# ITN 260 – Introduction to Machine Learning for Anomaly Detection (Phishing-Chapter 2)

|  |  |  |  |
| --- | --- | --- | --- |
| **Learning Objectives** | **Discussion(s)** | **Assignment(s)** | **Learning Resources + Media** |
| After this lesson, you 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 | Increase in web attacks are through spams, phishing, and malware. Identification of attack types is useful to take appropriate action.  First paper **(a)** talks about testing a dataset of URLs using Machine Learning (ML) techniques (Section 2). Focus on URL features (Section 3.2.1 & Tables 4 & 5). Note the split of the dataset to train & test (Table 15). Table 16 presents the confusion matrix.  Second paper **(b)** talks about multiple attack types: Phishing, Spam & Malware. Non-ML-based methods are discussed.  Describe each of these terms and give examples:   1. Features 2. Correlation 3. Accuracy of detection 4. Confusion Matrix 5. Regression | Phishing Attack: You will be working to use Machine Learning to distinguish between benign v. malicious URLs.  **Module 1: To prepare for this assignment**:   * Practice with Jupyter Notebook + Dataset #1   + Python libraries “**numpy, pandas, matplotlib**”   + Generate simple Histograms & determine bins   + Focus on basic visualizations * Review visualizations in Python for understanding **(c), (d), (e).**   + Focus on data fields (features)   + Explore “**seaborn**” distplot   **Assignment** - Phishing  **Module 2: Exploratory Analysis**   1. Dataset #2 2. Descriptive visualizations for EDA    1. **Histogram** – different kernel density plots for phishing & benign URLs: by length; number of special characters (dots, hyphens); domain name age    2. **Scatterplots** -domain length v. URL length    3. **Features Analysis** to determine anomaly thresholds       1. Univariate features analysis: outlier detection – 2x std dev       2. Multivariate features analysis: uncorrelated v. correlated    4. **Correlation matrix**   **Module 3: Prepare Data for ML tutorial**   1. Feature selection/engineering 2. Label feature (0 = not Phishing/legit)   **Module 4: Machine Learning**   1. Logistic Regression 2. Decision Tree Classifier 3. Train and test dataset 4. Report Recall, Precision, F-score/confusion matrix/model comparison/best fit   **End of course:** Reflection assignment on lessons learned & application to another problem (e.g., Risk Mitigation) | **SUMMER, 2020 – 8-week course; LMS: Canvas**  **Articles**   1. Dharmaraj R. Patil and Jayantrao B. Patil, 2018. Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification <http://www.isecure-journal.com/article_63041_7477a24f5e3a62ba0bf7a8416d70c51d.pdf> 2. Hyunsang Choi, Bin B. Zhu, Heejo Lee Gindi, 2016. Detecting Malicious Web Links and Identifying Their Attack Types <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/paper-65.pdf> 3. Brad Solomon, 2019. Visualizing Histograms with Matplotlib and Pandas from “Python Histogram Plotting: NumPy, Matplotlib, Pandas & Seaborn” <https://realpython.com/python-histograms/> 4. Jason Brownlee, (2019). Visualize Machine Learning Data in Python With Pandas <https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas/> 5. Jason Brownlee, (2019). Quick and Dirty Data Analysis with Pandas <https://machinelearningmastery.com/quick-and-dirty-data-analysis-with-pandas/> 6. Kavita Ganesan, (2019). Build Your First Text Classifier in Python with Logistic Regression <https://kavita-ganesan.com/news-classifier-with-logistic-regression-in-python/>   **Links for Visualization Student Guide**  An Introduction to Histograms <https://en.wikipedia.org/wiki/Histogram>  **Python Guidance**   1. Scikit-Learn <https://scikit-learn.org/> 2. Aurelien Geron, 2017. Hands-on Machine Learning with Scikit-Learn & TensorFlow. O’Reily Media, Inc.   **Datasets**  Dataset #1 - **for plotting & simple histograms; randomly generated**  Dataset #2 - from UMD Shady Grove Hackathon  **Jupyter Notebook**  **ZOOM Meetings: 2 synchronous sessions; recorded for tutorial** |