DTONOMY INTERNSHIP

DSA 5900 – PROFESSIONAL PRACTICE

SUMMER 2020

R. KEVIN OBERLAG

DTONOMY, INC.

- Cybersecurity company aiming to deliver meaningful insights of security related threats and alerts.
- Utilize automation and artificial intelligence to enable smarter organization of security alerts and incidents.

DTONOMY'S TECHNOLOGY

- "DTonomy’s technology includes a unique adaptive learning engine which continuously learns, provides contextual insights that are not easily discoverable, and makes relevant recommendations and automated workflows to guide IT teams through steps and procedures, resulting in up to 10 times quicker resolution of incidents, decreased downtime, and reduced alert fatigue for staff. "

EXPANDING USER REACH AND TOOLSET

- Expanding insights by integrating with the G-Mail web and mobile email applications.
- Potential to reach a larger user base and drive new customers to DTonomy's main platform.
- Phish AIR G-Mail add-in application

OBJECTIVES

- Email add-in software development of core features
  - Provide users with easy to use tool for analyzing potentially malicious phishing emails
- Analysis on important features of phishing emails
  - Perform feature selection to determine 2-3 important features that help classify phishing emails
- Visual representation of information
  - Email routing path map for more analyst insights
- Implement a recommendation engine
  - Recommended steps for user when faced with a potentially malicious phishing email

SOFTWARE DEVELOPMENT OF CORE FEATURES AND PUBLICATION

- Ability to extract URLs from an email body and list them to user
- Functionality to scan web pages of URLs for important information on determining maliciousness
- WHOIS web site data for domains
- Publication of initial version

ANALYSIS ON IMPORTANT FEATURES OF PHISHING EMAILS

- Research of phishing datasets
- Use relevant dataset to narrow down most important features with feature selection techniques
- Utilize Python packages to read, manipulate, visualize and analyze data
- Create report on analysis of top features from dataset

VISUAL REPRESENTATION OF INFORMATION

- Create a map visualization for the add-in
- Apply geolocation data of email server routing paths

IMPLEMENT A RECOMMENDATION ENGINE

- Provide recommendations to users for handling potentially malicious emails
- Recommendations are static as of now
- Analysis on phishing emails in the second objective may be used in future for more dynamic recommendations

EMAIL ADD-IN SOFTWARE DEVELOPMENT TECHNIQUES

- Scrum project management framework
  - Useful for small teams
  - Break up development into small goals in short time frames (sprints)
  - Two weekly 15-30-minute progress meetings on Mondays and Wednesdays
    - Helps keep team apprised of other members' work and provides allotted time for exchange of questions and ideas
  - One, 2-hour demo meeting on Fridays
    - Opportunity to present and receive feedback
  - Iterative process taken for Phish AIR development

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - DISCUSSIONS

- Initial discussions
  - Desired core features
  - Applications with similar functionality
  - Recommended APIs to integrate

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - APPS SCRIPT

- Research and learn API of Apps Script
- Subset of JavaScript; Some features not accessible
  - Continuity of design requires absence of client-side JavaScript and HTML
- UI created with pre-made components via Google's Card Service
  - Makes development easier but can be limiting
  - No asynchronous function calls
  - Less than desirable UX

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - URL EXTRACTION

- Write code for extraction URLs from a message body
  - Research regex and HTML methods
    - Due to possible complexity of embedded URLS, regex is deemed to risky for edge cases
  - Use HTML text with HTML parser to parse elements
    - <a> and <area> tags determined to be the relevant HTML elements

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - URLSCAN.IO

- Review urlscan.io documentation
- Integrate with methods exposed by Apps Script API for making HTTP requests
  - Setup RESTful API request to send each requested hyperlink to be scanned
  - Returns majority of data points determined to be essential to core functionality
    - Verdict on maliciousness of the web page
    - IP address of the web page host server
    - Country name of the web page host server
    - Domain of the effective web page
    - Screenshot URL of the effective web page
  - Cache data so subsequent requests do not make unnecessary calls to API
    - Helps avoid reaching rate limit
    - Make application more efficient and responsive

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - WHOIS.COM

- Use domain information retrieved from urlscan.io to create link

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - PUBLISH VERSION 1.0

- Not an easy process
  - Convoluted and confusing documentation
- Convert project to be hosted on Google Cloud Platform
  - Met with supervisor to obtain access and permissions
- Two-step approval process from Google
  - Security verification
    - YouTube video detailing various aspect of the app to help with security assessment
  - Brand verification
    - Ensure that the applications meets design specifications and guidelines

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - PUBLISH VERSION 1.0 (CONT.)

- Application description page link
- Drafting of privacy policy with link
- Drafting of Terms of Service agreement with link
- Work closely with graphic designer intern to help with branding and layout

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - MAP VISUALIZATION

- Develop email message header parser
  - Parse header section of the raw email content into JavaScript data structures
  - Setup "Received" header as an array object to more easily parse IP addresses
- Use regular expressions to parse IP4 and IP6 addresses
- Use IP addresses for ipgeolocationapi.com geolocate API
  - Pass IP address as parameter
  - Returns information on the country and coordinates

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - MAP VISUALIZATION (CONT.)

- Use coordinates in Google's Static Map API
- Accessed via Maps class and its methods, exposed via Apps Script
    - Coordinates can be used to create markers, paths, and bounds of visibility

EMAIL ADD-IN SOFTWARE DEVELOPMENT PROCEDURES - RECOMMENDATION STEPS

- Present static list of steps to user
  - Block IP of malicious URL
  - Contact effected receivers of email and notify of malicious nature
  - Evaluate impact of any compromised users
  - Contact DTonomy for more recommendations or security incident response solutions
- Eventually recommendations will be dynamic

TECHNIQUES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS

- Performed on UCI Machine Learning Repository dataset
  - Contained 30 engineered features related to phishing
- Create feature rankings to determine subset of features which provided the most importance to the predictive model
  - Random Forest Classification
  - Stratified K-Fold Cross-Validation
  - Recursive Feature Elimination

TECHNIQUES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - RANDOM FOREST CLASSIFICATION

- Used due to how the mechanics of Random Forests naturally rank by how well they improve the purity of a given node
  - Nodes with highest information gain occur at the top of the trees
  - Nodes with least information gain occur at the bottom of the trees
  - By pruning trees below a particular node, a subset of the most important features can be obtained

TECHNIQUES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - CROSS-VALIDATION

- Used for evaluating a machine learning model
    - Model is trained using a subset of data and evaluated on complementary subset
- K-Fold cross-validation partitions into k disjoint subsets of approximately equal size
    - Useful for limited data samples
- Stratified k-fold is important when there is a risk of sampling bias
    - Seeks to ensure that each fold is representative of all strata of the data.
    - Used since the dataset is slightly biased toward "Legitimate" results.

TECHNIQUES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - RECURSIVE FEATURE ELIMINATION

- Fits a model and removes the weakest feature(s)
- Used in combination with Random Forest classification model, and stratified k-fold cross-validation for determination of important features

PROCEDURES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS

- Utilize Jupyter notebooks and Python packages
    - Pandas
    - Sci-kit learn
    - Yellowbrick
    - Matplotlib

PROCEDURES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - PANDAS

- Used for reading data from CSV into data frame and manipulating
  - View attributes
  - Observe shape to understand number of observations and total attributes
  - Apply value_counts method to see values of target label
    - 4898 values of –1 (Phishy)
    - 6157 values of 1 (Legitimate)
    - Shows slight misbalance; Reason for Stratified k-fold cross-validation
    - Values of –1 (Phishy), 0 (Suspicious), and 1 (Legitimate) found for features
- Data manipulation
  - Two sets of analysis (Email related subset, all features including web page related)

PROCEDURES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - SCI-KIT LEARN

- Create 10-fold stratified cross-validation object
- Create Random Forest classifier

PROCEDURES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - YELLOWBRICK

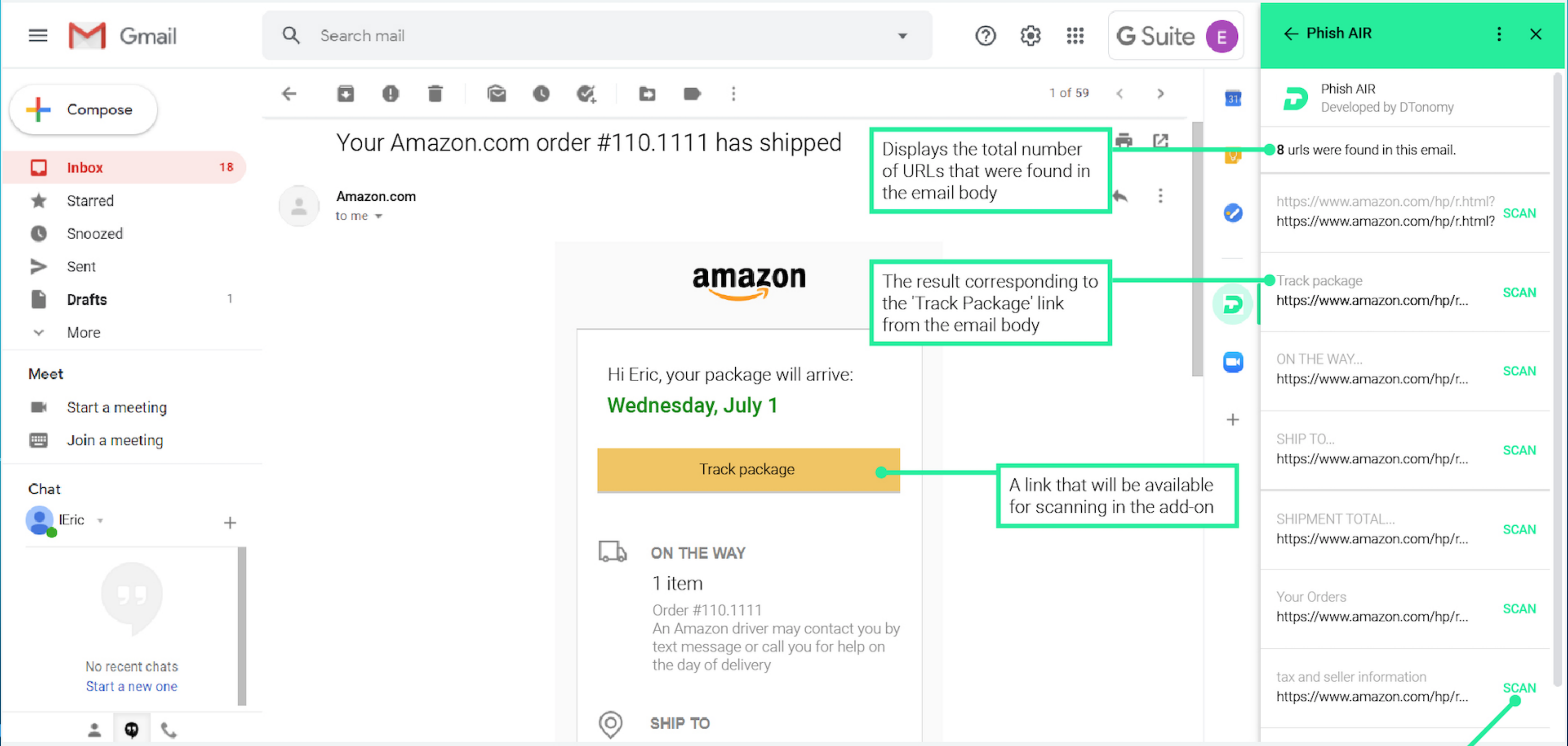
- Use Cross-validation object and Random Forest classifier in RFECV method
- Recursive Feature Elimination with Cross Validation method creates a visualizer for recursive feature elimination
  - Produces visualization of the scores achieved with varying number of features
  - Produces optimal number of features and the score
- False negative and False Positives are crucial for phishing classification
  - Use F1 scoring measure to reduce chance of incorrectly predicting that legitimate observations are phishing, or phishing observations are legitimate

PROCEDURES FOR ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS - MATPLOTLIB

- Use feature importance data returned from visualizer object to obtain top 3 significant features and visualize them as bar plots

EMAIL ADD-IN APPLICATION RESULTS

- There isn't any analysis related to the development of the add-in, so I will walk through the details of the product



Your Amazon.com order #110.1111 has shipped

Amazon.com
to me

amazon

Hi Eric, your package will arrive:

Wednesday, July 1

Track package

ON THE WAY

1 item

Order #110.1111
An Amazon driver may contact you by text message or call you for help on the day of delivery

SHIP TO

Phish AIR
Developed by DTonomy

8 urls were found in this email.

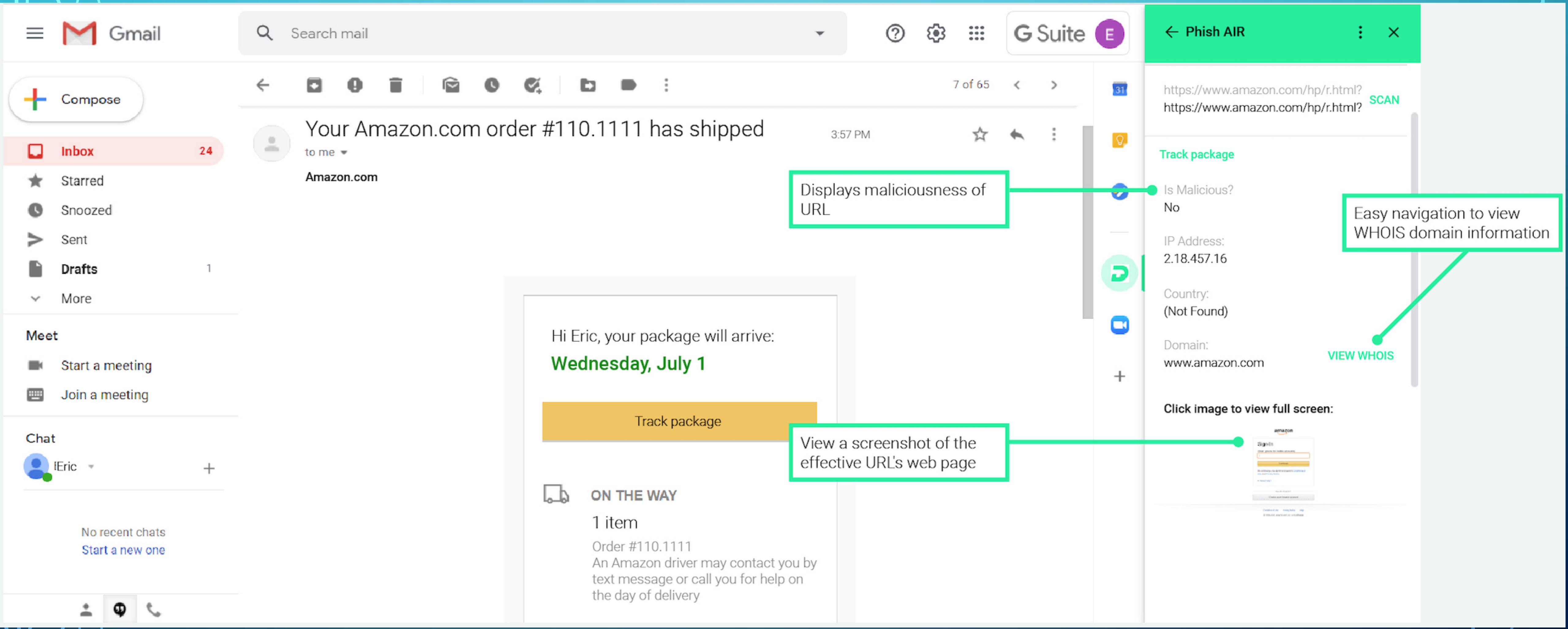
https://www.amazon.com/hp/r.html? https://www.amazon.com/hp/r.html?	SCAN
Track package https://www.amazon.com/hp/r...	SCAN
ON THE WAY... https://www.amazon.com/hp/r...	SCAN
SHIP TO... https://www.amazon.com/hp/r...	SCAN
SHIPMENT TOTAL... https://www.amazon.com/hp/r...	SCAN
Your Orders https://www.amazon.com/hp/r...	SCAN
tax and seller information https://www.amazon.com/hp/r...	SCAN

Displays the total number of URLs that were found in the email body

The result corresponding to the 'Track Package' link from the email body

A link that will be available for scanning in the add-on

Click to analyze URL for maliciousness and view a screen shot



Your Amazon.com order #110.1111 has shipped

3:57 PM

to me

Amazon.com

Hi Eric, your package will arrive:

Wednesday, July 1

Track package

ON THE WAY

1 item

Order #110.1111
An Amazon driver may contact you by text message or call you for help on the day of delivery

https://www.amazon.com/hp/r.html?
https://www.amazon.com/hp/r.html?

Track package

Is Malicious?
No

IP Address:
2.18.457.16

Country:
(Not Found)

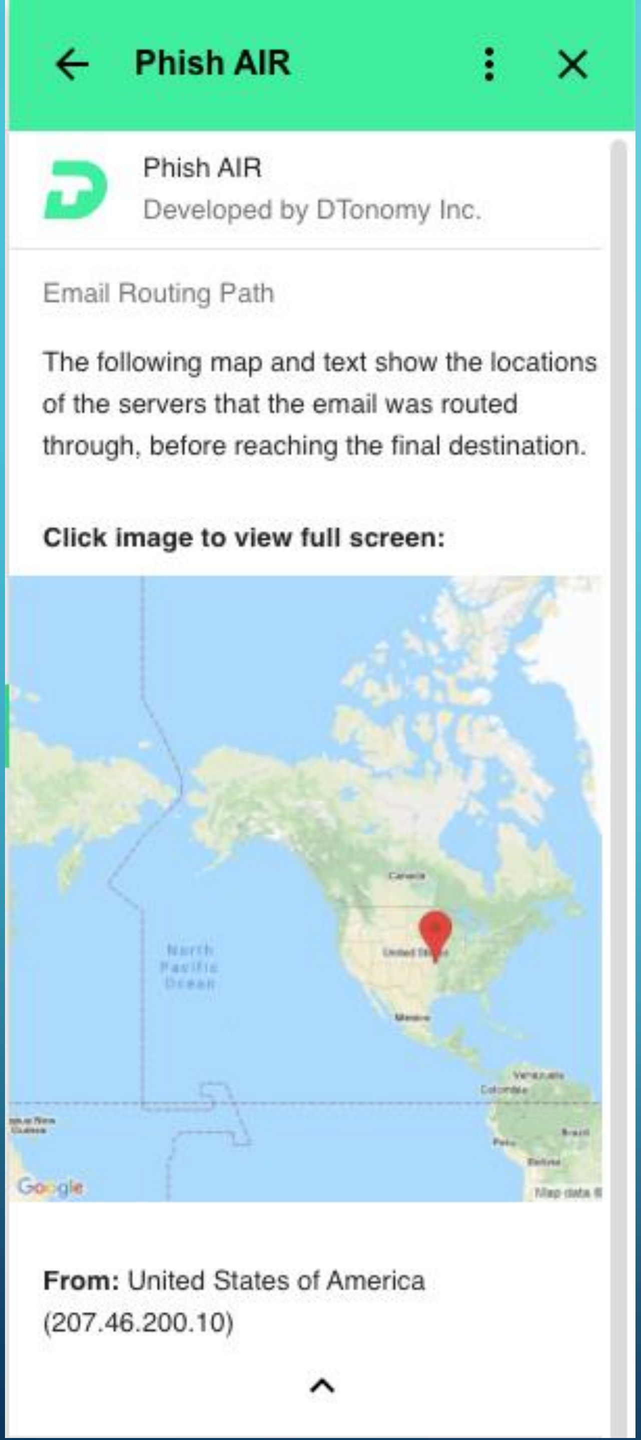
Domain:
www.amazon.com

Click image to view full screen:

Displays maliciousness of URL

Easy navigation to view WHOIS domain information

View a screenshot of the effective URL's web page



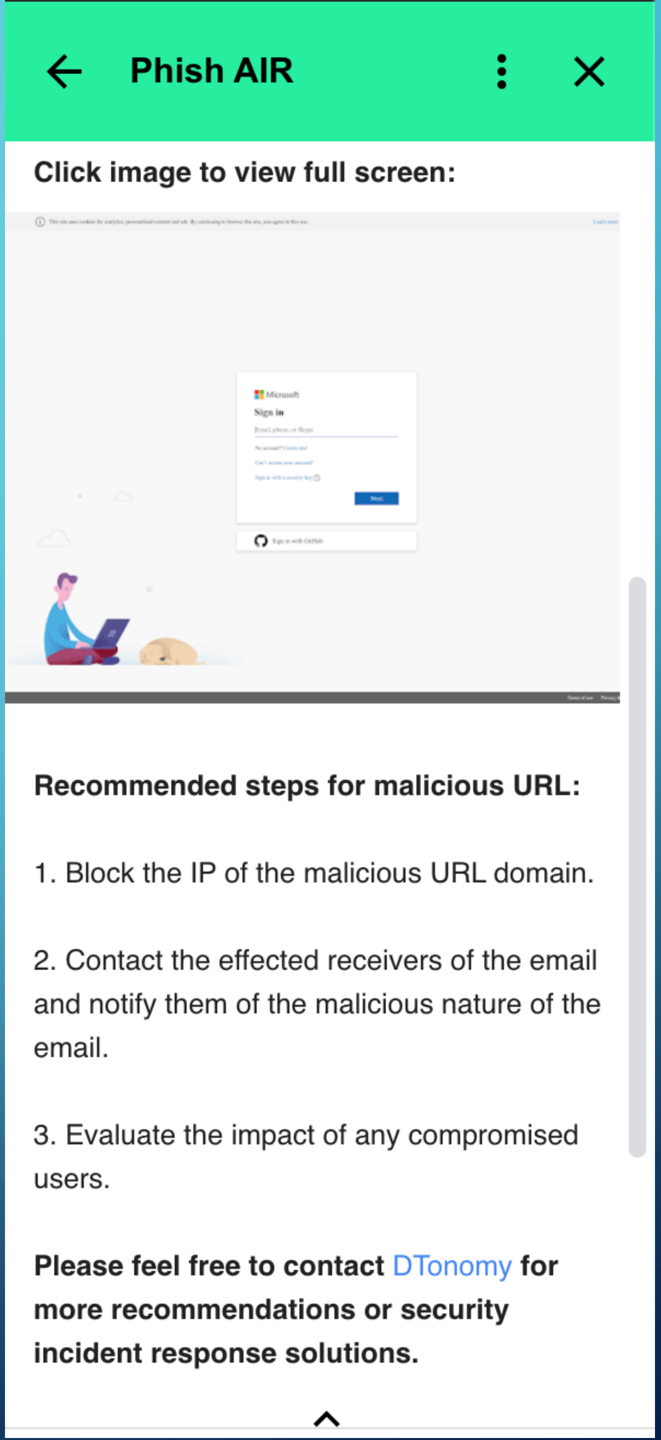
Phish AIR
Developed by DTonomy Inc.

Email Routing Path

The following map and text show the locations of the servers that the email was routed through, before reaching the final destination.

Click image to view full screen:

From: United States of America (207.46.200.10)



Click image to view full screen:

Recommended steps for malicious URL:

1. Block the IP of the malicious URL domain.

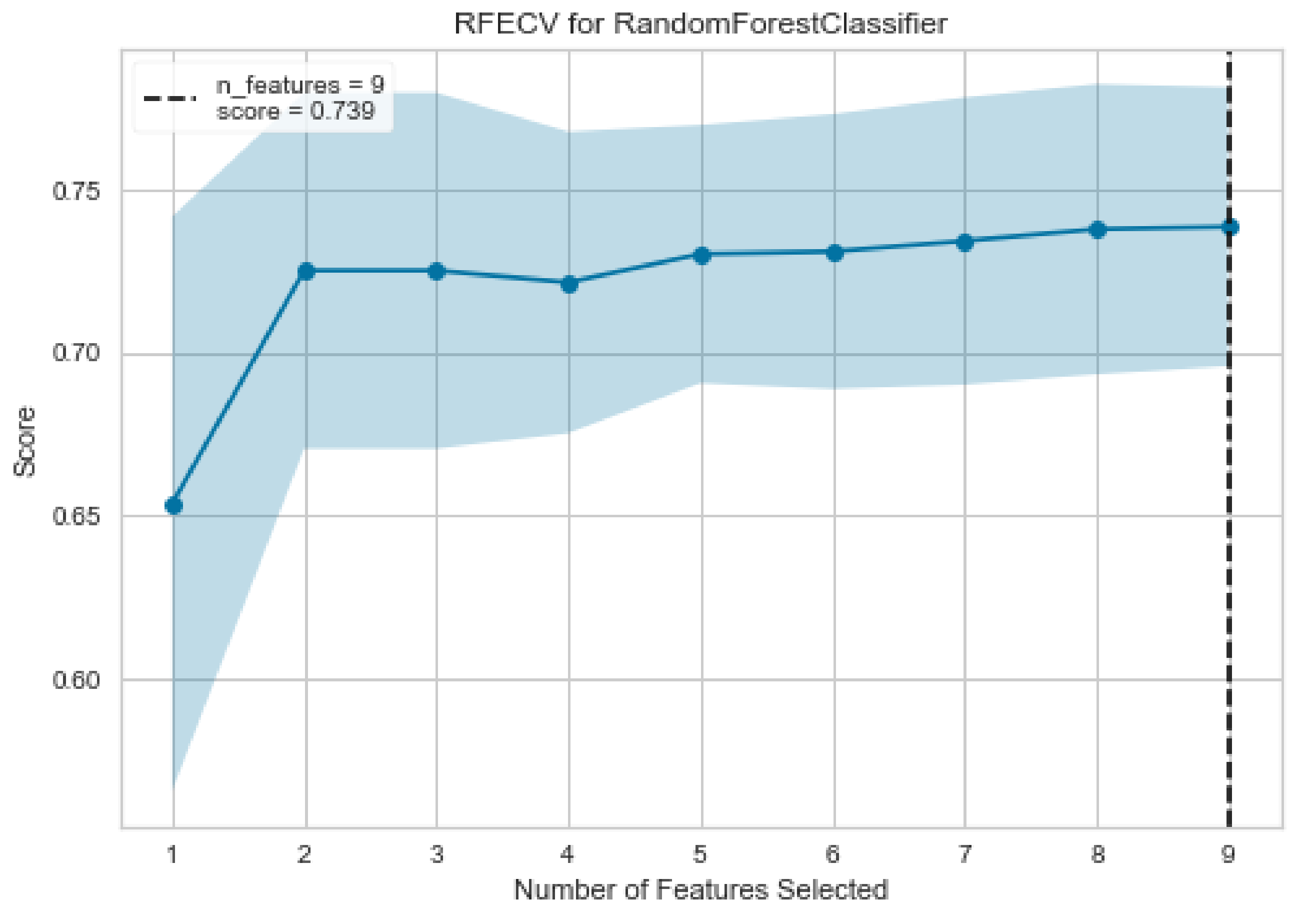
2. Contact the effected receivers of the email and notify them of the malicious nature of the email.

3. Evaluate the impact of any compromised users.

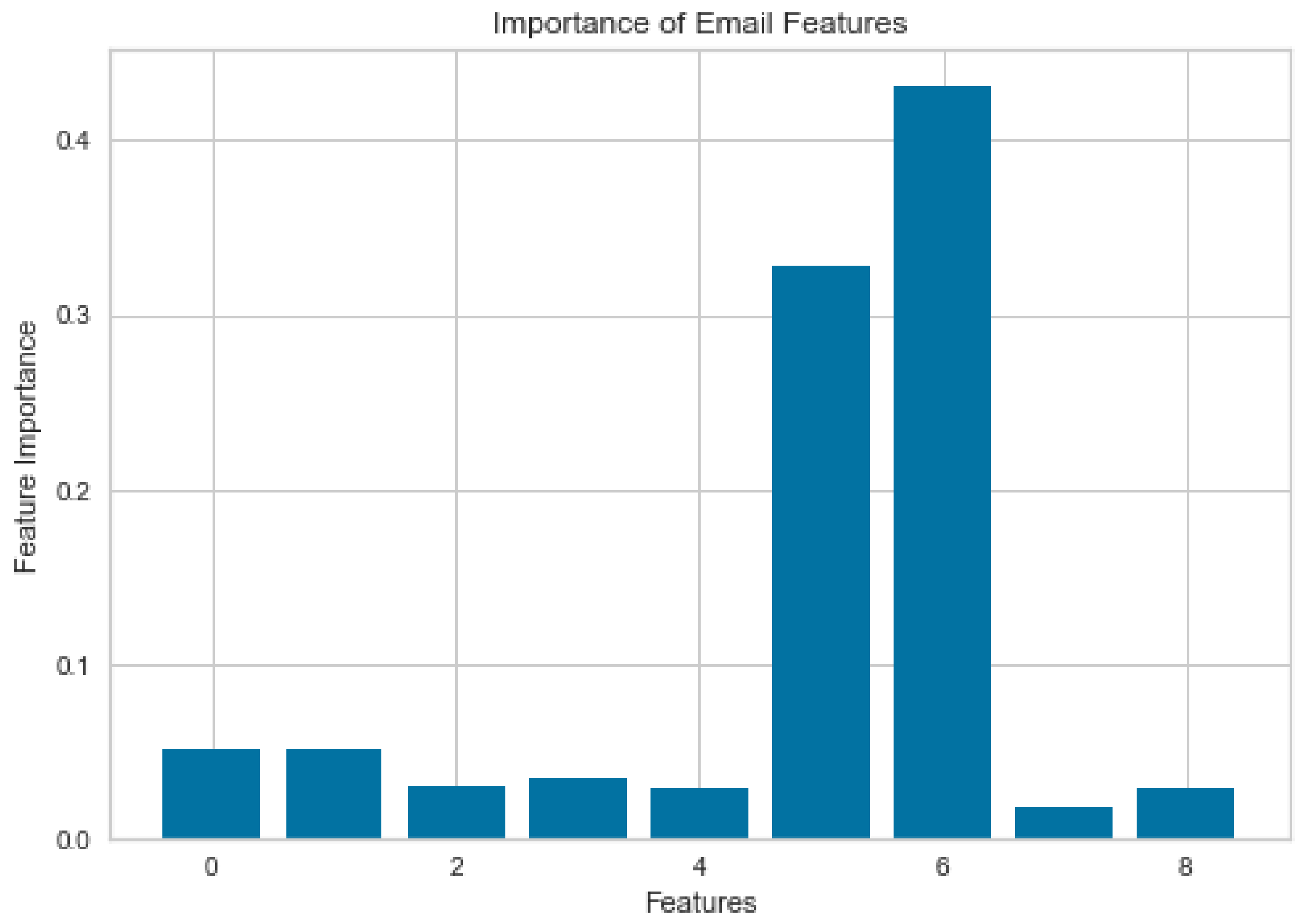
Please feel free to contact DTonomy **for more recommendations or security incident response solutions.**

ANALYSIS OF IMPORTANT FEATURES OF PHISHING EMAILS

RECURSIVE FEATURE ELIMINATION

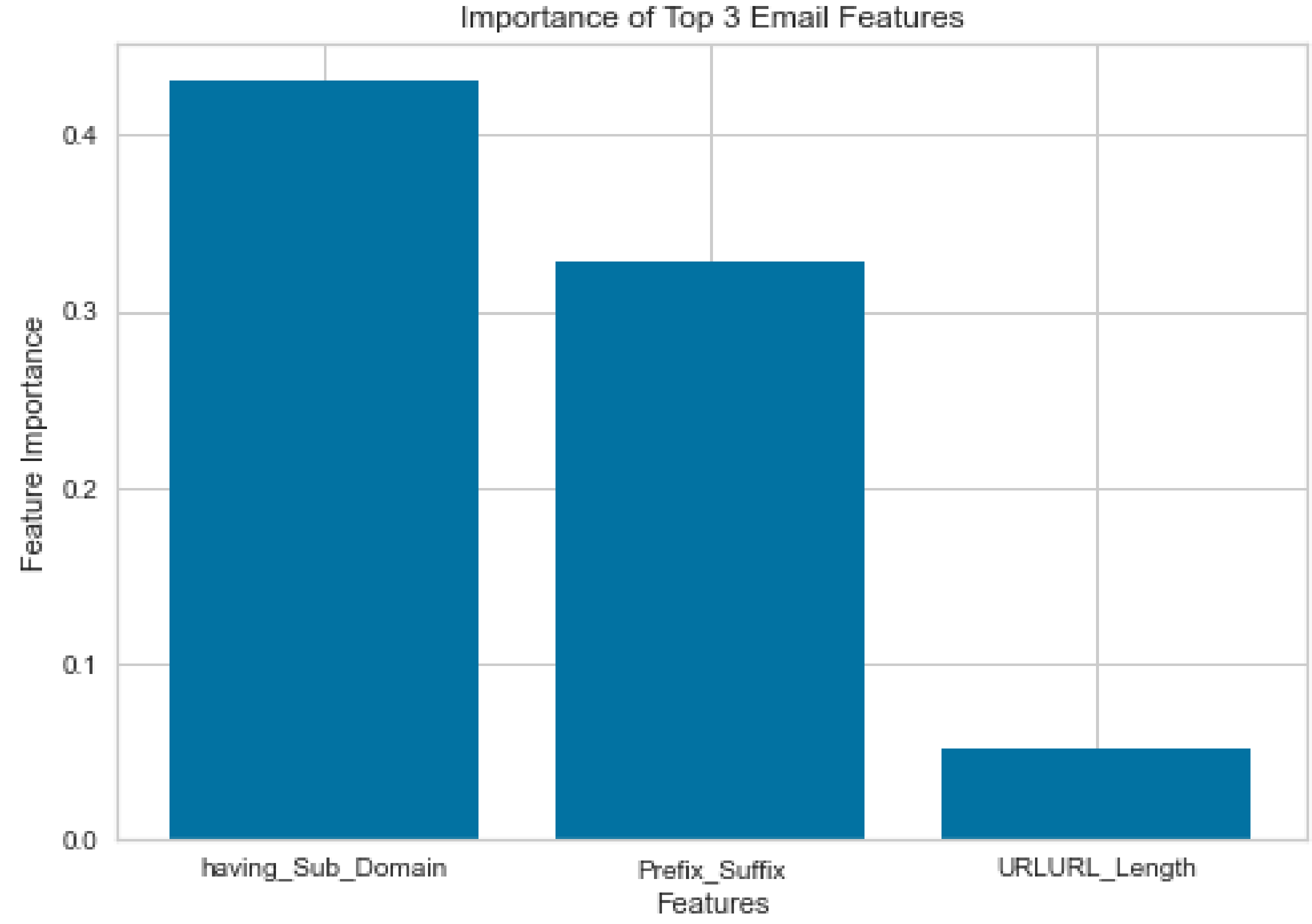


- Fit with Random Forest classification model
- Fit with subset of email related features
- Consists of 9 of original 30 features
- All 9 are used for optimal RFE cross-validation score of 0.739
- We see 2-3 provide near optimal score



INDIVIDUAL FEATURE IMPORTANCE SCORES

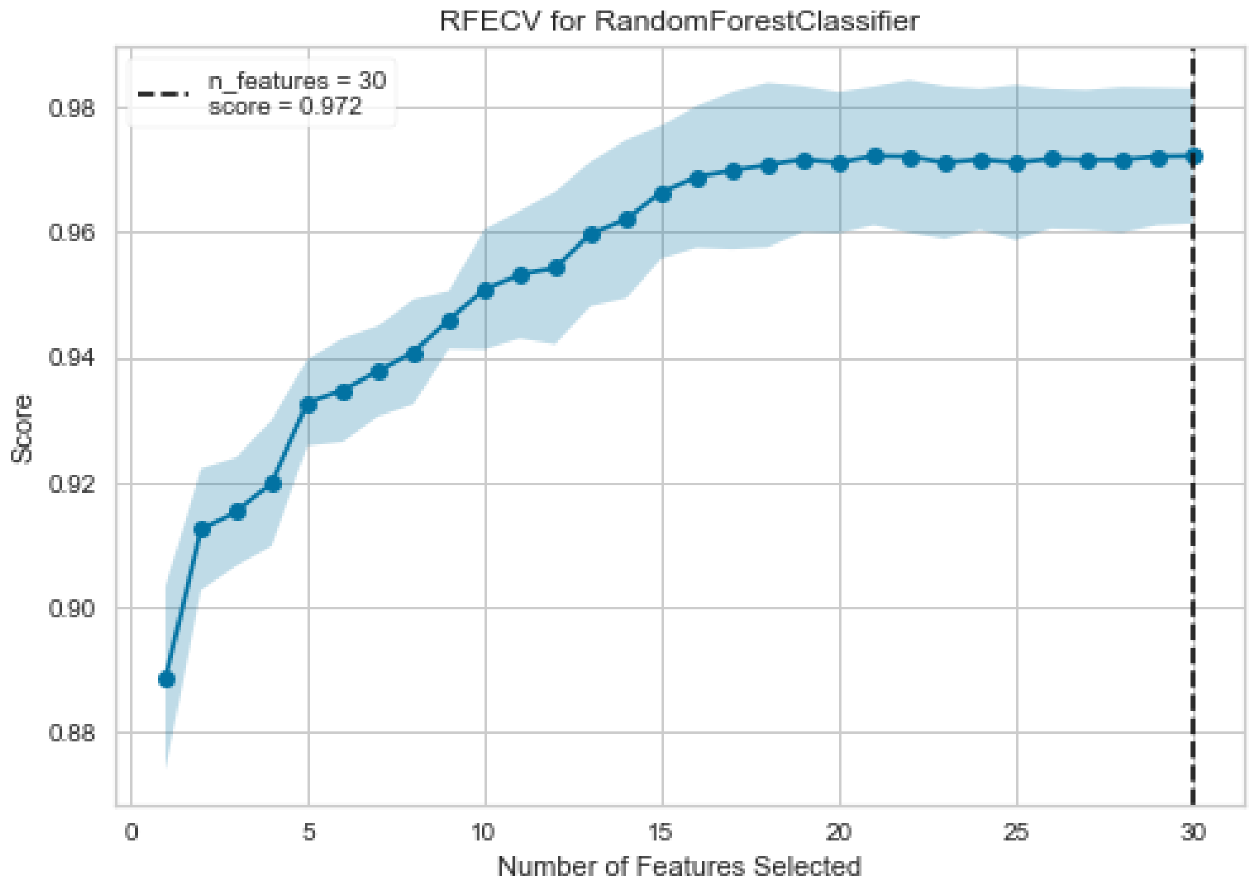
- Feature importance scores for each of the 9 email related features
- We can see that two features clearly standout
- These features provide the most information gain



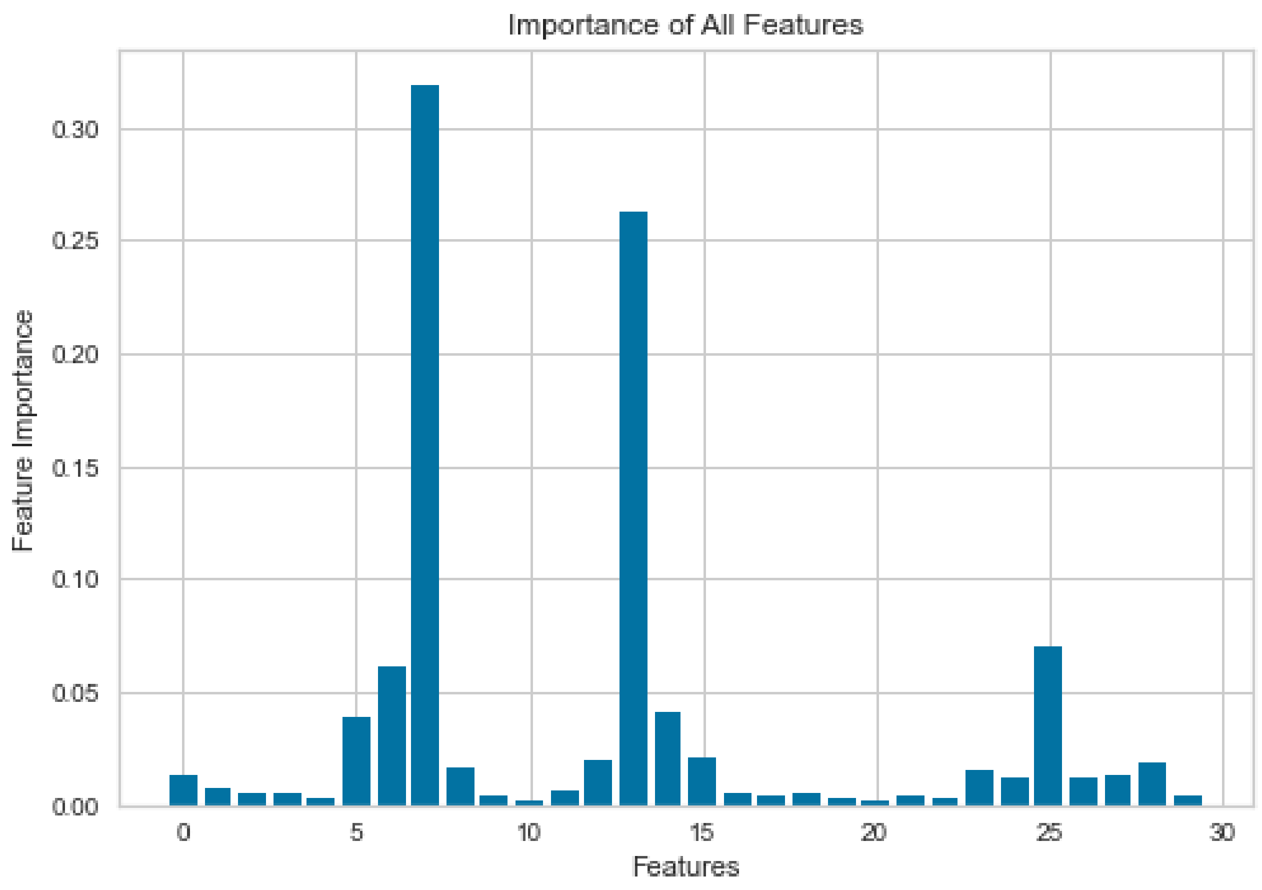
INDIVIDUAL FEATURE IMPORTANCE SCORES

- We see that the having_Sub_Domain, Prefix_Suffix, and URLURL_Length features provide the most benefit to the classification model
- These would be the most interesting features for further analysis and usage for recommendations and insights for the email add-in

RECURSIVE FEATURE ELIMINATION

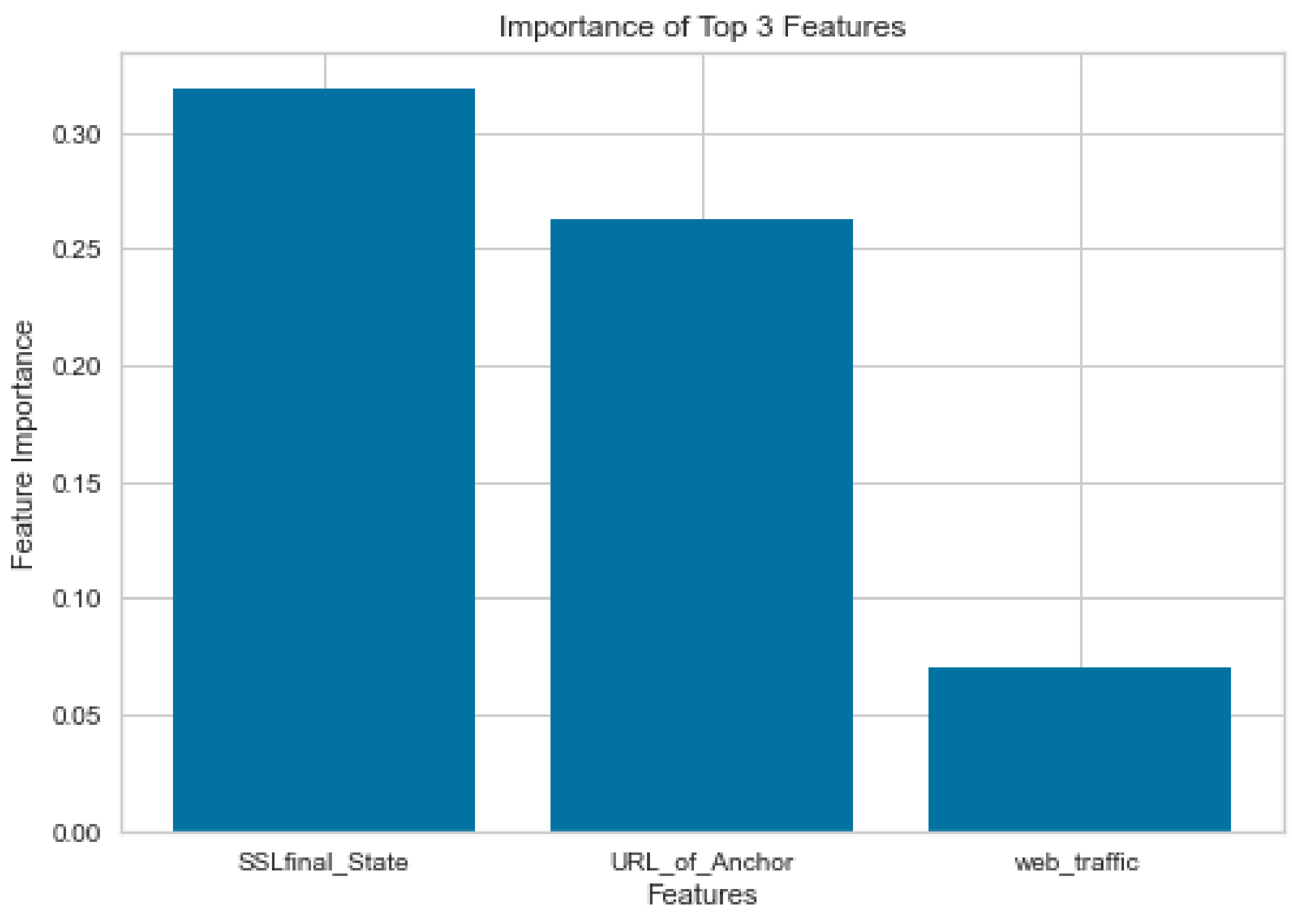


- Fit with Random Forest classification model
- Fit with all 30 features
- All 30 are used for optimal RFE cross-validation score of 0.972
- We see 2-3 provide excellent score



INDIVIDUAL FEATURE IMPORTANCE SCORES

- Feature importance scores for each of the 30 features
- We can see that two features clearly standout and a few others have moderate importance
- These features provide the most information gain



INDIVIDUAL FEATURE IMPORTANCE SCORES

- We see that the SSLfinal_State, URL_of_Anchor, and web_traffic features provide the most benefit to the classification model
- These would be the most interesting features for further analysis and usage for recommendations and insights for the email add-in

DETAILS OF FEATURES

- Please see report for more details about top features

EMAIL ADD-IN APPLICATION DELIVERABLE

- Development and publication of G-Mail add-in, which delivers augmented intelligence to security analysts as an easy to use tool for investigating suspicious and potentially malicious phishing emails
- Detailed documentation
  - Apps Script project access and code
  - Deployment steps to production environment
  - Necessary steps and components involved in preparing for verification and approval in the Marketplace
    - Due to confusing and convoluted official documentation, created a shortlist of requirements involved to prepare for initial publication

ANALYSIS ON IMPORTANT FEATURES OF PHISHING EMAILS DELIVERABLE

- Found useful academic research on engineered features for phishing analysis
- Delivered organized and documented report on important features from dataset
- Insights from report help to create a starting point for delivering reasons about an email's "phishing" verdict
  - Analysis can be applied to future versions of the add-in
  - Analysis can be applied to DTonomy's main platform, DTonomy AIR

WHAT WAS LEARNED?

- Software engineering skills
  - Learned new useful and job marketable platform, Apps Script
  - Hands-on experience with developing a real-world project and useable tool
  - Practice working with documentation and third-party API integration
- Data Analysis skills
  - Practice with Python packages and Jupyter notebooks
  - Augmented skills for reports, visualizations, and machine learning modeling
  - Feature selection analysis
  - Use of skills from DSA program for interpretation of results
- Start-up environment exposure
  - SCRUM framework
  - Source control collaboration and workflow
    - Branching
    - Pull-requests (entailed code reviews)

QUESTIONS?