# **Teacher Notes for Anomaly Detection (Phishing) \*DRAFT\***

## **Motivation and Essential Understandings**

Increase in web attacks are through spams, phishing, and malware. Identification of attack types is useful to take appropriate action.

## **Context and Dataset**

Students will use Machine Learning on a randomized dataset to distinguish between benign v. malicious URLs.

## **Learning Objectives**

Students will be able to:

## Describe how machine learning is used to detect phishing websites

## Use machine learning to determine phishing websites

* Compare machine learning methods for use in other domains

## **Data Science Concepts and Skills**

1. Summary statistics
2. Exploratory data analysis; Static data visualization
3. Predictive modeling using Logistic Regression; Decision tree

## **Students**

This lesson is for early and late undergraduate students. Students should be familiar with statistical concepts, basic data visualizations, and have worked in Excel.

## **Time to Teach this Lesson**

This lesson can be taught in 2 sessions in one week

1-hour prep, 2-hour class session; guided data analysis and independent work

## **Lesson Materials**

You will find all the lesson materials in the GenAI GitHub repository. The Jupyter notebook is not necessary to teach this lesson but is available to those who wish to teach more hands-on Data Science.

|  |  |  |
| --- | --- | --- |
| **Materials** | **File** | **Description** |
| Lecture |  | PPT lecture |
| Handout 1 |  |  |
| Handout 2 |  |  |
| Worksheet |  |  |
| Dataset | url\_data.csv |  |
| Image 1 | Not\_Facebook | Photos used in Jupyter Notebook |
| Image 2 | URL\_Breakdown | Photos used in Jupyter Notebook |
| Data dictionary |  | pdf of data dictionary explaining the column headings (data fields) in the datasets |
| Jupyter notebook | JupyterNotebook\_Jupyter\_Anomaly\_Detection\_Phishing\_2020 | Jupyter notebook |
| Jupyter notebook pdf | JupyterNotebook\_pdf\_Anomaly\_Detection\_Phishing\_2020 | pdf of annotated notebook |
| Template | Lesson\_Template\_Anomaly\_Detection\_Phishing\_2020.docx | Lesson planner \*DRAFT |

Photos used in Jupyter notebooks

## **Teaching Strategies**

* Review concepts in Statistics such as descriptive stats, correlation, and hypothesis testing
* Discuss the difference between Inferential vs. Predictive modeling

## **Lesson Narrative**

|  |
| --- |
| **Module 0: Pre-lesson** |

Review concepts of descriptive & multivariate Statistics and correlation of features

|  |
| --- |
| **Module 1: Dataset and Exploratory Data Analysis** |

**Histogram** – different kernel density plots for phishing & benign URLs: by length; number of special characters (dots, hyphens); domain name age. **Scatterplots** -domain length v. URL length.

**Features Analysis** to determine anomaly thresholds.

* Univariate features analysis: outlier detection – 2x std dev
* Multivariate features analysis: uncorrelated v. correlated

**Correlation matrix**

Prepare Data for ML tutorial (a) Feature selection/engineering(b) Label feature (0 = not Phishing/legit)

|  |  |  |
| --- | --- | --- |
| **Module 2: Explain Predictive Modeling** | | |

* Logistic Regression
* Decision Tree Classifier
* Train and test dataset
* Report Recall, Precision, F-score/confusion matrix/model comparison/best fit

|  |
| --- |
| **ASSIGNMENT:** Students write an end of course reflection on lessons learned & application to another problem (e.g., Risk Mitigation). |

|  |
| --- |
| **Module 3: Close Out** |

Post-assessment questions. TBD