models running in production need to be **fast** and **inexpensive**. Large Language Models (LLMs) are slow and expensive, while Transformer Models are not interpretable.

Dataset Summary

We trained our models on a 2,897-entry dataset. We focused on 14 main columns: "text" contains the email texts, "PHISH-GT" contains a the associated labels (1 for "phish", 0 for "not phish"), and 12 columns ("billing", ..., "work") each containing a binary label. The data from the latter 12 columns is noisy and was generated by a LLM. We refer to them as "soft labels".

	Unnamed: 0	text	PHISH-GT	billing	account	generic	attachment	typos	click-link	grammar	login	urgency	phish	unsolicited	work
0	42	Please take a look at the attached and give me...	0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
1	44	Rod,\nI wanted to forward this to you. Arthur...	0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

Detailed dataset description: https://monkey.org/~jose/tmp/PHISHING-FINAL-03-KN.pdf

What is a 'phishing' email?

A "phishing" email is malicious; it aims to create an adverse outcome for the receiver by deception. A "phishing" email is often (not exclusively) characterised by:

- It creates a sense of emergency and calls for immediate action.
- It doesn't include the receiver's personal information (such as name)

Please Click Here to Validate your email account Thank you, IT-Service Help Desk This email contains PRIVILEGED and CONFIDENTIAL INFORMATION intended only for the use of the recipient named above. The information may be protected by state and federal laws, including, without limitation, the provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which prohibit unauthorized disclosure. If you are not the intended recipient, you are hereby notified that any use or dissemination of this information is strictly prohibited. If you have received this email in error, please immediately notify the sender by reply email at the address provided above and delete this message. Thank you. Please¬†Click Here¬†to Validate your email account Thank you, IT-Service

Figure 1: Example of a "phishing" email from the dataset.

Interpretability and Explainability

An interpretable model is one whose predictions or decisions can be easily understood and traced back to the input features. This allows us to easily see how decisions were made by a classifier. In cybersecurity, this allows us to explain what features of an email made it appear as "phishing", which is very important due to analyst having to understand why exactly a model is making those decisions and having to attempt to explain it to others.

What is the 'best' model for detecting 'phishing' activity?

Precision, Recall and F1 Score in Cybersecurity

In cybersecurity, it is key to detect fraudulent activity, thus have a high precision. It is also important fraudulent activity aren't miscategorized. Precision and recall are extremely important, with the former being about the accuracy of those labeled as scams and the latter being the accuracy of those categorized as good emails. With precision and recall being linked, we can use a F1 score to find the best balance between both, with the best F1 score being 100%.

Predicting "Phish" Labels Via "Soft" Labels

Our data had 12 labels, each analyzing the content of the emails. These labels included grammar structure and content, using a 0 or 1 for whether or not the label category criteria was met. For example, an email calling for immediate action would be labeled 1 in the "urgency" category. Using these labels, we created a decision tree ML model and constructed a Precision-Recall curve.

```
Noisy Labels Confusion Matrix With Threshold of 0.96
[[264   9]
 [ 26 281]]
Accuracy: 0.8724137931034482
Precision: 1.0
Recall: 0.758957654723127
F1 Score: 0.862962962962963
```

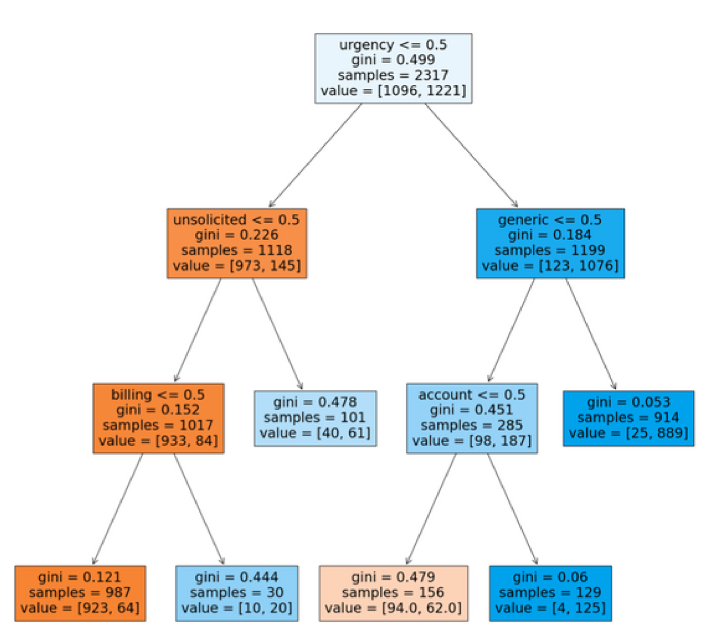


Figure 5: Noisy Label Decision Tree

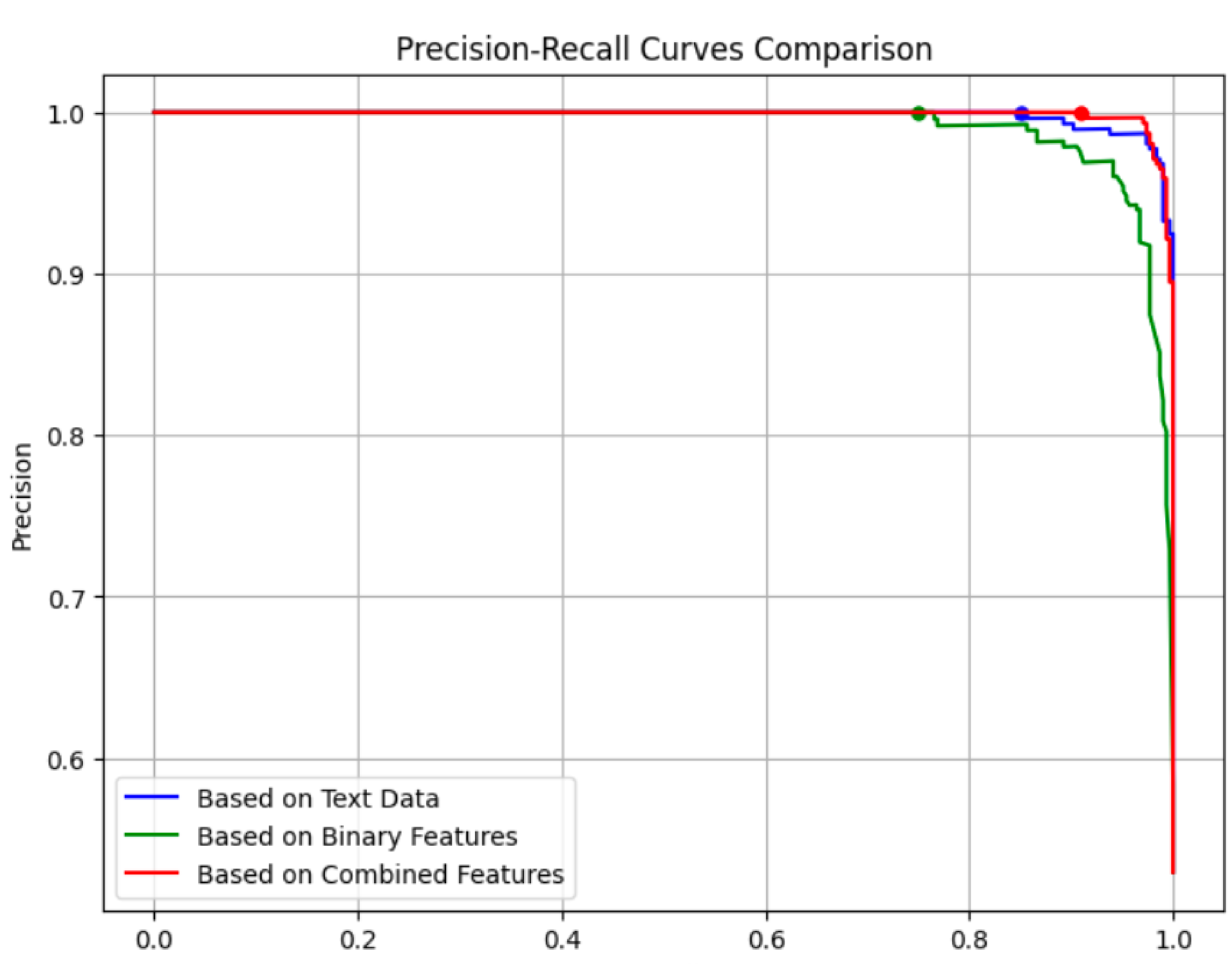


Figure 4: Precision-Recall curves for the three models.

Predicting "PHISH-GT" Label Via Email Text

When only the email bodies are available, we can make predictions using TF-IDF. TF-IDF stands for Term Frequency - Inverse Document Frequency. It evaluates how relevant a token (ex: a word) is to a document in a collection of documents (in our example, to a single email in the the entire dataset).
We used Logistic Regression for building our models.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

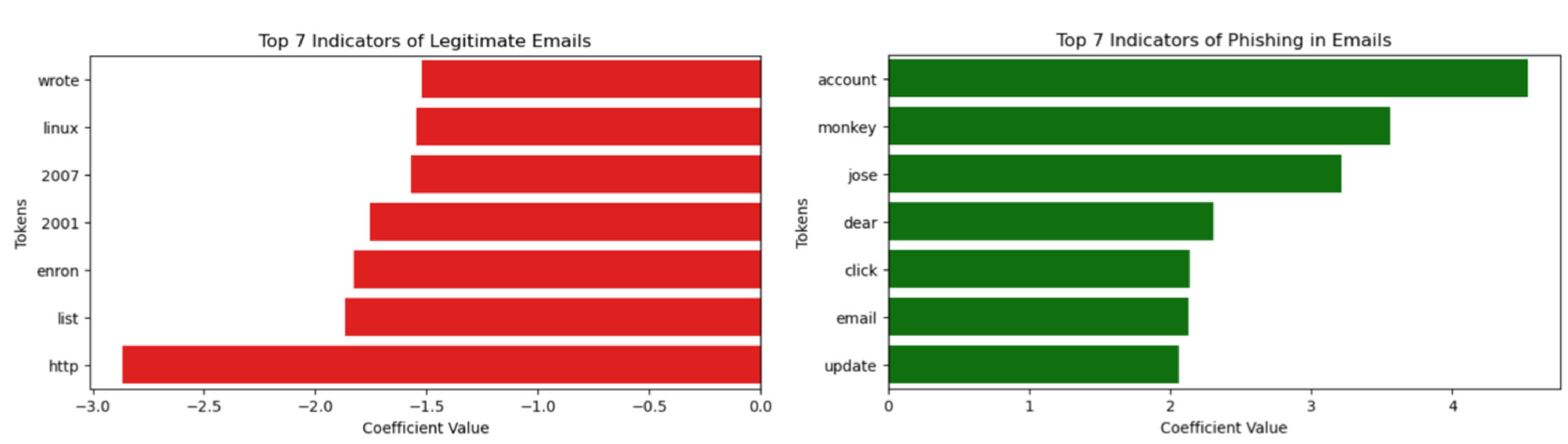
TF-IDF Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Figure 2: Equation for computing TF-IDF

```
TF-IDF Confusion Matrix With Threshold of 0.85
[[273   0]
 [ 54 253]]
Accuracy: 0.906896551724138
Precision: 1.0
Recall: 0.8241042345276873
F1 Score: 0.9035714285714287
```

Figure 3: Confusion Matrix for predicting "PHISH-GT" based on "text"



Add TF-IDF and 'soft' Labels into One Feature

```
TF-IDF With Noisy Labels Confusion Matrix With Probability Threshold 0.85
[[273   0]
 [ 26 281]]
Accuracy: 0.9551724137931035
Precision: 1.0
Recall: 0.9153094462540716
F1 Score: 0.95578231292517
```

Results

The points on Figure 4 indicate where our confusion matrices lie. Combining TF-IDF and the 'soft' labels into one feature resulted in the least amount of trade off between precision and recall, with a 100% precision score and 91.53% recall score. The **combined model** also resulted in the **highest accuracy**: 95.51724137931035%.

Despite good accuracy, precision and recall, the Logistic Regression model lacks interpretability, as its formula is very mathematically heavy.

In conclusion, we were able to develop a classification model that can detect 'phishing' based on email texts as well as the 'soft' labels. The definition of success for our project was to build a model that had a 100% recall score to make the job of cybersecurity analyst easier, to which we achieved.