July 26, 2022

The University of Texas at Austin
McCombs School of Business

Web Page Phishing Detection

A model to protect us from cybercriminals

PRESENTERS:
Aniket Patil, Praneet Kumar Alamuri, Anushka Iyer, Jiaxi Wang, Aakash Talathi, Biyun Yuan
The University of Texas at Austin

Meet Our Team

Aniket Patil

Praneet Kumar Alamuri

Anushka Iyer

Jiaxi Wang

Aakash Talathi

Biyun Yuan

We'll Talk About...

1. What is Phishing? How can we help?
2. The datasets leveraged to identify and resolve the problem
3. ML Models used to identify Phishing Websites
4. Choose the best model and our recommendation to help mitigate Phishing

What is Phishing

- Phishing is a **cybercrime**
- A phishing website is a domain **similar in name and appearance** to an official website
- A link as a legitimate institution to lure individuals into providing sensitive data
- Result in **identity theft and financial loss**
- Not-So-Fun facts:
  - Phishing is involved in **90% breaches**
  - The average cost of a breach is **$3.86M dollars**

Objective

- We want to identify features that will help identify Phishing webpages
- We do not want to **miss any suspicious phishing website**. Therefore, we want False-Negatives to be as low as possible**, thus we are prioritizing *Recall***

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **Eg: A recall** of 0.9 means that 90% of actual phishing websites are **correctly identified** as phishing

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Eg: A precision** of 0.8 means that 80% of all phishing predictions are **actually phishing websites**

Understanding Data and Features..

87 features were extracted for ~11.4k Websites and analyzed to design prediction models

URL Features

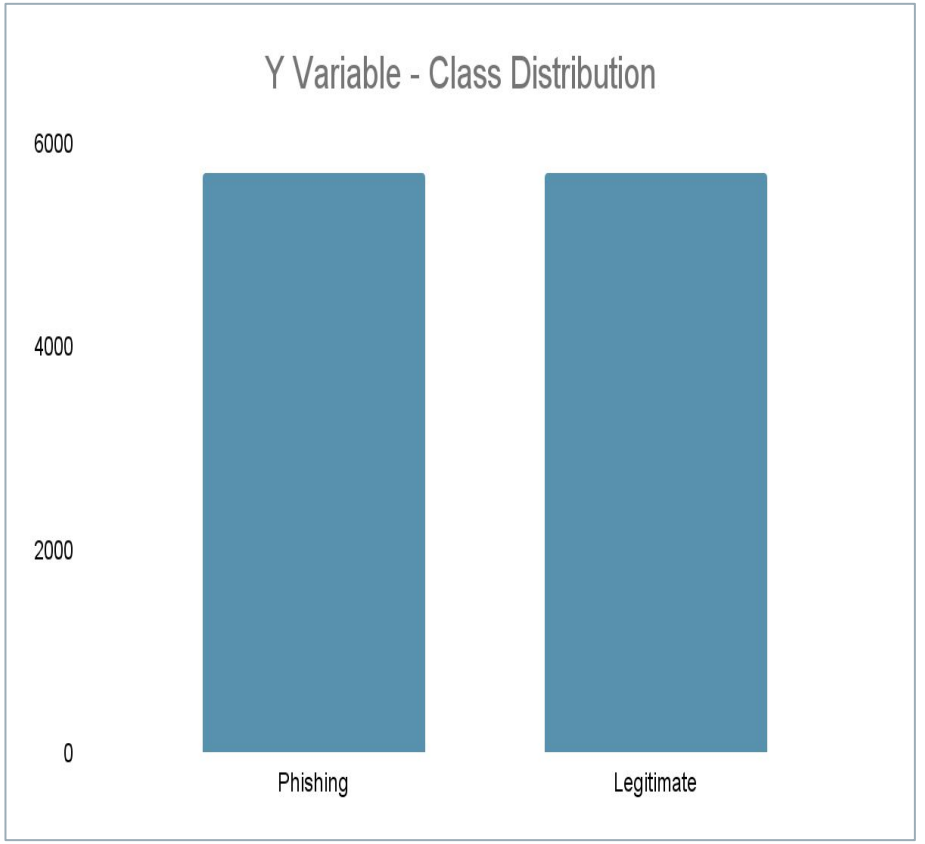
- Features quantifying the contents of the URL
- The features were built using the URL string available for each website
- Eg - URL Length, number of hyphens in URL, etc

Content Features

- Qualitative and quantitative features used to describe the content of webpage
- Eg - Number of hyperlinks, numbers of images, etc

External Features

- Features designed to indicate Webpage relevance with respect to other Webpages in the analysis universe
- Eg - Google Search Rank, DNS registration, etc



EDA Key Steps

Correlation and check for Multicollinearity

Categorical to Numerical
&
Scaling (Min-Max Scaler)

Feature Engineering

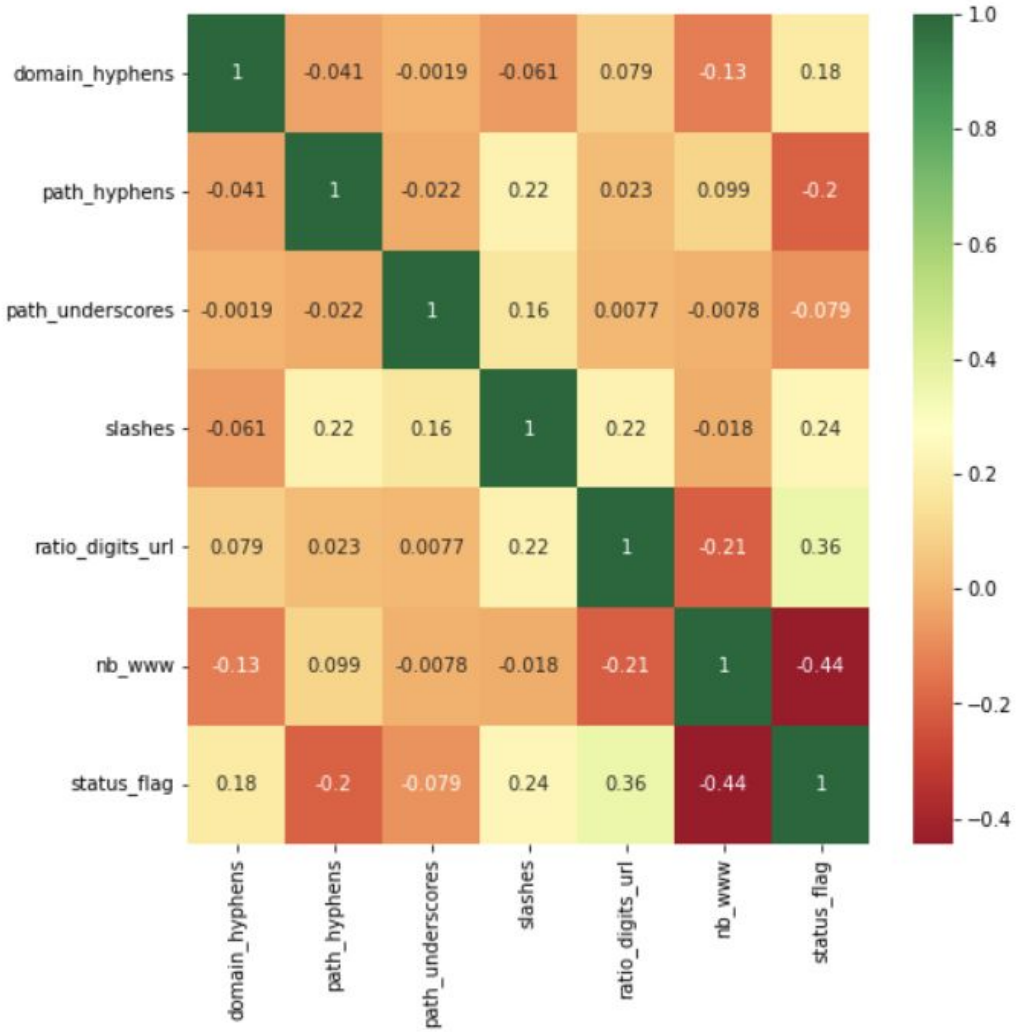
Missing Values

9 Features were finalized to run initial Model iterations

Selecting Features that indicate a strong relationship with Phishing websites

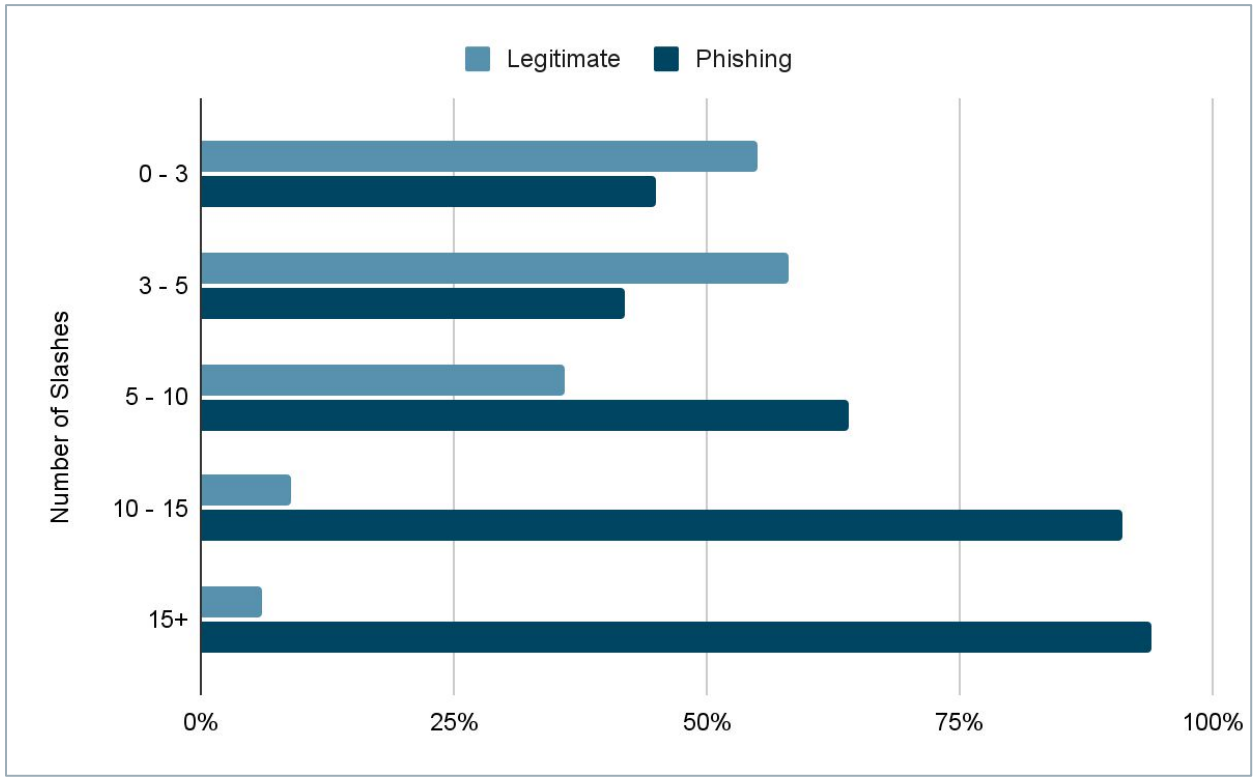
Features based on URL composition

- **Domain Hyphens -** Number of Hyphens in the URL Domain Name
- **Path Hyphens -** Number of Hyphens in the URL Path
- **Path Underscores -** Number of Underscores in the URL Path
- **Slashes -** Number of Slashes in the URL
- **Ratio Digits to URL -** Ratio of Digits to total characters in URL
- **Number of "WWW"s -** Number of "WWW"s used in the URL Name

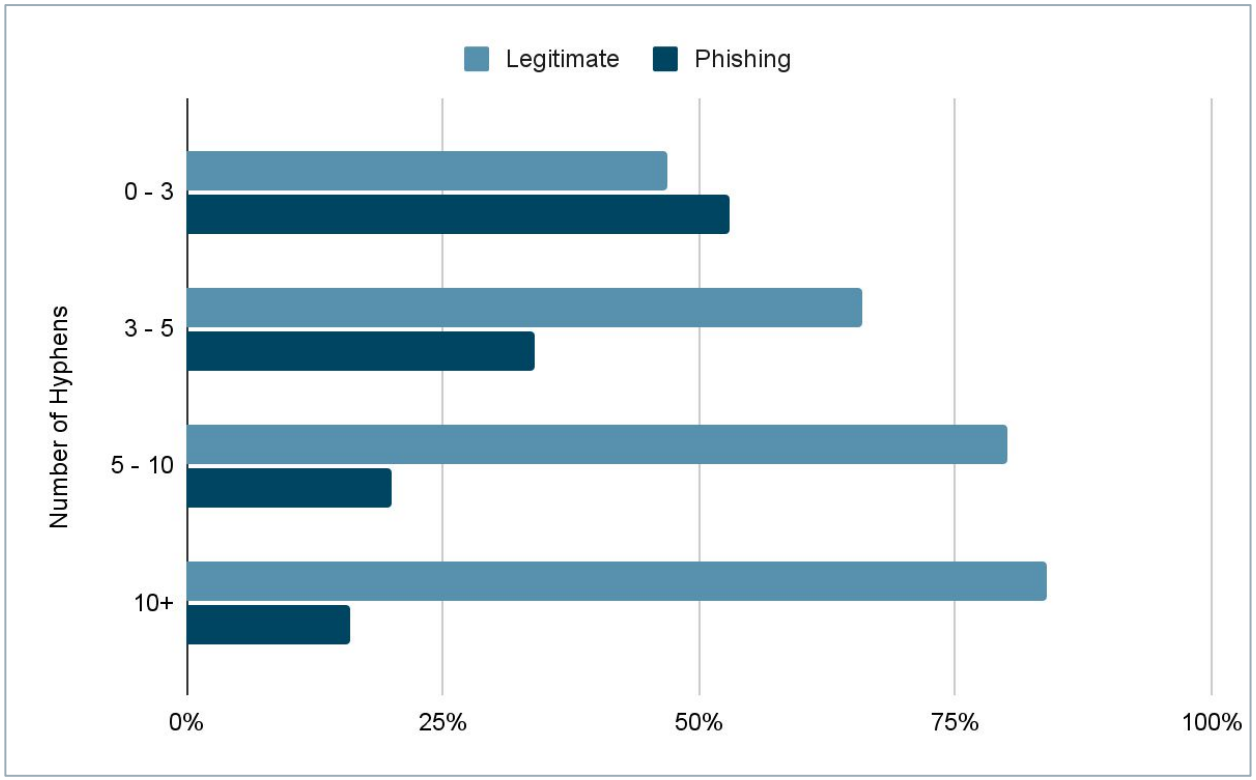


Impact of chosen metrics on Phishing (Y-Variable)

Concentration of Phishing Websites <u>increases</u> in buckets with higher slashes



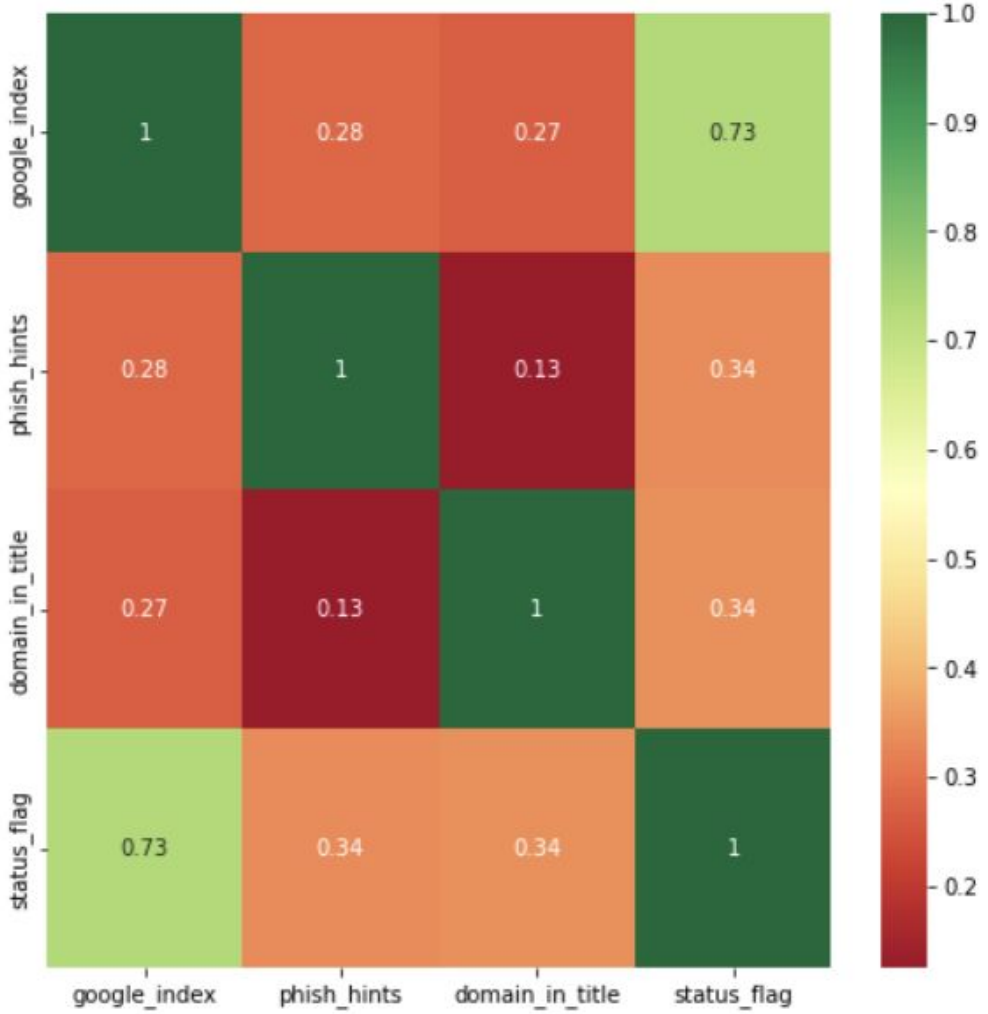
Concentration of Phishing Websites <u>decreases</u> in buckets with higher Hyphens



Selecting Features that indicate a strong relationship with Phishing websites

Features with high correlation with Phishing websites

- **Google Index -** Indicates if the URL is indexed by Google
- **Phishing Hints -** Indicates the presence of common Phishing words in the URL Name
- **Title in Content -** Indicates if the title keywords are also present in the content of the webpage

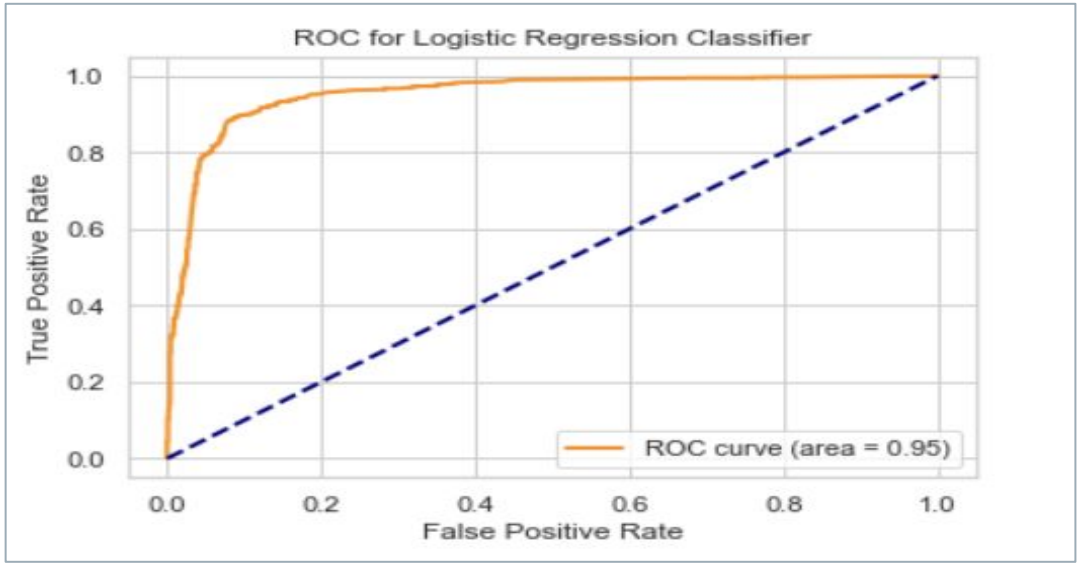


Logistic Regression Classifier (Without RFE)

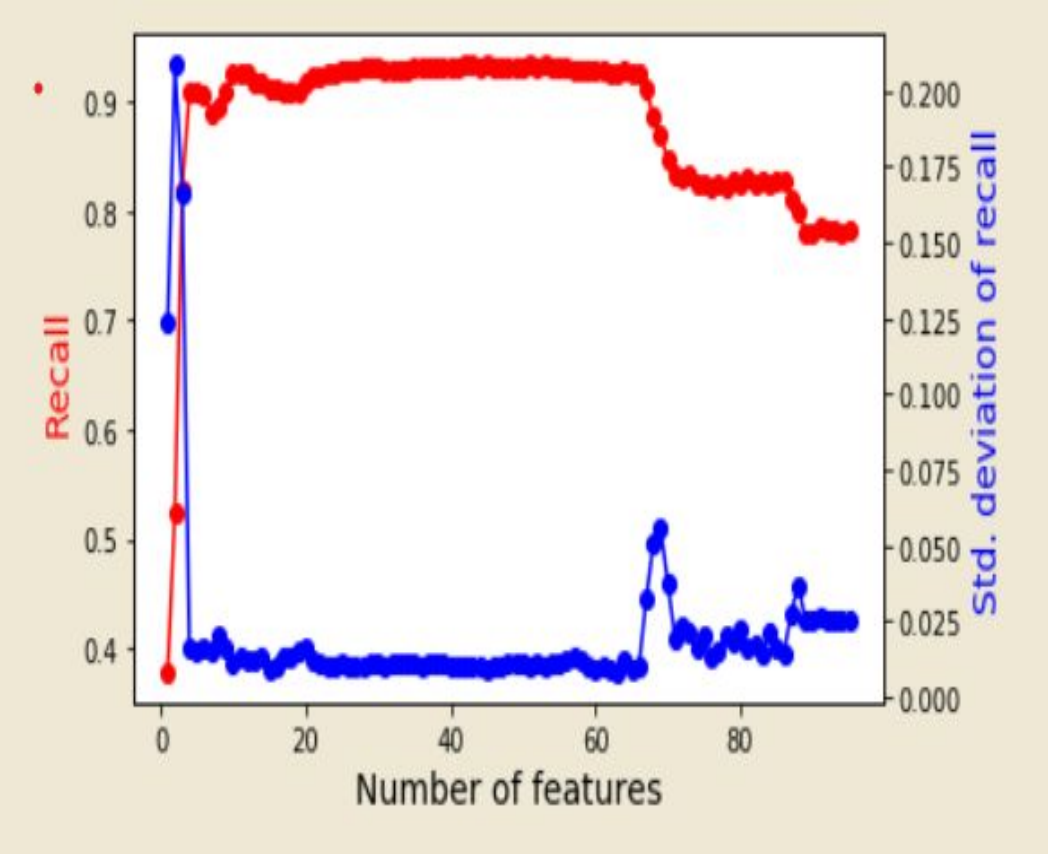
	precision	recall	f1-score
0	0.90	0.89	0.89
1	0.89	0.90	0.90
accuracy			0.90
macro avg	0.90	0.89	0.89
weighted avg	0.90	0.90	0.90

	coef	std err	z
domain_hyphens	0.5227	0.072	7.259
path_hyphens	-0.5841	0.029	-19.916
path_underscores	-0.5589	0.044	-12.561
slashes	-0.1580	0.021	-7.421
google_index	2.9659	0.071	41.681
ratio_digits_url	6.6105	0.473	13.987
phish_hints	1.7591	0.086	20.542
nb_www	-2.1820	0.059	-36.776
domain_in_title	-0.4061	0.056	-7.285

- Multiple Iterations were performed to arrive at the final set of features with |Z-score| > 2
- Cross validation results had low standard deviation of recall
- Ratio of digits to length of URL is a strong predictor



Logistic Regression Classifier (With RFE)

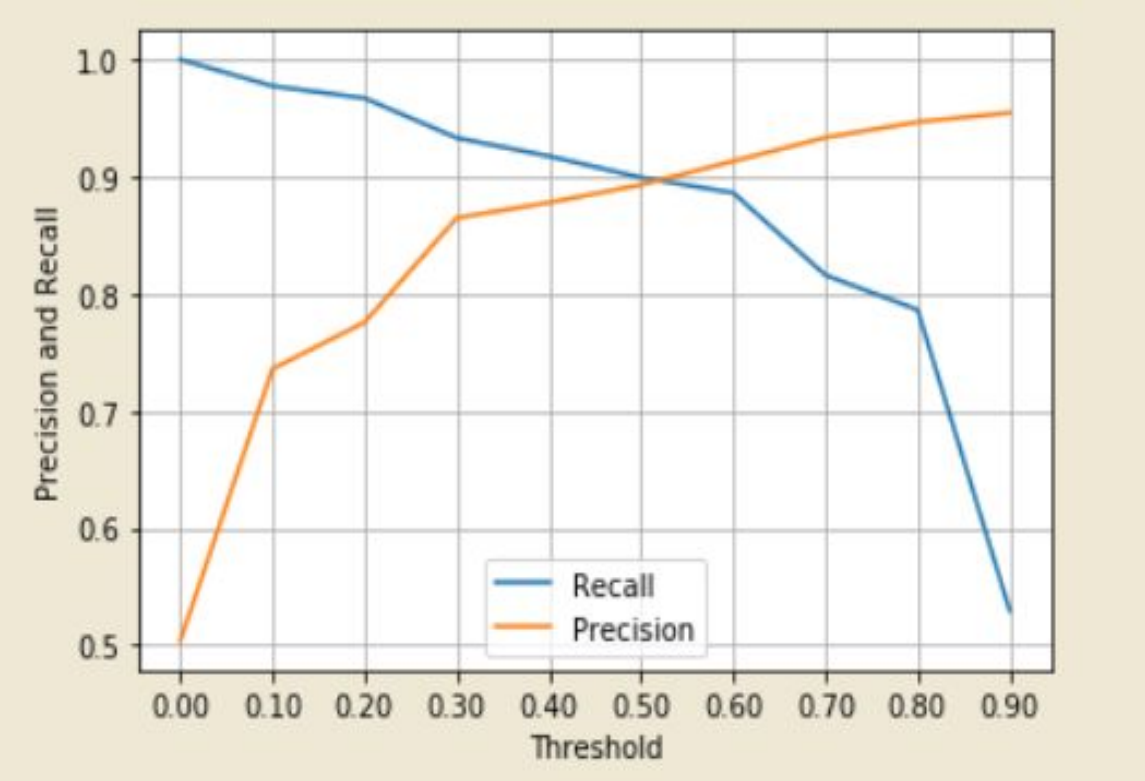


- The RFE with cross-validation algorithm recommends ***50 features*** that generate the highest Recall in the Logistic Regression model
- However, only marginal incremental recall was obtained after adding 10+ features, leading to an over-complicated model

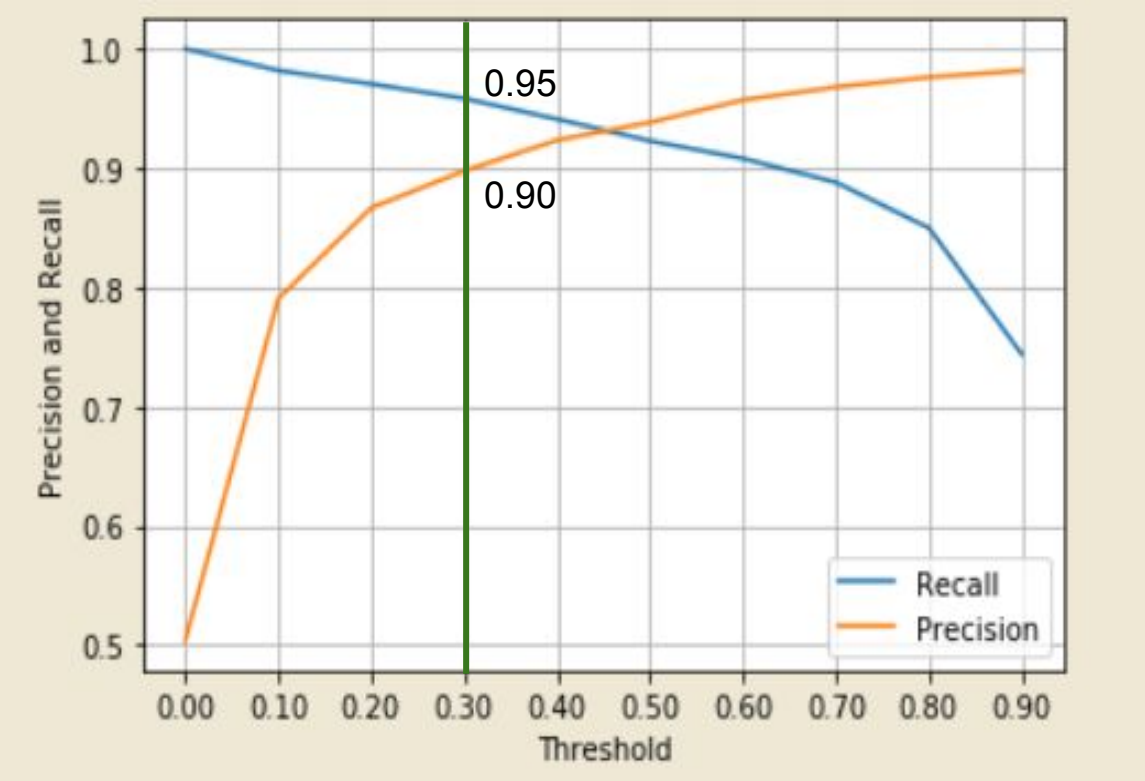
	precision	recall	f1-score
0	0.93	0.93	0.93
1	0.93	0.93	0.93
accuracy			0.93
macro avg	0.93	0.93	0.93
weighted avg	0.93	0.93	0.93

Features recommended by RFE with cross-validation increased the recall by *3%*

Logistic Regression Threshold Analysis



MODEL WITHOUT RFE

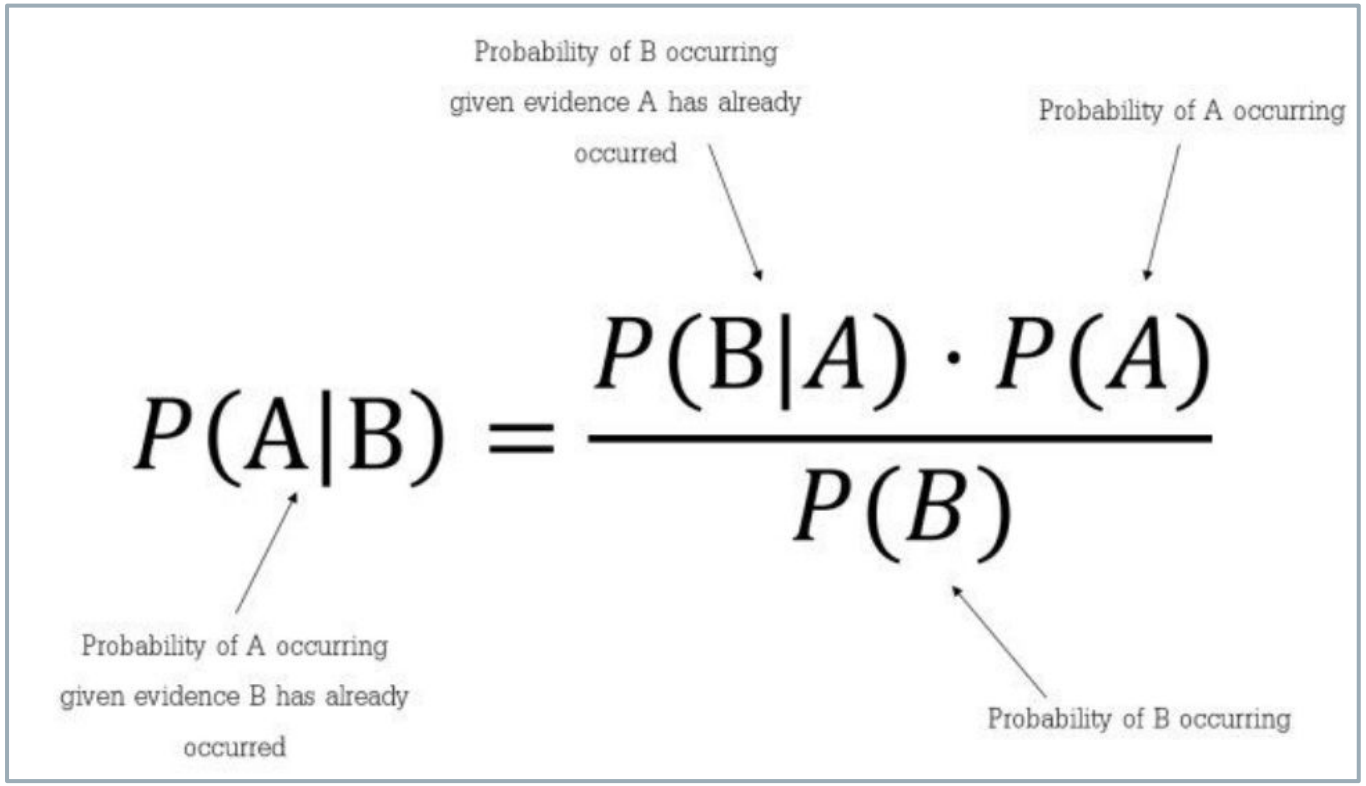


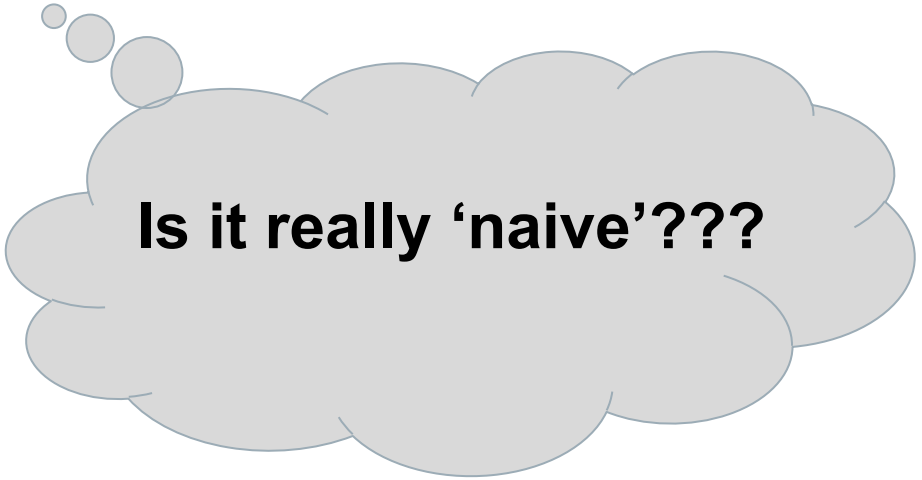
MODEL WITH RFE

We can reduce classification threshold to 0.3 to increase recall while maintaining 90% precision

Naive Bayes Model

Baye's Theorem:





B - predictors or variables
[$X = (x_1, x_2 \ldots . x_n)$]
(length of url, domain hyphens, domain title, etc)

A - Value to be predicted [Y]
(website link is a phishing link or not)

Naive Bayes Model

- Model provides a higher recall of 92%
- Faster processing

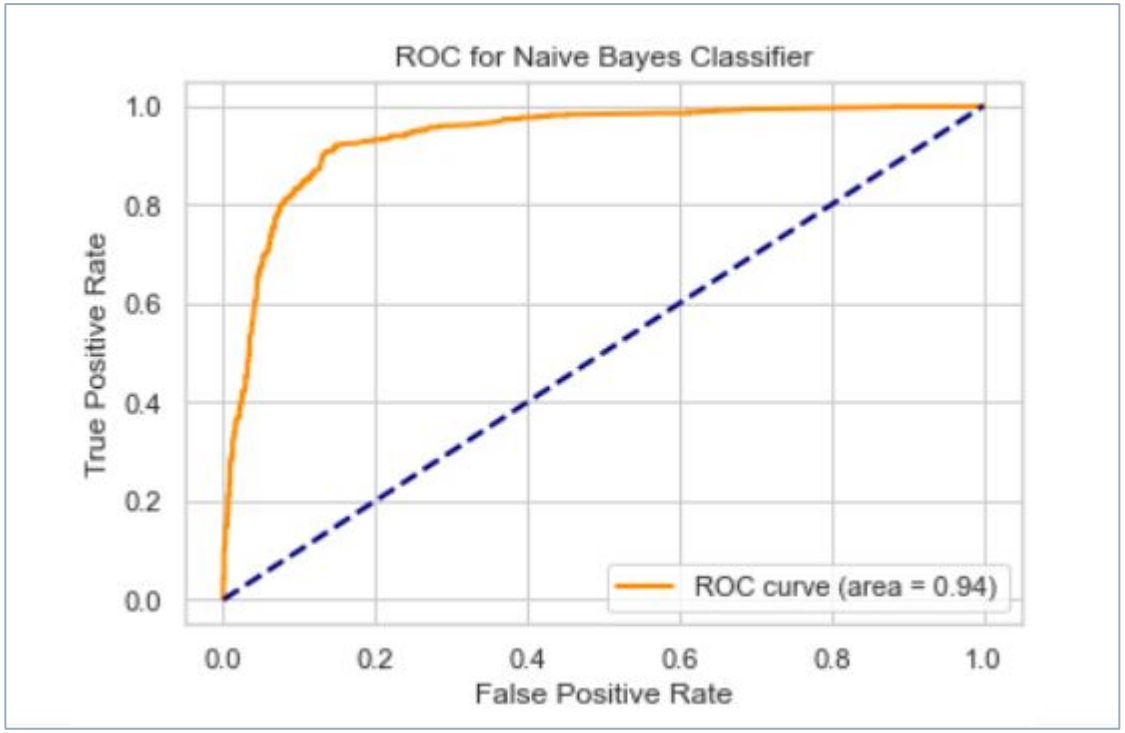
```
Cross-validation on the train dataset:
Mean train dataset's Recall for Naive Bayes Classifier : 0.9243
Mean test dataset's Recall for Naive Bayes Classifier  : 0.9242
```

```
confusion_matrix for train dataset:
 [[3817  764]
 [ 345 4218]]
confusion_matrix for test dataset:
 [[ 953  181]
 [  88 1064]]
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      1134
           1       0.85      0.92      0.89      1152

    accuracy                           0.88      2286
   macro avg       0.89      0.88      0.88      2286
weighted avg       0.88      0.88      0.88      2286
```



Random Forest Parameter Tuning

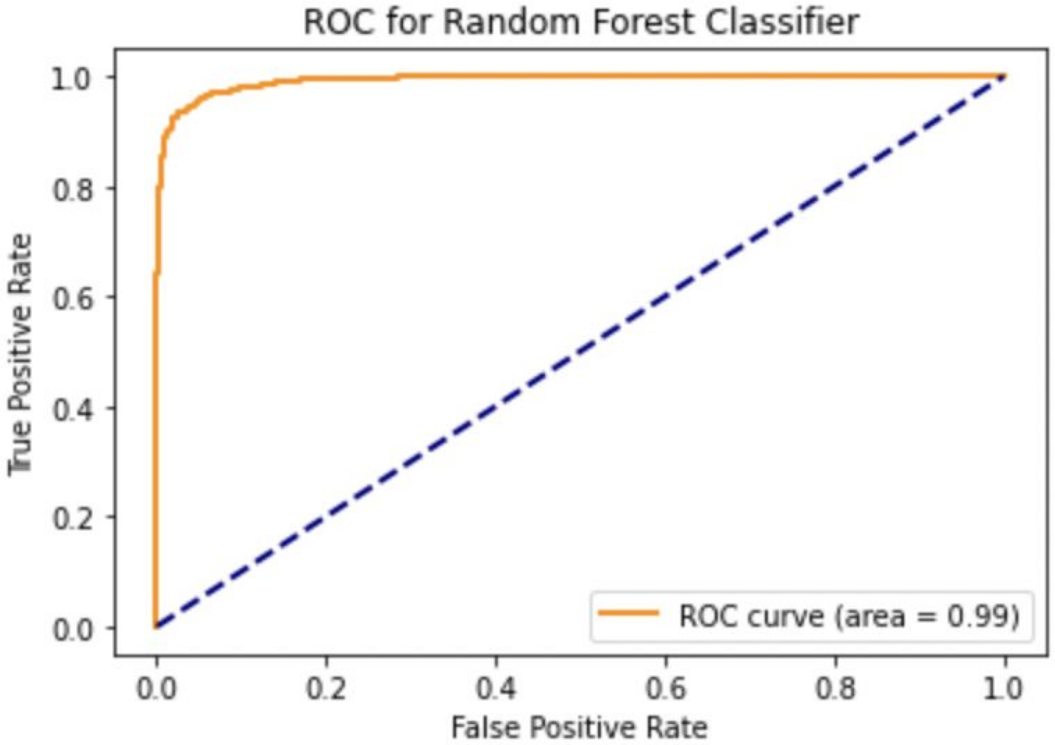
Random Search

Grid Search

```
{'n_estimators': 800,
 'max_features': 'auto',
 'max_depth': 9,
 'bootstrap': True}
```

```
{'bootstrap': True,
 'max_depth': 11,
 'max_features': 'auto',
 'n_estimators': 780}
```

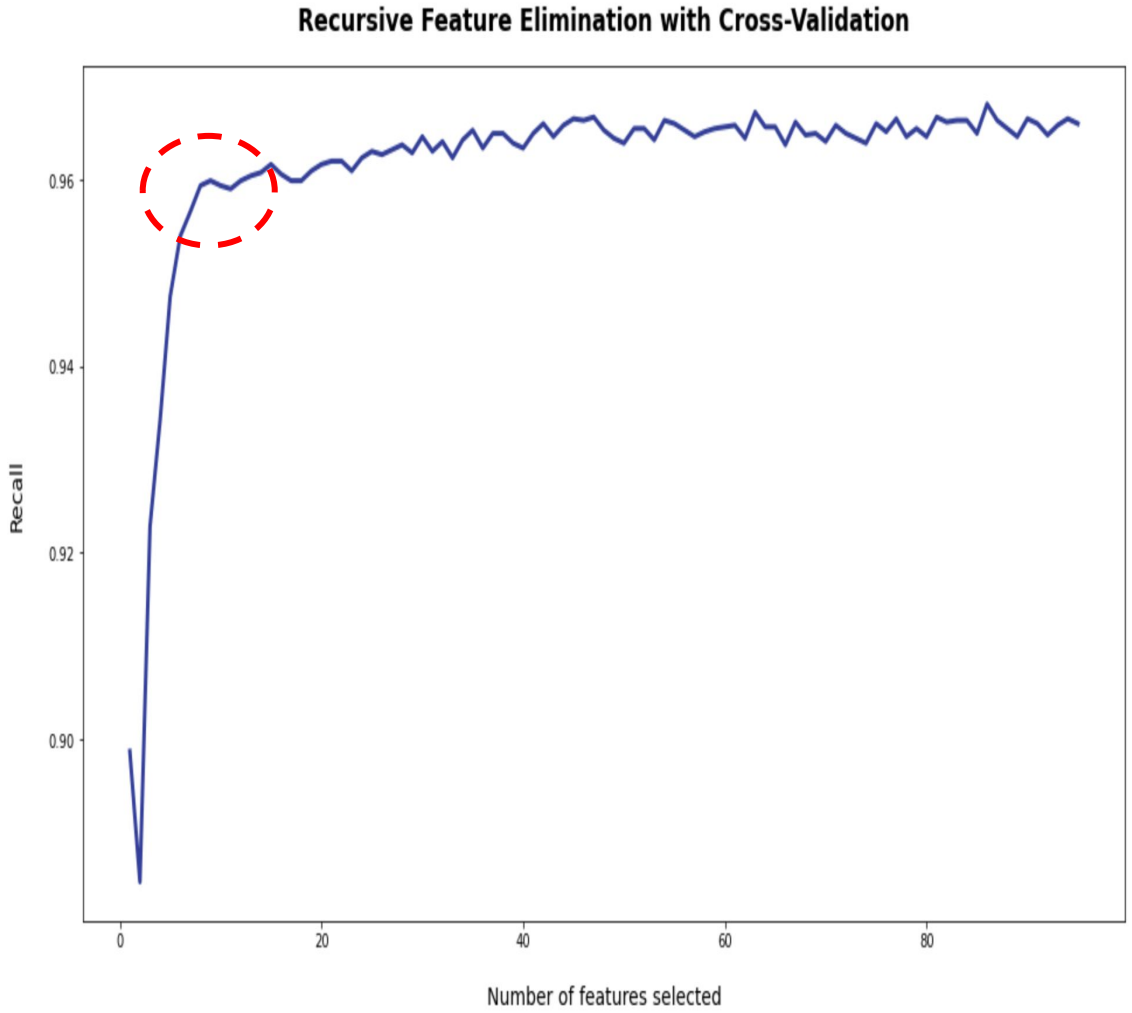
Random Forest Model 1



- Using the original **9 features**, with the optimal hyperparameters, a recall of **94%** was achieved

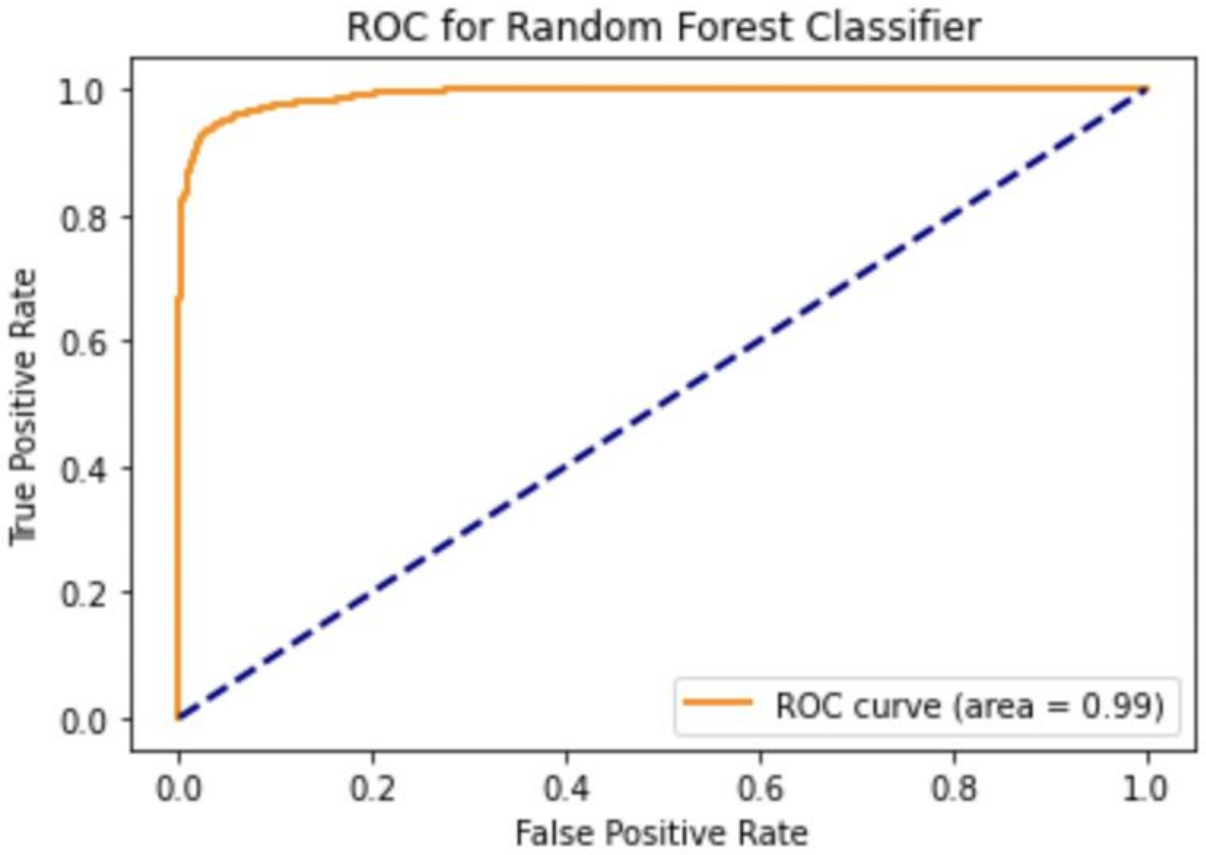
	precision	recall	f1-score	support
0	0.94	0.96	0.95	1134
1	0.96	0.94	0.95	1152
accuracy			0.95	2286
macro avg	0.95	0.95	0.95	2286
weighted avg	0.95	0.95	0.95	2286

Random Forest Feature Elimination



- **86 Features** were selected by the RFE algorithm
- Improvement is marginal after the first **10** features

Random Forest Model 2



- Another model was ran with all **86 recommended features**
- A recall of **97%** was reached using this technique, which was a welcome improvement over the other methods

	precision	recall	f1-score	support
0	0.97	0.98	0.97	1134
1	0.98	0.97	0.97	1152
accuracy			0.97	2286
macro avg	0.97	0.97	0.97	2286
weighted avg	0.97	0.97	0.97	2286

Models' Performance Summary

ML Algorithm	Model	Recall	Precision	F1-score
Naive Bayes	Model 1 - 9 Features	0.92	0.85	0.89
Logistic Regression Classifier	Model 1 - 9 Features	0.90	0.89	0.90
	Model 2 - RFE Recommended Features	0.93	0.93	0.93
Random Forests	Model 1 - 9 Features	0.94	0.96	0.95
	Model 2 - RFE Recommended Features	0.97	0.98	0.97

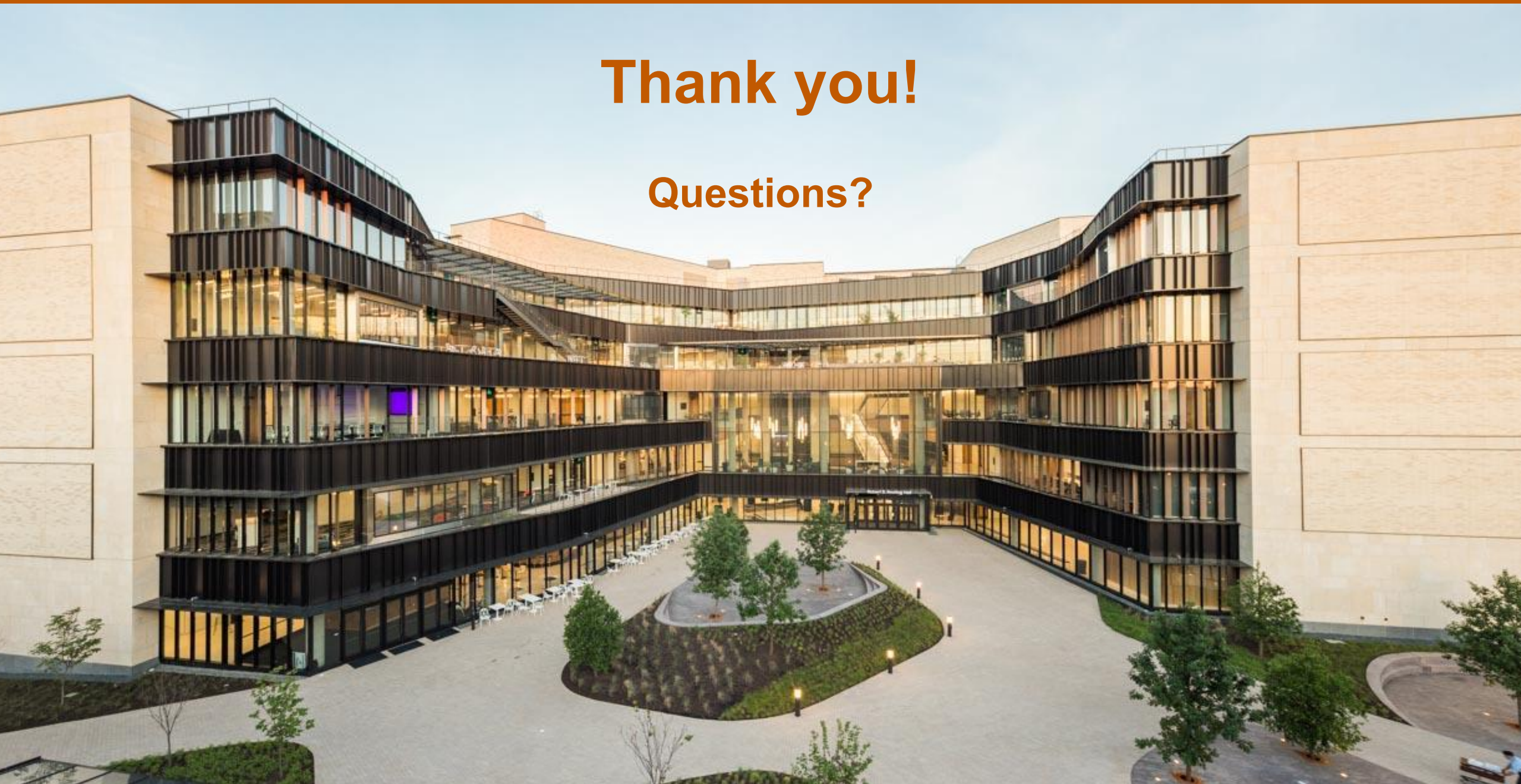
Conclusion & Next Steps

Improved Random Forest model has best performance with highest precision and recall scores (0.98 and 0.97 respectively)

Limitations: Improved Random Forest recommends large number of features, which may be difficult to obtain data for all and trade-off with model complexity

Next steps:

Determine which features are powerful factors and raise awareness	Inform relevant personnel for update in cyber security protection



Thank you!

Questions?